Pure Exploration Beyond Reward Feedback: The Role of Post-Action Context

Mohammad Shahverdikondori *
EPFL
mohammad.shahverdikondori@epfl.ch

Amir Mohammad Abouei *
EPFL
amir.abouei@epfl.ch

Alireza Rezaeimoghadam
EPFL
alireza.rezaeimoghadam@epfl.ch

Negar Kiyavash
EPFL
negar.kiyavash@epfl.ch

Abstract

We introduce the problem of best arm identification (BAI) with post-action context, a new BAI problem in a stochastic multi-armed bandit environment and the fixed-confidence setting. The problem addresses the scenarios in which the learner receives a *post-action context* in addition to the reward after playing each action. This post-action context provides additional information that can significantly facilitate the decision process. We analyze two different types of the post-action context: (i) *separator*, where the reward depends solely on the context, and (ii) *non-separator*, where the reward depends on both the action and the context. For both cases, we derive instance-dependent lower bounds on the sample complexity and propose algorithms that asymptotically achieve the optimal sample complexity. For the separator setting, we propose a novel sampling rule called *G-tracking*, which uses the geometry of the context space to directly track the contexts rather than the actions. For the non-separator setting, we do so by demonstrating that the Track-and-Stop algorithm can be extended to this setting. Moreover, in both settings, we theoretically and empirically show that algorithms that ignore the post-action context are sub-optimal. Finally, our empirical results showcase the advantage of our approaches compared to the state of the art.

1 Introduction

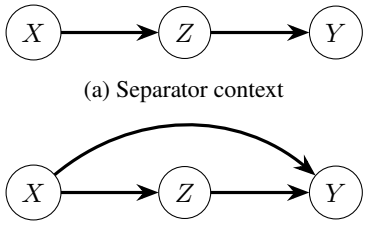
Multi-armed bandit (MAB) refers to a class of sequential decision-making problems, where a learner selects actions (arms) in order to maximize a reward. MAB has widespread applications in various domains, such as clinical trials [Tho33], dynamic pricing [KL03, BZ09], recommender systems [LCLS10], and resource allocation [GKJ12]. Depending on the learner's goal and constraints, different objectives may be pursued. For example, if the learner's goal is to minimize cumulative regret, they must balance the exploration-exploitation trade-off [Aue02, GC11]. Alternatively, in pure exploration problems, the learner aims to identify a desired object rather than maximize cumulative reward during learning [Ste23, DK19, SEK26a]. In the *Best Arm Identification (BAI)* setting, a canonical pure exploration problem, the learner seeks the arm with the highest expected reward and must minimize the sample complexity, i.e., the number of rounds needed to identify this arm [MT04, AB10, GK16]. We focus on BAI with *fixed-confidence*, where the error probability δ for identifying the best arm is fixed, and the objective is to minimize the sample complexity [GK16, AB10, Kau20, JN14, MT04].

*Authors contributed equally.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

While the classic MAB model is suitable for a broad range of applications, additional side information about the environment can lead to more efficient algorithms. In the bandit literature, various types of side information have been considered. For example, *causal bandits* [LLR16, LMTY20] or *linear bandits* [Aue02, AYPS11] impose specific structures on the actions, and *contextual bandits* [TM17, LZ07, RKB$^+$21] allow the learner to observe a context before choosing an action. In this work, we study a form of side information known as *mediator feedback* or *post-action context*, which has been considered in the setting of cumulative regret minimization by several recent studies [MGG$^+$19, MPDR21, BWR22, EMQ$^+$23, ECBMR24, LAR24]. Specifically, after choosing an action in each round, the learner receives intermediate feedback from the environment along with the reward. This post-action context can substantially accelerate the process of identifying the best arm. We consider this new problem in the fixed-confidence setting, aiming to show how leveraging post-action context can reduce the sample complexity for different environments.

Figure 1 illustrates two structures pertaining to post-action context, which may also be interpreted as causal graphs. In both structures, variables X, Z, and Y represent the action, the post-action context, and the reward, respectively. In Figure 1a, once Z is observed, the distribution of Y depends only on Z, making X and Y conditionally independent given Z. We call this the *separator* context. By contrast, in Figure 1b, the context Z is informative but not a sufficient statistic of X for determining Y. In other words, the distribution of Y depends on both X and Z. We refer to this as the *non-separator* context. Although the separator context can be viewed as a special case of the non-separator, an algorithm optimized for the latter may not be optimal for the former. This is because a separator context provides the important side information that X and Y are conditionally independent given Z. We analyze both settings in this work.



(a) Separator context

(b) Non-separator context

Figure 1: Two possible structures for the post-action context.

Example 1. Consider an advertising platform that compares a finite set of randomized serving policies. Here, X is the selected policy, Z is the advertisement (or slate) actually displayed, and Y is the resulting reward, such as a click or a conversion. Each policy is implemented through programmed randomization, so $P(Z \mid X)$ is known from the policy specification. Users respond to the advertisement they see, regardless of which policy selected it. The goal is to identify the best policy, for instance during the testing phase of a recommender system, before deployment. Since an observed advertisement is informative about every policy that may display it, a policy with a lower expected reward can still be valuable during exploration if it displays advertisements whose rewards must be estimated more accurately.

Example 2. Consider a healthcare scenario for diabetes management where X denotes a chosen treatment regimen (e.g., medication or dosage), Z is an intermediate biomarker such as short-term blood sugar levels, and Y is the clinical target of improvement (e.g., glucose stability and range). In each round, a clinician assigns a treatment X to a patient, measures Z after a short interval, and later observes Y. Tracking this intermediate observation may provide additional information about the alternative treatments. The distribution $P(Z|X)$ can often be reliably estimated from electronic health records or prior clinical studies, making it valuable prior knowledge. By incorporating Z, clinicians can adapt treatment strategies more efficiently, reducing the number of trials needed to identify an optimal regimen.

Comparison to closely related work. [LMTY20, ECBMR24] and [LLR16] study similar settings with separating contexts, focusing on regret minimization and simple regret optimization, respectively. Several works study best-arm identification (BAI) in contextual bandits [LRJ$^+$22, KA21], including the proportional mode of [RKB$^+$21]. Although these problems are related to ours, there is a fundamental distinction: in contextual bandits, the context is generated before the action is selected, and the action is chosen based on the observed context. [RKB$^+$21] also considers more general scenarios, including active, agnostic, and oblivious modes. For example, in the agnostic mode, the context is independent of the action but is observed only after an action is chosen. However, none of these works captures the structure of our problem, where the context is generated as a consequence of the chosen action. Finally, [CBJD24] considers a similar problem; however, in their setting, the

learner can choose Z in each round, whereas in ours the learner cannot directly choose Z. For a more detailed discussion of related work, see Appendix A.

The presence of post-action context can fundamentally alter the optimal strategy for selecting arms. In classic BAI problems, the optimal algorithms proposed by [GK16] focus more time on pulling the arms with the higher observed rewards, which is natural when no additional side information is available. However, in the presence of post-action context, collecting more samples from a certain context may expedite the identification of the best arm. In such scenarios, choosing a suboptimal arm that increases the probability of observing that context is more effective than selecting the arm with the maximum reward. We show the suboptimality of the methods that ignore the post-action context both theoretically (Section 4) and empirically (Section 5).

Contributions. We introduce *best-arm identification with post-action context*, a fixed-confidence pure exploration problem in which each action generates both a reward and an informative post-action context. We study two structural regimes: a *separator* setting, where the reward depends on the action only through the post-action context, and a *non-separator* setting, where the reward may depend on both the action and the context. For both regimes, we derive instance-dependent lower bounds and design algorithms that achieve the corresponding asymptotically optimal sample complexity. In the separator setting, the lower bound depends only on the frequencies with which contexts are observed; this motivates our new sampling rule, *G-tracking*, which tracks optimal context frequencies directly rather than tracking arm pull frequencies. In the non-separator setting, we show that the classical Track-and-Stop framework can be extended to exploit post-action contexts optimally. The main novelty lies in the separator setting, where the relevant allocation shifts from action frequencies to induced context frequencies, while the non-separator setting complements it by covering rewards that depend on both the action and the context. Finally, we show that algorithms that ignore the post-action context can be strictly sub-optimal and characterize this suboptimality.

2 Preliminaries and Problem Setup

Notations. The set $[n]$ denotes $\{1, 2, \dots, n\}$ and Δ^{n-1} represents the $(n-1)$-dimensional standard simplex, defined as $\{w \in \mathbb{R}^n \mid w_i \geq 0 \wedge \sum_{i=1}^n w_i = 1\}$. The notation $d(P, Q)$ denotes the Kullback–Leibler (KL) divergence between two probability measures P and Q. Additionally, $d_B(\delta, 1-\delta)$ refers to the KL divergence between two Bernoulli random variables with parameters δ and $1-\delta$. The convex hull of n vectors $\mathcal{A}_i \in \mathbb{R}^d$ is denoted by $\text{conv}(\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n\}) = \{\sum_{i=1}^n \lambda_i \mathcal{A}_i \mid \lambda_i \geq 0 \ \wedge \sum_{i=1}^n \lambda_i = 1\}$. A table summarizing the notation is provided at the beginning of the appendix.

Best arm identification with post-action context problem. In this problem, a player interacts with a multi-armed bandit environment $\mathcal{V}$ with n arms. In each round t, the player selects an action $X = i_t \in [n]$ and observes a pair (z_t, y_t), where z_t is the value of the post-action context variable Z, and y_t is the realization of the reward variable Y.

We assume the context variable Z is discrete and takes values in $[k]$, depending only on the arm pulled. The context probability matrix $\mathcal{A} = [\mathcal{A}_1 | \mathcal{A}_2 | \dots | \mathcal{A}_n] \in \mathbb{R}^{k \times n}$, where $\mathcal{A}_{j,i} = P(Z = j \mid X = i)$, encodes the probability of contexts given each action. We further assume that the reward distribution $P(Y \mid X = i, Z = j)$ follows a Gaussian distribution with unit variance. The matrix $\boldsymbol{\mu} = [\boldsymbol{\mu}_1 | \boldsymbol{\mu}_2 | \dots | \boldsymbol{\mu}_n] \in \mathbb{R}^{k \times n}$ represents the mean values of the reward distributions, such that $\mu_{j,i} = \mathbb{E}(Y | X = i, Z = j)$.

In the case of the separator context, the reward depends only on Z, that is, $P(Y \mid X = i, Z = j) = P(Y \mid Z = j)$, for all $i \in [n]$ and $j \in [k]$. This implies that all columns of $\boldsymbol{\mu}$ are identical. We denote the joint distribution of the context and reward for a given action i by $P_i^{\boldsymbol{\mu}, \mathcal{A}} = P(Z, Y \mid X = i)$. The expected reward for a given action i is computed as

$$\mathbb{E}[Y \mid X = i] = \sum_{j=1}^{k} P(Z = j | X = i) \mathbb{E}[Y | Z = j, X = i] = \mathcal{A}_i^\top \boldsymbol{\mu}_i.$$

The best arm is the arm with the maximum expected reward, i.e., $i^*(\boldsymbol{\mu}, \mathcal{A}) = \text{argmax}_{i \in [n]} \mathcal{A}_i^\top \boldsymbol{\mu}_i$, where the mean matrix $\boldsymbol{\mu}$ and the context probability matrix $\mathcal{A}$ characterize the instance. In this work, we assume that $\mathcal{A}$ is known to the learner and that the best arm is unique. Let $\mathcal{I}(\mathcal{A}) \subset \mathbb{R}^{k \times n}$ denote

the set of $\boldsymbol{\mu}$ matrices that imply a unique best arm for a given $\mathcal{A}$. For simplicity, as $\mathcal{A}$ is known, we often use $\mathcal{I}, i^*(\boldsymbol{\mu})$, and P^{μ} instead of $\mathcal{I}(\mathcal{A}), i^*(\boldsymbol{\mu}, \mathcal{A})$, and $P^{\mu,\mathcal{A}}$, respectively. When the best arm is used as an index, we write i^* for $i^*(\boldsymbol{\mu}, \mathcal{A})$; for instance, $\mathcal{A}_{j,i^*}$ and $\mathcal{A}_{i^*}$ denote $\mathcal{A}_{j,i^*(\boldsymbol{\mu},\mathcal{A})}$ and the $i^*(\boldsymbol{\mu}, \mathcal{A})$-th column of $\mathcal{A}$.

We consider the best arm identification problem in the *fixed-confidence* setting, where the player aims to identify the best arm $i^*(\boldsymbol{\mu})$ for any $\boldsymbol{\mu} \in \mathcal{I}$ with a pre-specified error probability $\delta \in (0, 1)$. A sequential learning algorithm for this setting consists of three main components: (i) A *sampling rule* (deterministic or stochastic) that determines which arm to pull in round t based on the history up to round $t-1$, denoted by $\mathcal{H}_{t-1} = (i_1, (z_1, y_1), i_2, (z_2, y_2), \ldots, i_{t-1}, (z_{t-1}, y_{t-1}))$, (ii) A *stopping rule* τ, based on $\mathcal{H}_t$, determines when to stop the process, and (iii) A *decision rule* $\hat{i}_\tau$, which recommends the best action.

For any $\delta \in (0, 1)$, an algorithm is δ-*correct* if for any $\mathcal{A}$ and $\boldsymbol{\mu} \in \mathcal{I}(\mathcal{A})$ it satisfies $P_{\boldsymbol{\mu},\mathcal{A}}(\tau_\delta < \infty, \hat{i}_\tau \neq i^*(\boldsymbol{\mu})) \leq \delta$. The goal of the learner is to find a δ-correct algorithm that minimizes the expected number of samples $\mathbb{E}_{\boldsymbol{\mu},\mathcal{A}}[\tau_\delta]$.

The sub-optimality gap of arm i is defined as

$$\Delta_i \triangleq \mathbb{E}[Y \mid X = i^*(\boldsymbol{\mu})] - \mathbb{E}[Y \mid X = i] = \mathcal{A}_{i^*}^\top \boldsymbol{\mu}_{i^*} - \mathcal{A}_i^\top \boldsymbol{\mu}_i.$$

During the learning process, let $\hat{\boldsymbol{\mu}}(t)$ denote the matrix containing the empirical estimates of all entries of $\boldsymbol{\mu}$ based on the samples collected up to round t and define

$$\hat{\Delta}_i(t) \triangleq \mathcal{A}_{i^*(\hat{\boldsymbol{\mu}}(t))}^\top \hat{\boldsymbol{\mu}}_{i^*(\hat{\boldsymbol{\mu}}(t))}(t) - \mathcal{A}_i^\top \hat{\boldsymbol{\mu}}_i(t),$$

which is the gap of arm i relative to the arm with the best empirical mean up to round t estimated using $\hat{\boldsymbol{\mu}}(t)$. Finally, $N_i^X(t)$, $N_j^Z(t)$, and $N_{j,i}(t)$ represent the number of times action i, context j, and joint action-context pair (i, j) have been observed up to round t, respectively.

Summary of Assumptions. We assume that the set of contexts is finite. The context probability matrix $\mathcal{A}$ is known, and the best arm is unique. In Appendix H, we provide a discussion on the complexity of the problem when $\mathcal{A}$ is unknown. The known-$\mathcal{A}$ assumption is natural when each arm is a predefined randomized policy, as in Example 1. Note that when $\mathcal{A}$ is unknown, simply plugging an estimate $\hat{\mathcal{A}}_t$ into our algorithms and treating it as the true $\mathcal{A}$ does not work in general; see Appendix H for details and a counterexample. The reward distribution for each action-context pair follows a Gaussian distribution with unit variance; however, our results can be easily generalized to the one-parameter exponential family. Additionally, we assume positivity, meaning that for each action, each context occurs with positive probability, i.e., $a_{\min} = \min_{i,j} \mathcal{A}_{i,j} > 0$.

2.1 General Lower Bound

For the classical best arm identification problem, [GK16] established a lower bound on the expected sample complexity of any δ-correct algorithm. Extending this lower bound to our problem setting is straightforward, as we show in Proposition 2.1. Before stating the proposition, let us define $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ as the set of alternative parameter matrices $\boldsymbol{\mu}' \in \mathcal{I}$ such that $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\mu}')$.

Proposition 2.1. *For any bandit environment with parameters $\mathcal{A}$ and $\boldsymbol{\mu} \in \mathcal{I}(\mathcal{A})$ and any δ-correct algorithm, $\mathbb{E}_{\boldsymbol{\mu},\mathcal{A}}[\tau_\delta] \geq T^*(\boldsymbol{\mu}, \mathcal{A}) d_B(\delta, 1-\delta)$, where $T^*(\boldsymbol{\mu}, \mathcal{A})$ is*

$$T^*(\boldsymbol{\mu}, \mathcal{A})^{-1} \triangleq \sup_{\mathbf{w} \in \Delta^{n-1}} \inf_{\boldsymbol{\mu}' \in Alt(\boldsymbol{\mu},\mathcal{A})} \sum_{i \in [n]} w_i d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}). \tag{1}$$

This lower bound indicates that the optimal sampling strategy is to play each action i in proportions to w_i^*, a solution of Equation (1). Note that the bound is similar to a previous result [GK16], with the key distinction that it depends on the context variable through $P_i^{\boldsymbol{\mu}} = P(Z, Y \mid X = i)$.

We derive lower bounds for the BAI problem under both separator and non-separator post-action context settings by explicitly characterizing Equation (1). The key difference lies in the structure of the alternative set $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$, which varies between the two settings. In the separator case, all $\boldsymbol{\mu}_i$ vectors are identical, imposing additional constraints on $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ and yielding a strictly smaller lower bound. This tighter bound motivates the design of a tailored sampling algorithm to achieve optimal sample complexity in the separator setting. The analysis for the non-separator case is deferred to Appendix B, as the derivation closely follows the separator case and known sampling rules such as D-tracking are sufficient to achieve optimality.

Algorithm 1 Separator Track and Stop (STS)

Input: Context probability matrix $\mathcal{A}$.
Initialization: Pull arms until collecting at least one sample from $P(Y|Z=j)$ for each $j \in [k]$, then update t accordingly.
while $\hat{\Lambda}(t) \leq \hat{c}_t(\delta)$ **do**
  Find policy $\mathbf{p}(t+1) \in \text{conv}(\mathcal{A})$ based on G-tracking rule.
  Play according to a distribution π such that $\mathcal{A}\pi = \mathbf{p}(t+1)$.
  Update $\hat{\boldsymbol{\mu}}(t)$ and $t \leftarrow t+1$.
end while
Output: $i^*(\hat{\boldsymbol{\mu}}(t))$.

3 Separator Context

This section presents the results for the *bandit with separator post-action context*, where the reward distribution depends solely on the value of the context, i.e., $P(Y \mid X=i, Z=j) = P(Y \mid Z=j)$, as illustrated in Figure 1a. This condition implies that all columns of the matrix $\boldsymbol{\mu}$ are identical. For ease of notation, in this section, we represent $\boldsymbol{\mu}$ as a vector in $\mathbb{R}^k$.

3.1 Lower Bound

The following theorem provides a lower bound on the expected sample complexity of any δ-correct algorithm in the separator setting. This lower bound is derived by explicitly solving the minimization problem defined in (1).

Theorem 3.1 (Separator Lower Bound)**.** *Let $\delta \in (0,1)$. Consider a bandit instance with a separator context and Gaussian reward distribution with unit variance, parameterized by the matrix $\mathcal{A}$ and the vector $\boldsymbol{\mu}$. Then, any δ-correct algorithm with stopping time τ_δ satisfies $\mathbb{E}[\tau_\delta] \geq T^*_S(\boldsymbol{\mu}, \mathcal{A}) d_B(\delta, 1-\delta)$, where*

$$T^*_S(\boldsymbol{\mu}, \mathcal{A})^{-1} = \sup_{\mathbf{w} \in \Delta^{n-1}} \min_{i \neq i^*(\boldsymbol{\mu})} \frac{\Delta_i^2}{2 \sum_{j \in [k]} \frac{(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,i})^2}{\sum_{l \in [n]} w_l \mathcal{A}_{j,l}}}. \tag{2}$$

The values $w_{z,j} \triangleq \sum_{l \in [n]} w_l \mathcal{A}_{j,l}$ in Equation (2) represent the expected frequency of observing context $Z = j$. Note that the optimization problem in (2) depends only on $w_{z,j}$s. Consequently, defining the vector $\mathbf{w}_z = (w_{z,1}, w_{z,2}, \ldots, w_{z,k})$, Equation (2) can be expressed as

$$T^*_S(\boldsymbol{\mu}, \mathcal{A})^{-1} = \sup_{\mathbf{w}_z \in \text{conv}(\mathcal{A})} \min_{i \neq i^*(\boldsymbol{\mu})} \frac{\Delta_i^2}{2 \sum_{j \in [k]} \frac{(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,i})^2}{w_{z,j}}}, \tag{3}$$

where $\text{conv}(\mathcal{A})$ is the convex hull of the points $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$, and $\mathbf{w}_z \in \text{conv}(\mathcal{A})$ because $\mathbf{w}_z$ is a convex combination of these points. The optimal solution to this new formulation represents the expected frequency of observing context values generated by any optimal algorithm.

It can be shown that the objective functions in both Equations (2) and (3) are concave. Since the domains Δ^{n-1} and $\text{conv}(\mathcal{A})$ are convex, both optimization problems can be tackled using convex optimization techniques. For more details on solving such optimization problems in BAI settings, see [WTP21, Mén19].

3.2 Learning Algorithm

Stopping Rule. The stopping rule operates independently of the sampling rule and determines, at each step, whether sufficient information has been gathered to terminate the algorithm or if further sampling is necessary. A widely used approach for designing stopping rules in one-parameter exponential family bandits, including Gaussian bandits, is based on Generalized Likelihood Ratio (GLR) tests [GK16, KK21]. We describe how we adapt these tests to our setting in Appendix D. The GLR statistic in this setting is defined as

$$\hat{\Lambda}(t) \triangleq \inf_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \sum_{j \in [k]} N_j^Z(t) \frac{(\hat{\boldsymbol{\mu}}_j(t) - \boldsymbol{\mu}'_j)^2}{2}, \tag{4}$$

provided that $\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}(\mathcal{A})$ and otherwise $\hat{\Lambda}(t)$ is set to zero. $\hat{\Lambda}(t)$ reflects the current confidence in the best arm identified based on the observations so far. A higher value of $\hat{\Lambda}(t)$ indicates greater confidence. This definition can be simplified to

$$\hat{\Lambda}(t) = \min_{i \neq i^*(\hat{\boldsymbol{\mu}}(t))} \frac{\hat{\Delta}_i^2}{2\sum_{j\in[k]} \frac{(\mathcal{A}_{j,i^*(\hat{\boldsymbol{\mu}}(t))} - \mathcal{A}_{j,i})^2}{N_j^Z(t)}}.$$

Appendix D and D.2 provide a proof of how to compute the GLR statistic and a justification of the above simplification.

To decide whether to terminate or continue playing, $\hat{\Lambda}(t)$ is compared to a sequential threshold $\hat{c}_t(\delta)$, which determines if the confidence is sufficiently high. To design the sequential thresholds, we apply the result by [KK21] for adaptive sequential testing, which suggests

$$\hat{c}_t(\delta) = 2\sum_{j\in[k]} \ln\left(4 + \ln\left(N_j^Z(t)\right)\right) + kC^g\left(\frac{\ln\left(\frac{1}{\delta}\right)}{k}\right), \tag{5}$$

where C^g is the same function introduced in Equation (11). The stopping time τ_δ is then defined as

$$\tau_\delta \triangleq \inf\{t \in \mathbb{N} \mid \hat{\Lambda}(t) > \hat{c}_t(\delta)\}. \tag{6}$$

At the stopping time τ_δ, the final suggestion $\hat{i}_\tau$ is equal to the unique estimated best arm $i^*(\hat{\boldsymbol{\mu}}(\tau_\delta))$. The following lemma establishes the correctness of this stopping rule.

Lemma 3.2. *Consider a bandit instance with a separator context and Gaussian reward distribution with unit variance, parameterized by the matrix $\mathcal{A}$ and the vector $\boldsymbol{\mu}$. Any algorithm with the stopping rule of* (6) *is δ-correct, that is, $P_{\boldsymbol{\mu},\mathcal{A}}(\tau_\delta < \infty, \hat{i}_\tau \neq i^*(\boldsymbol{\mu})) \leq \delta$.*

Sampling Rule. As mentioned earlier, the lower bound and stopping rule depend solely on how often each context value is observed, regardless of the actions that result in these contexts. Motivated by this observation, we introduce our new sampling method, called *Geometric Tracking (G-tracking)*, which directly tracks the optimal frequency of observed contexts.

Note that this setting can be viewed as an active off-policy learning problem in k-armed bandit, where each arm corresponds to a context value. However, the agent cannot directly select a context to play. Instead, they have access to n distinct policies (the original n arms in our problem), where choosing policy i, probabilistically selects a context according to the known distribution $\mathcal{A}_i$. Each vector $\mathcal{A}_i$ lies in the $(k-1)$-dimensional simplex and encodes the probability of selecting a context by playing policy i. Under this interpretation, the policy space is the convex hull of the points $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$, denoted by $\text{conv}(\mathcal{A})$. In each round t, choosing a policy $\mathbf{p}(t) \in \text{conv}(\mathcal{A})$ is equivalent to fixing a distribution on context Z. Notably, there exists (although not necessarily unique) a distribution over arms X that is compatible with $\mathbf{p}(t)$, and the agent samples an arm at time t according to this distribution.

We now propose a new sampling rule that directly tracks the optimal frequency of observing contexts and asymptotically achieves optimal sample complexity when combined with the stopping rule in Equation (6). For any vector $\boldsymbol{\mu} \in \mathcal{I}(\mathcal{A})$ (i.e., a unique best arm exists), we define the set $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ as the set of solutions to the following optimization problem

$$\operatorname*{argmax}_{\mathbf{w}_z \in \text{conv}(\mathcal{A})} \min_{i \neq i^*(\boldsymbol{\mu})} \frac{\Delta_i^2}{2\sum_{j\in[k]} \frac{(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,i})^2}{w_{z,j}}}, \tag{7}$$

which corresponds to the formulation in Equation (3). Unlike the non-separator setting, in the separator setting, the optimal weights may not be unique, i.e., $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ is not a singleton. The optimization problem (7) is convex for each $\boldsymbol{\mu} \in \mathcal{I}$ and can be solved using numerical methods such as Frank-Wolfe [Jag13, WTP21]. Although (7) is non-smooth, the objective is the minimum of a finite number of smooth functions. The standard Frank–Wolfe algorithm does not provide theoretical convergence guarantees in this case, but [WTP21] establish such guarantees for optimization problems whose objectives admit this structure. Each iteration costs $\mathcal{O}(nk)$, namely a gradient evaluation of cost $\mathcal{O}(\max(k, n))$ and a linear subproblem over $\text{conv}(\mathcal{A})$ of cost at most $\mathcal{O}(nk)$. Since $\mathcal{O}(1/\varepsilon)$ iterations typically suffice to reach an ε-accurate solution, the per-round cost of solving (7) is $\mathcal{O}(nk/\varepsilon)$.

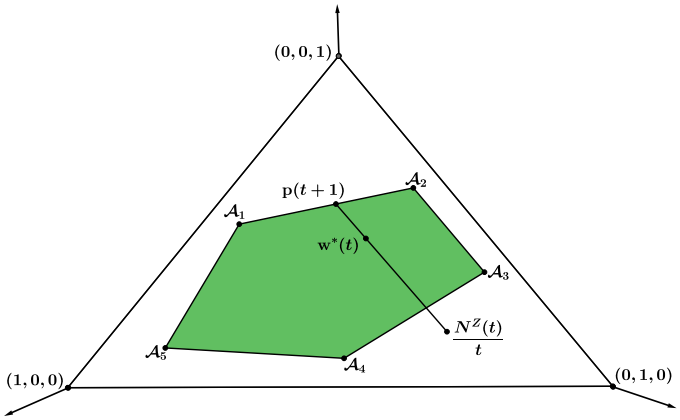


Figure 2: Illustration of G-tracking rule for an instance with $k = 3, n = 5$. The triangle depicts the two-dimensional simplex, and the green area shows the policy space conv($\mathcal{A}$).

At each round t, our sampling rule first solves the optimization problem (7) using the current empirical estimate $\hat{\boldsymbol{\mu}}(t)$ to find a solution $\mathbf{w}^*(t) \in \mathbf{w}_z^*(\hat{\boldsymbol{\mu}}(t), \mathcal{A})$. This vector lies within conv($\mathcal{A}$). Recall that vector $\frac{\mathbf{N}^Z(t)}{t}$ denotes the actual frequency of observed contexts up to round t and it lies in the simplex Δ^{k-1}. Our sampling rule, called G-tracking, determines the next policy $\mathbf{p}(t+1) \in$ conv($\mathcal{A}$) by connecting $\frac{\mathbf{N}^Z(t)}{t}$ to $\mathbf{w}^*(t)$ and extending this line until it intersects the boundary of conv($\mathcal{A}$). The algorithm then pulls arms according to a distribution $\pi \in \Delta^{n-1}$, where $\mathcal{A}\pi = \mathbf{p}(t+1)$. A compatible distribution π always exists because $\mathbf{p}(t+1) \in$ conv($\mathcal{A}$), but it may not be unique. For the formal mathematical formulation of G-tracking, see Appendix F.1.

Figure 2 illustrates the sampling rule for an instance with $k = 3$ and $n = 5$. Note that while the policy points always lie within conv($\mathcal{A}$), $\frac{\mathbf{N}^Z(t)}{t}$ may lie outside this set (as shown in Figure 2). In such cases, the policy point $\mathbf{p}(t+1)$ is the intersection point of the line passing through $\frac{\mathbf{N}^Z(t)}{t}$ and $\mathbf{w}^*(t)$ with the boundary of policy space, ensuring that $\mathbf{w}^*(t)$ lies between $\mathbf{p}(t+1)$ and $\frac{\mathbf{N}^Z(t)}{t}$.

The main advantage of the G-tracking rule is that it ensures $\mathbf{w}^*(t)$ is a convex combination of the observed context frequencies $\frac{\mathbf{N}^Z(t)}{t}$ and the next action $\mathbf{p}(t+1)$. Thus, if a context j is under-explored $\left(\frac{N_j^Z(t)}{t} \leq w_j^*(t)\right)$, then $\mathbf{p}(t+1)$ allocates a probability greater than $w_j^*(t)$ to that context. Conversely, for each over-explored context j $\left(\frac{N_j^Z(t)}{t} \geq w_j^*(t)\right)$, the next policy allocates a probability lower than $w_j^*(t)$. This "balancing" property holds for all contexts simultaneously. Note that the balancing property holds for every point on the segment connecting $\mathbf{w}^*(t)$ to $\mathbf{p}(t+1)$. Therefore, our proof remains valid for any algorithm that picks a point on this segment. In our algorithm, we choose the $\mathbf{p}(t+1)$ on the boundary of the convex hull to achieve the fastest possible tracking of the optimal weights.

The pseudocode of our algorithm called *Separator Track and Stop (STS)* is presented in Algorithm 1 which uses G-tracking for sampling with the stopping rule of (6). The following theorem states that our algorithm achieves the optimal sample complexity.

Theorem 3.3 (Separator Upper Bound)**.** *Algorithm 1 applied to a bandit instance with a separator context and Gaussian reward distribution with unit variance, parameterized by the matrix $\mathcal{A}$ and the vector $\boldsymbol{\mu}$ attains*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu},\mathcal{A}}[\tau_\delta]}{\ln(1/\delta)} \leq T_S^*(\boldsymbol{\mu}, \mathcal{A}),$$

where $T_S^*(\boldsymbol{\mu}, \mathcal{A})$ is defined in Equation (2)*.*

4 Sub-Optimality of Ignoring Post-Action Context

In this section, we explain why algorithms that ignore the post-action context are suboptimal in the separator setting. The same discussion for the non-separator setting is deferred to Appendix B.3

4.1 Separator Context

In the non-separator setting, the sub-optimality of ignoring Z stems solely from the increased variance in the reward distributions (i.e., the increase in the sub-Gaussian parameter). In contrast, the separator setting involves more substantial structural differences. The key distinction is that, in the separator setting, the action and reward are conditionally independent given the post-action context. As a result, any observation with context $Z = j$ is informative about μ_j, regardless of which action produced it. This property implies that algorithms that ignore Z can suffer arbitrarily large sub-optimality in the separator setting, even when the range of reward means is small. In the following, we describe two potential sources of sub-optimality in this case:

1. In the separator setting with known context probability matrix $\mathcal{A}$, the learner only needs to estimate the means of k different unit-variance Gaussian variables (i.e., the entries of $\boldsymbol{\mu}$) with high confidence to identify the best arm. In cases where $n \gg k$, ignoring Z forces the learner to estimate the means of n different random variables (sub-Gaussian with variance greater than 1) with high confidence, which can significantly increase the required sample complexity.

2. The correlation among the arms in the separator setting implies that, even when the mean values lie within a small range and $n = k$, algorithms that ignore Z may still be highly sub-optimal. Consider the following example with $k = n = 3$ and the parameters:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0.5 \\ 1 \end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix} \epsilon & 2\epsilon & \frac{1}{3} \\ 2\epsilon & \epsilon & \frac{1}{3} \\ 1-3\epsilon & 1-3\epsilon & \frac{1}{3} \end{bmatrix},$$

where ϵ is a small positive number. In this example, the first arm is optimal, and the gap between arms 1 and 2 is $\Delta_2 = \mathcal{A}_1^\top \boldsymbol{\mu} - \mathcal{A}_2^\top \boldsymbol{\mu} = \frac{\epsilon}{2}$. According to known lower bounds in the BAI literature, any algorithm solving this instance with confidence δ must take more than $\frac{4}{\epsilon^2} \ln(\frac{1}{\delta})$ samples [KCG16]. However, by using the post-action context, repeatedly pulling the third arm (which has a large sub-optimality gap and is thus rarely played by algorithms ignoring Z) can stop after just $\mathcal{O}(\ln(\frac{1}{\delta}))$ samples. This follows from a standard concentration argument, since the components of $\boldsymbol{\mu}$ are separated by a constant gap, $\mathcal{O}(\ln(\frac{1}{\delta}))$ samples are enough to recover their ordering with high probability. This example illustrates that the cost of ignoring Z in the separator setting can be substantial.

4.2 Reduction to Linear Bandit

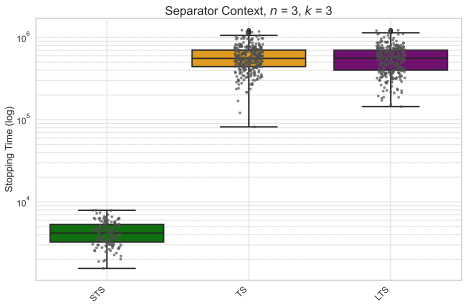
A more sophisticated approach to solve the separator setting that ignores Z but takes advantage of knowing the context probability matrix $\mathcal{A}$ is to reduce the problem to a linear bandit problem.

In this reduction, each arm i is mapped to $\mathcal{A}_i \in \mathbb{R}^k$. Then the reward from pulling action i can be written as

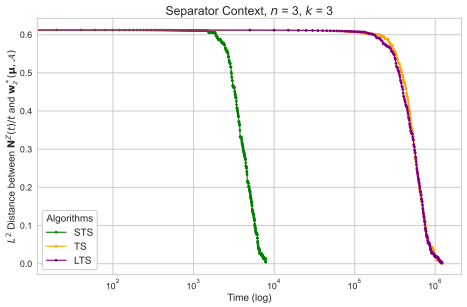
$$Y_i = \sum_{j=1}^{k} \mathcal{A}_{j,i} \boldsymbol{\mu}_j + R + \varepsilon = \langle \mathcal{A}_i, \boldsymbol{\mu} \rangle + R + \varepsilon,$$

where ε is a standard Gaussian, and R is a discrete variable over $\{\boldsymbol{\mu}_j - \mathcal{A}_i^\top \boldsymbol{\mu} \mid j \in [k]\}$ such that $\forall j \in [k] : P(R = \boldsymbol{\mu}_j - \mathcal{A}_i^\top \boldsymbol{\mu}) = \mathcal{A}_{j,i}$. If we assume the entries of $\boldsymbol{\mu}$ are bounded from both sides, then similar to the non-separator case, we conclude that $R + \varepsilon$ is a sub-Gaussian noise but with a variance greater than 1 (which means that the sub-Gaussian parameter is bigger than 1). Note that a similar reduction also works for the non-separator setting by mapping each action $X = i$ to a vector in $\mathbb{R}^{nk}$ whose entries in positions $(i-1)k + 1$ to ik are equal to $\mathcal{A}_i$ and whose other entries are zero.

In the following analysis, we disregard the increase in variance due to the reduction to linear bandit, as discussed above, and assume $\sigma = 1$. Even with this simplification, we show that by incorporating the post-action context, the sample complexity lower bound is always less than (or equal to) that of



(a) Comparison of stopping times over 150 runs.



(b) Average L^2 distance between the vectors $\frac{\mathbf{N}^Z(t)}{t}$ and $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ during the learning process.

Figure 3: The results of different algorithms for an instance in Equation (8).

linear bandits on the same instance. This implies that the theoretical guarantees for linear bandit do not yield tight bounds in our setting, as we discuss below.

There is no sample complexity bound for BAI in linear bandits with sub-Gaussian rewards. However, if we approximate the sample complexity bound for the Gaussian case using the result of [Soa15], we obtain

$$T_1^*(\boldsymbol{\mu}, A)^{-1} \approx \sup_{\boldsymbol{w} \in \Delta^{n-1}} \min_{i \neq i^*} \frac{\Delta_i^2}{2(A_{i^*} - A_i)^\top \left(\sum_j w_j A_j A_j^\top \right)^{-1} (A_{i^*} - A_i)},$$

while for BAI with post-action context based on Equation (2), we have

$$T_2^*(\boldsymbol{\mu}, A)^{-1} = \sup_{\boldsymbol{w} \in \Delta^{n-1}} \min_{i \neq i^*} \frac{\Delta_i^2}{2(A_{i^*} - A_i)^\top D^{-1} (A_{i^*} - A_i)},$$

where $D = \sum_j w_j \mathrm{diag}(A_j)$. To show that $T_1^* \geq T_2^*$, it suffices to prove that

$$D - \sum_j w_j A_j A_j^\top = \sum_j w_j \left(\mathrm{diag}(A_j) - A_j A_j^\top \right)$$

is positive semi-definite. This follows from the Cauchy–Schwarz inequality

$$\forall v \in \mathbb{R}^k : \quad v^\top \mathrm{diag}(A_j) v = \sum_i v_i^2 A_{i,j} = \left(\sum_i v_i^2 A_{i,j} \right) \left(\sum_i A_{i,j} \right) \geq (v^\top A_j)^2 = v^\top A_j A_j^\top v.$$

Therefore, the linear bandit reduction necessarily increases the sample complexity.

To illustrate the difference numerically, consider the small instance

$$\boldsymbol{\mu} = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix}, \quad A = \begin{bmatrix} 0.9 & 0.2 & 0.5 & 0.1 \\ 0.05 & 0.4 & 0.4 & 0.8 \\ 0.05 & 0.4 & 0.1 & 0.1 \end{bmatrix}.$$

In this case, we compute $T_2^*(\boldsymbol{\mu}, A) \approx 165$, while the corresponding linear bandit lower bound is $T_1^*(\boldsymbol{\mu}, A) \approx 1054$, showing a substantial gap.

5 Experiments

This section presents a representative experiment for the separator setting, comparing our proposed algorithm with two benchmark algorithms. We use the Python CVXPY library [DB16, AVDB18] to solve the required optimization problems in all experiments. Additional experiments on synthetic instances and a semi-synthetic separator instance constructed from a real-world dataset, together with a more comprehensive discussion, are provided in Appendix I. The implementation is available at `https://github.com/ban-epfl/BAI-with-Post-Action-Context`.

We consider an instance with $n = 3$ arms and $k = 3$ context values, specified by the following parameters

$$\delta = 0.01, \quad \boldsymbol{\mu} = \begin{bmatrix} 1.0 \\ 0.1 \\ 0.3 \end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix} 0.9 & 0.9 & 0.1 \\ 0.09 & 0.01 & 0.45 \\ 0.01 & 0.09 & 0.45 \end{bmatrix}. \tag{8}$$

This instance is interesting because the first and second arms have very close average rewards, making it difficult to distinguish the best arm with high probability. To do so, one must observe a substantial number of samples from contexts 2 and 3, which have low probabilities of occurring when pulling arms 1 and 2. Consequently, an effective strategy involves frequently selecting arm 3, even though it has a notably lower expected reward than arms 1 and 2.

This behavior differs from classic BAI, where suboptimal arms are typically chosen less often. In instances like this, algorithms designed for standard BAI (e.g., track-and-stop) become suboptimal because they do not take full advantage of the post-action context information.

We compare Algorithm 1 (STS) with the following two baselines: (i) *Track-and-Stop (TS).* This classic algorithm ignores the context values and applies D-tracking to the actions. (ii) *Lazy Track-and-Stop (LTS).* It can be shown that the BAI problem with a separator post-action context can be reduced to a BAI problem in linear bandits, while the mean values of the reward distributions are bounded. For more details on this reduction, refer to Section 4.2. Consequently, identifying the best action translates into best-arm identification in a finite-arm linear bandit. We then apply Lazy Track-and-Stop [JP20], which is an optimal algorithm for BAI in linear bandits.

Figure 3a illustrates the average number of arm pulls for each algorithm on the instance in Equation (8), aggregated over 150 runs for each algorithm. Figure 3b shows the average L^2 distance of the vector $\frac{\mathbf{N}^z(t)}{t}$ from the optimal frequency of contexts $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ over time, which captures the convergence speed of each algorithm. As the figures demonstrate, ignoring the post-action contexts can lead to significant sub-optimality. Moreover, no incorrect recommendation was observed for any algorithm in any of our experiments, including those in Appendix I

6 Conclusion

We introduced a new BAI problem with post-action context in the fixed-confidence setting. We considered two settings for post-action context. For both settings, non-separator and separator context, we derived lower bounds on sample complexity and proposed algorithms that achieve asymptotic optimal sample complexity.

An interesting direction for future work is to consider the setting where context probability matrix $\mathcal{A}$ is unknown. In this case, as discussed in Appendix H, the optimization problem in Equation (1) becomes non-convex which renders designing a computationally efficient algorithm challenging.

7 Acknowledgement

We thank the anonymous reviewers for their feedback. This research was in part supported by the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40_180545 and Swiss SNF project 200021_204355 /1.

References

[AB10] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p, 2010.

[AMGP21] Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems*, 34:25852–25864, 2021.

[AMP21] Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes. In *International Conference on Machine Learning*, pages 7459–7468. PMLR, 2021.

[AMTK23] Aymen Al-Marjani, Andrea Tirinzoni, and Emilie Kaufmann. Towards instance-optimality in online pac reinforcement learning. *arXiv preprint arXiv:2311.05638*, 2023.

[Aue02] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[AVDB18] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[Ber63] Claude Berge. *Topological spaces: including a triatment of mltivalued functions, vector spaces and convexity*. Oliver and Boyd, 1963.

[BGK22] Antoine Barrier, Aurélien Garivier, and Tomáš Kocák. A non-asymptotic approach to best-arm identification for gaussian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 10078–10109. PMLR, 2022.

[BWR22] Blair Bilodeau, Linbo Wang, and Dan Roy. Adaptively exploiting d-separators with causal bandits. *Advances in Neural Information Processing Systems*, 35:20381–20392, 2022.

[BZ09] Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations research*, 57(6):1407–1420, 2009.

[CBJD24] Emil Carlsson, Debabrota Basu, Fredrik Johansson, and Devdatt Dubhashi. Pure exploration in bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 334–342. PMLR, 2024.

[CGL+17] Lijie Chen, Anupam Gupta, Jian Li, Mingda Qiao, and Ruosong Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Conference on Learning Theory*, pages 482–534. PMLR, 2017.

[CL16] Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.

[CWM+22] Romain Camilleri, Andrew Wagenmaker, Jamie H Morgenstern, Lalit Jain, and Kevin G Jamieson. Active learning with safety constraints. *Advances in Neural Information Processing Systems*, 35:33201–33214, 2022.

[DB16] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[DK19] Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

[DMSV20] Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

[ECBMR24] Khaled Eldowa, Nicolò Cesa-Bianchi, Alberto Maria Metelli, and Marcello Restelli. Information capacity regret bounds for bandits with mediator feedback. *arXiv preprint arXiv:2402.10282*, 2024.

[EMQ+23] Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk Van Der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: when do intermediate observations help? *arXiv preprint arXiv:2305.19036*, 2023.

[FJJR19] Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

[FN22] Fathima Zarin Faizal and Jayakrishnan Nair. Constrained pure exploration multi-armed bandits with a fixed budget. *arXiv preprint arXiv:2211.14768*, 2022.

[GC11] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[GGL12] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.

[GK16] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

[GKJ12] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

[HZW22] Wen Huang, Lu Zhang, and Xintao Wu. Achieving counterfactual fairness for causal bandit. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6952–6959, 2022.

[Jag13] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[JEK24] Fateme Jamshidi, Jalal Etesami, and Negar Kiyavash. Confounded budgeted causal bandits. In *Causal Learning and Reasoning*, pages 423–461. PMLR, 2024.

[JK19] Sandeep Juneja and Subhashini Krishnasamy. Sample complexity of partition identification using multi-armed bandits. In *Conference on Learning Theory*, pages 1824–1852. PMLR, 2019.

[JN14] Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th annual conference on information sciences and systems (CISS)*, pages 1–6. IEEE, 2014.

[JP20] Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

[KA21] Masahiro Kato and Kaito Ariu. The role of contextual information in best arm identification. *arXiv preprint arXiv:2106.14077*, 2021.

[Kau20] Emilie Kaufmann. *Contributions to the Optimal Solution of Several Bandit Problems*. PhD thesis, Université de Lille, 2020.

[KCG16] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

[KK13] Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251. PMLR, 2013.

[KK21] Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

[KKG18] Emilie Kaufmann, Wouter M Koolen, and Aurélien Garivier. Sequential test for the lowest mean: From thompson to murphy sampling. *Advances in Neural Information Processing Systems*, 31, 2018.

[KKS13] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*, pages 1238–1246. PMLR, 2013.

[KL03] Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.

[KTAS12] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[LAR24] Ziyi Liu, Idan Attias, and Daniel M Roy. Causal bandits: The pareto optimal frontier of adaptivity, a reduction to linear bandits, and limitations around unknown marginals. *arXiv preprint arXiv:2407.00950*, 2024.

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[LLR16] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.

[LMTY20] Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pages 141–150. PMLR, 2020.

[LRJ+22] Zhaoqi Li, Lillian Ratliff, Kevin G Jamieson, Lalit Jain, et al. Instance-optimal pac algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 35:37590–37603, 2022.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[LZ07] John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

[MDJ+21] Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

[Mén19] Pierre Ménard. Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*, 2019.

[MGG+19] Timothy Arthur Mann, Sven Gowal, Andras Gyorgy, Huiyi Hu, Ray Jiang, Balaji Lakshminarayanan, and Prav Srinivasan. Learning from delayed outcomes via proxies with applications to recommender systems. In *International Conference on Machine Learning*, pages 4324–4332. PMLR, 2019.

[MJTN20] Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all ϵ-good arms in stochastic bandits. *Advances in Neural Information Processing Systems*, 33:20707–20718, 2020.

[MMSB24] Rahul Madhavan, Aurghya Maiti, Gaurav Sinha, and Siddharth Barman. Causal contextual bandits with adaptive context. *arXiv preprint arXiv:2405.18626*, 2024.

[MPDR21] Alberto Maria Metelli, Matteo Papini, Pierluca D'Oro, and Marcello Restelli. Policy optimization as online learning with mediator feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8958–8966, 2021.

[MT04] Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

[MZG22] Avinash Mohan Mohammadi Zaki and Aditya Gopalan. Improved pure exploration in linear bandits with no-regret learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3709–3715, 2022.

[NWRJ24] Adhyyan Narang, Andrew Wagenmaker, Lillian Ratliff, and Kevin Jamieson. Sample complexity reduction via policy difference estimation in tabular reinforcement learning. *arXiv preprint arXiv:2406.06856*, 2024.

[PMR23] Riccardo Poiani, Alberto Maria Metelli, and Marcello Restelli. Pure exploration under mediators' feedback. *arXiv preprint arXiv:2308.15552*, 2023.

[RKB+21] Yoan Russac, Christina Katsimerou, Dennis Bohle, Olivier Cappé, Aurélien Garivier, and Wouter M Koolen. A/b/n testing with control in the presence of subpopulations. *Advances in Neural Information Processing Systems*, 34:25100–25110, 2021.

[RKKN23] Kota Srinivas Reddy, PN Karthik, Nikhil Karamchandani, and Jayakrishnan Nair. Best arm identification in bandits with limited precision sampling. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 1466–1471. IEEE, 2023.

[SEK26a] Mohammad Shahverdikondori, Jalal Etesami, and Negar Kiyavash. Active context selection improves simple regret in contextual bandits. *arXiv preprint arXiv:2605.20040*, 2026.

[SEK26b] Mohammad Shahverdikondori, Jalal Etesami, and Negar Kiyavash. Graph learning is suboptimal in causal bandits. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 3088–3096. PMLR, 2026.

[SLM14] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

[Soa15] Marta Soare. *Sequential resource allocation in linear stochastic bandits*. PhD thesis, Université Lille 1-Sciences et Technologies, 2015.

[SSZ+23] Zhongxiang Sun, Zihua Si, Xiaoxue Zang, Dewei Leng, Yanan Niu, Yang Song, Xiao Zhang, and Jun Xu. Kuaisar: A unified search and recommendation dataset. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 5407–5411, 2023.

[Ste23] Connor J Stephens. Pure exploration in multi-armed bandits. 2023.

[TD25] Adrienne Tuynman and Rémy Degenne. The batch complexity of bandit pure exploration. *arXiv preprint arXiv:2502.01425*, 2025.

[Tho33] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

[TJNN24] Dengwang Tang, Rahul Jain, Ashutosh Nayyar, and Pierluigi Nuzzo. Pure exploration for constrained best mixed arm identification with a fixed budget. *arXiv preprint arXiv:2405.15090*, 2024.

[TM17] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. *Mobile health: sensors, analytic methods, and applications*, pages 495–517, 2017.

[TNHK20] Koji Tabata, Atsuyoshi Nakamura, Junya Honda, and Tamiki Komatsuzaki. A bad arm existence checking problem: How to utilize asymmetric problem structure? *Machine learning*, 109(2):327–372, 2020.

[VSST23] Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Causal bandits for linear structural equation models. *Journal of Machine Learning Research*, 24(297):1–59, 2023.

[WBSJ21] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, pages 10686–10696. PMLR, 2021.

[WSJ22] Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, pages 358–418. PMLR, 2022.

[WTP21] Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.

[XC22] Nuoya Xiong and Wei Chen. Combinatorial pure exploration of causal bandits. *arXiv preprint arXiv:2206.07883*, 2022.

[XHS18] Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR, 2018.

[YMVT24] Zirui Yan, Arpan Mukherjee, Burak Varıcı, and Ali Tajer. Robust causal bandits for linear models. *IEEE Journal on Selected Areas in Information Theory*, 2024.

Appendix

The appendix is organized as follows. In Appendix A, we review the literature on related work. In Appendix B, we provide the results of non-separator setting. Appendix C contains the proof of the general lower bound. In Appendix D, we discuss the GLR test. Appendices E and F provide proofs for the non-separator and separator cases, respectively. Appendix G addresses how to solve the optimization problem in Equation (9), and Appendix H discusses the scenario where the context matrix is unknown. Finally, Appendix I provides additional details about our experiments and presents further numerical results.

Table 1: Table of notations.

Notation	Description
X	Action
Z	Post-action context
Y	Reward variables
n	Size of action set
k	Size of post-action context
$[n]$	$\{1, 2, \ldots, n\}$
Δ^{n-1}	$(n-1)$-dimensional standard simplex = $\{w \in \mathbb{R}^n \mid w_i \geq 0 \wedge \sum_{i=1}^n w_i = 1\}$
$d(P, Q)$	Kullback–Leibler (KL) divergence between two probability measures P and Q
$d_B(p, 1-p)$	KL between two Bernoulli with parameters p and $1-p$
$\text{conv}(\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n\})$	Convex hull of n vectors $\mathcal{A}_i \in \mathbb{R}^d = \left\{\sum_{i=1}^n \lambda_i \mathcal{A}_i \mid \lambda_i \geq 0 \ \wedge \sum_{i=1}^n \lambda_i = 1\right\}$
$\mathcal{A} = [\mathcal{A}_1 \mid \mathcal{A}_2 \mid \ldots \mid \mathcal{A}_n]$	Context probability matrix, where $\mathcal{A}_{j,i} = P(Z = j \mid X = i)$
$\mathcal{A}_i$	The i-th column of $\mathcal{A}$, i.e., the distribution of the context when arm i is played
$\boldsymbol{\mu} = [\boldsymbol{\mu}_1 \mid \boldsymbol{\mu}_2 \mid \ldots \mid \boldsymbol{\mu}_n]$	Mean value matrix, where $\mu_{j,i} = \mathbb{E}[Y \mid X = i, Z = j]$
$\boldsymbol{\mu}_i$	The i-th column of $\boldsymbol{\mu}$ (in the separator setting, $\boldsymbol{\mu}$ is a vector in $\mathbb{R}^k$)
$P_i^{\boldsymbol{\mu}}, P_i^{\mathcal{A},\boldsymbol{\mu}}$	Joint distribution of (Z, Y) after pulling arm $X = i$ in environment $\mathcal{A}, \boldsymbol{\mu}$
$i^*(\boldsymbol{\mu}), i^*(\boldsymbol{\mu}, \mathcal{A})$	Best arm for an instance parameterized by $\boldsymbol{\mu}$ and $\mathcal{A}$
i^*	Shorthand for $i^*(\boldsymbol{\mu}, \mathcal{A})$ when the best arm is used as an index, e.g., $\mathcal{A}_{j,i^*}$
$\mathcal{I}, \mathcal{I}(\mathcal{A})$	Set of $\boldsymbol{\mu}$ matrices that imply a unique best arm for a given $\mathcal{A}$
$\mathcal{I}^s, \mathcal{I}^s(\mathcal{A})$	Set of $\boldsymbol{\mu}$ vectors that imply a unique best arm for a given $\mathcal{A}$ in separator setting
Δ_i	Sub-optimality gap of arm i
$\hat{\Delta}_i(t)$	Estimated sub-optimality gap of arm i in round t
$\hat{\boldsymbol{\mu}}(t)$	Matrix containing the empirical estimates of all entries of $\boldsymbol{\mu}$ up to round t
$\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$	Set of alternative parameter matrices $\boldsymbol{\mu}' \in \mathcal{I}$ such that $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\mu}')$
$a_{\min}$	$\min_{i,j} \mathcal{A}_{i,j}$
$N_i^X(t)$	Total number of times arm i has been played up to round t
$N_j^Z(t)$	Total number of times context j has been observed up to round t
$N_{j,i}(t)$	Total number of times arm i has been played and context j has been observed up to round t
$\hat{\Lambda}(t)$	GLR statistic at around t
$\mathbf{w}$	Weights (proportions) of the arms, $\mathbf{w} \in \Delta^{n-1}$
$\mathbf{w}_z$	Expected frequencies of the contexts, $w_{z,j} = \sum_{l \in [n]} w_l \mathcal{A}_{j,l}$; $\mathbf{w}_z \in \text{conv}(\mathcal{A})$
$\mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})$	The optimal weights (proportions) for actions in non-separator context
$\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$	The set of optimal weights (proportions) for contexts

A Further Discussion on Related Work

Best Arm Identification. The literature on pure exploration in multi-armed bandit problems is extensive [LS20, Ste23]. Here, we provide a brief overview of the most closely related works. The standard best-arm identification problem is primarily studied in two settings: the fixed-confidence setting, where the probability of error is predetermined [GK16, AB10, Kau20, JN14, MT04], and the fixed-budget setting, where the time horizon is fixed [GGL12, KKS13, CL16].

[GK16] propose the first (asymptotically) optimal algorithm for the best-arm identification problem, known as track and stop, which attempts to track the optimal proportion of arm plays. The track-and-stop idea has since been widely adopted to design algorithms for various identification problems, including the algorithms presented in this paper.

In the fixed-confidence setting, other identification problems with different objectives have been explored in the literature, such as subset selection [KTAS12, KK13], checking for the existence of

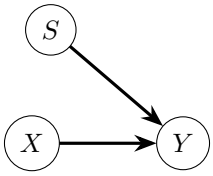


Figure 4: The structure of variables in [RKB$^+$21].

an arm with a low mean [TNHK20, KKG18], and finding all the good arms [MJTN20]. The most general form of the identification problem with bandit feedback has also been studied under various assumptions and is sometimes called *sequential identification* [KK21], *General-Samp* [CGL$^+$17], *partition identification* [JK19], and is typically encompassed by the term pure exploration [WTP21, DK19]. Our problem can be viewed as a general identification problem in which the agent cannot take arbitrary actions but can only choose probability distributions over them.

Best-arm identification in contextual bandits has also received considerable attention, with optimal algorithms and different constraints [LRJ$^+$22, KA21]. In these works, however, the context is assumed to be generated (independently) before taking an action, which differs from our setting.

In particular, [RKB$^+$21] introduce an additional observation, called a *sub-population*, and study four different modes. In the first mode, the *action mode*, the learner chooses both the action and the sub-population. In the other modes, however, the sub-population is sampled independently of the chosen action. The third mode, called the *agnostic mode*, is structurally related to the causal graph in Figure 4: the learner chooses an action and observes both the sub-population and the reward afterward. However, since the sub-population distribution is fixed and independent of the chosen action, this mode corresponds in our framework only to the special case where all columns of $\mathcal{A}$ are identical. Moreover, their objective is to identify the subset of actions that outperform a given control action. Therefore, their agnostic formulation only overlaps with our setting in the restricted two-action case with identical columns of $\mathcal{A}$. In contrast, our setting allows the post-action context distribution to depend on the chosen action and aims to identify a single best action. These differences make the two problems intrinsically distinct.

A more closely related line of research, recently explored, involves the bandit problem with mediator feedback. This setting is similar to our separator context setting but investigates different objectives [MPDR21, PMR23, RKKN23, ECBMR24].

In the setting of BAI with fixed-confidence, deriving a closed-form expression for GLR requires that the reward distributions belong to a one-parameter exponential family, which includes the Gaussian distribution with known variance. This assumption is commonly found in most existing works. However, recently [TD25] extend the GLR-based approach to the more general class of sub-Gaussian distributions. [TD25] show that by adjusting the threshold in the stopping rule, one can adapt the GLR expression originally designed for Gaussian distributions to achieve an appropriate stopping rule for sub-Gaussian distributions (for more details, please refer to Lemmas 3.1 and 3.2 in [TD25]). This result shows we can further relax our assumption from Gaussian distribution to sub-Gaussian distributions but for simplicity we consider Gaussian distributions in this paper.

Causal Bandits. Another line of research in bandits, where the agent receives additional information beyond the reward, is the field of *causal bandits*, first introduced by [LLR16] and subsequently explored in [MMSB24, JEK24, HZW22, XC22, SEK26b]. In causal bandit problems, a (known or unknown) causal graph is assumed among a set of variables, and the actions correspond to interventions on a subset of variables within the graph. After choosing each intervention (i.e., action), the agent observes the values of all nodes, with the goal of identifying interventions that maximize the expected value of a reward node.

In this paper, we study the best-arm identification problem for two particular graphs with three variables in a causal bandit setting. [BWR22, LAR24] introduce the notions of post-action context and *conditionally benign* property, which in absence of latent variables is identical to separator case in Section 3. However, their main focus is on cumulative regret minimization in the minimax setting, aiming to design algorithms that perform well in both conditionally benign and non-conditionally benign environments.

Linear Bandits. The separator setting of our problem can be reduced to a linear bandit problem through the following steps. We can disregard the context variable and assume that the reward value of action i is drawn from a distribution with mean $\mathcal{A}_i^\top \boldsymbol{\mu}_i$ and a zero-mean noise ϵ_i. Note that in this case, the mean of this new variable is the same as the mean reward for action i, and if the reward means are bounded, the noise becomes sub-Gaussian (for more details on this reduction, refer to Section 4.2).

In this setting, action i corresponds to a point $\mathcal{A}_i$ in $\mathbb{R}^k$, resulting in a best-arm identification problem in linear bandits with finitely many arms. This problem was first introduced by [SLM14] and has been widely studied. Optimal algorithms for best arm identification in linear bandits are provided in [JP20, DMSV20, MZG22]. [JP20] propose an algorithm inspired by track and stop, called lazy track and stop, while [DMSV20] present an algorithm based on game-theoretic intuition. For additional algorithms on best arm identification in linear bandits, see [XHS18, FJJR19]. The causal bandit with linear assumption on the causal model is also been studied recently [VSST23, YMVT24].

Best Policy Identification in RL. Another direction related to our problem is the best-policy identification problem in reinforcement learning, which has gained considerable attention in recent years [MDJ+21, AMGP21, WSJ22, AMTK23]. In best-policy identification, an agent interacts with an environment consisting of actions and states, and the goal is to select, from a given set of policies, the one that maximizes the expected reward. This problem generalizes best-arm identification in bandits, and many problems in the bandit literature can be viewed as special cases of this framework. For instance, in best-arm identification in contextual bandits, if we treat different contexts as distinct starting states, set the time horizon $H = 1$, and consider the actions as the set of policies, the algorithms for best-policy identification can solve the problem; for a detailed discussion, see Section 6 of [NWRJ24].

One might consider reducing a bandit with post-action contexts to an RL setting by treating post-action contexts as states. However, this approach does not apply to either the non-separator or the separator settings. The key difference lies in how the reward is determined: in RL, the reward depends on the previous action and the state from which the action is taken. In contrast, in our problem, the reward depends on the last action and the new state reached after the action (non-separator) or only on the new state (separator). This fundamental difference prevents a direct application of RL algorithms to solve our problem.

Pure exploration with safety constraints. The classic best-arm identification algorithm fails to address many problems in which constraints on action selection arise for critical reasons. Such constraints are often used to ensure safety in learning algorithms and can restrict different phases of learning—whether in exploration or recommendation. Furthermore, constraints can be known or unknown, depending on the application. Prior work explores best-arm identification under various settings [WBSJ21, TJNN24, FN22, CWM+22].

[CBJD24] also tackles a related problem, but with the key distinction that the goal is to find an optimal policy, a distribution over actions, rather than identifying a single arm. In fact, [CBJD24] is the closest work to ours, although they can choose any action (i.e., setting Z to a specific value in our setting) during exploration. This flexibility is not available in our setting, constituting the primary difference from their work.

B Non-Separator Context

In this section, we assume the context variable is a non-separator as depicted in Figure 1b, meaning the distribution of the reward variable can depend on both the action and the context. We first derive an instance-dependent lower bound on the number of samples required for any δ-correct algorithm and then propose an algorithm that asymptotically achieves this lower bound. The omitted proofs for this section appear in Appendix E.

B.1 Lower Bound

We establish a lower bound by explicitly solving the minimization problem defined in Equation (17).

Theorem B.1 (Non-Separator Lower Bound)**.** *Let $\delta \in (0, 1)$. Consider a bandit instance with a non-separator context and Gaussian reward distribution with unit variance, parameterized by matrices $\mathcal{A}$*

and $\boldsymbol{\mu}$. Then, any δ-correct algorithm with stopping time τ_δ satisfies $\mathbb{E}[\tau_\delta] \geq T^*(\boldsymbol{\mu}, \mathcal{A}) d_B(\delta, 1-\delta)$, where

$$T^*_{NS}(\boldsymbol{\mu}, \mathcal{A})^{-1} = \sup_{\mathbf{w} \in \Delta^{n-1}} \min_{i \neq i^*(\boldsymbol{\mu})} \frac{\Delta_i^2}{2} \left(\frac{w_{i^*(\boldsymbol{\mu})} w_i}{w_{i^*(\boldsymbol{\mu})} + w_i} \right). \tag{9}$$

To design the sampling rule for the algorithm, one must solve the optimization problem given in Equation (9). Solving this problem reduces to finding the root of a strictly decreasing function within a known interval, which can be efficiently done using methods such as binary search. For further details, refer to Appendix G.

B.2 Learning Algorithm

In this part, we introduce the components of our algorithm and show that it achieves optimal sample complexity.

Stopping Rule. Similar to the separator setting, we use a Generalized Likelihood Ratio (GLR) test for the stopping rule. The GLR statistic in this setting is defined as

$$\hat{\Lambda}(t) \triangleq \inf_{\boldsymbol{\mu}' \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \sum_{i \in [n], j \in [k]} N_{j,i}(t) \frac{(\hat{\boldsymbol{\mu}}_{j,i}(t) - \boldsymbol{\mu}'_{j,i})^2}{2}, \tag{10}$$

provided that $\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}(\mathcal{A})$ (i.e., $\hat{\boldsymbol{\mu}}(t)$ induces a unique best arm). Otherwise, $\hat{\Lambda}(t)$ is set to zero. We can simplify Equation (10) to

$$\hat{\Lambda}(t) = \min_{i \neq i^*(\hat{\boldsymbol{\mu}}(t))} \frac{\hat{\Delta}_i^2}{2\left(\sum_{j \in [k]} \frac{\mathcal{A}^2_{j, i^*(\hat{\boldsymbol{\mu}}(t))}}{N_{j, i^*(\hat{\boldsymbol{\mu}}(t))}(t)} + \frac{\mathcal{A}^2_{j,i}}{N_{j,i}(t)} \right)},$$

which allows $\hat{\Lambda}(t)$ to be computed efficiently. Appendix D.1 provides a detailed proof of this simplification.

To derive the sequential thresholds $\hat{c}_t(\delta)$, we use the deviation inequalities proposed by [KK21], based on *mixture martingales*. For each $\delta \in (0, 1]$ and $t \in \mathbb{N}$, the threshold $\hat{c}_t(\delta)$ is derived as

$$\hat{c}_t(\delta) = 4k \ln\left(4 + \ln\left(\frac{t}{2k}\right)\right) + 2kC^g\left(\frac{\ln\left(\frac{n-1}{\delta}\right)}{2k}\right), \tag{11}$$

where $C^g(x) \simeq x + \ln(x)$ (refer to Definition E.1). For more details on the derivation of the thresholds, see the proof of Lemma B.2 in Appendix E. The stopping time τ_δ is then defined as

$$\tau_\delta \triangleq \inf\{t \in \mathbb{N} \mid \hat{\Lambda}(t) > \hat{c}_t(\delta)\}. \tag{12}$$

At the stopping time τ_δ, the final suggestion $\hat{i}_\tau$ is equal to the unique estimated best arm $i^*(\hat{\boldsymbol{\mu}}(\tau_\delta))$. The following lemma establishes the correctness of this stopping rule when combined with any sampling rule.

Lemma B.2. *Consider a bandit instance with a non-separator context and Gaussian reward distribution with unit variance, parameterized by matrices $\mathcal{A}$ and $\boldsymbol{\mu}$. Any algorithm with the stopping rule* (12) *is δ-correct, that is, $P_{\boldsymbol{\mu}, \mathcal{A}}(\tau_\delta < \infty, \hat{i}_\tau \neq i^*(\boldsymbol{\mu})) \leq \delta$.*

Sampling Rule. We now propose a sampling strategy, which asymptotically achieves optimal sample complexity with the stopping rule in (12). Define $\mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})$ as the solution to the optimization problem

$$\operatorname*{argmax}_{\mathbf{w} \in \Delta^{n-1}} \min_{i \neq i^*(\boldsymbol{\mu})} \frac{\Delta_i^2}{2} \left(\frac{w_{i^*(\boldsymbol{\mu})} w_i}{w_{i^*(\boldsymbol{\mu})} + w_i} \right).$$

For any $\boldsymbol{\mu} \in \mathcal{I}$, the optimal weights $\mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})$ are unique and can be efficiently estimated (see Appendix G). In our theoretical analysis, we assume the existence of an oracle that computes $\mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})$ exactly. However, in practice, we estimate the optimal weights $\mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})$. These weights correspond to the fraction of times each arm is pulled by an optimal algorithm. Based on this intuition, various algorithms have been developed for different identification problems in the literature. These

algorithms compute the optimal weights at each step given the estimated reward matrix $\hat{\boldsymbol{\mu}}(t)$ and track these sequential weights. We adopt the *D-tracking* sampling rule first introduced by [GK16]. At round t of the algorithm, the D-tracking rule defines $\mathcal{U}_t = \{i \in [n] \mid N_i^X(t) \leq \max\left(\sqrt{t} - \frac{n}{2}, 0\right)\}$, which contains the arms that are under-sampled. The algorithm then selects the next action X_{t+1} as

$$X_{t+1} = \begin{cases} \operatorname{argmin}_{i \in \mathcal{U}_t} N_i^X(t) & \text{if } \mathcal{U}_t \neq \emptyset, \\ \operatorname{argmax}_{i \in [n]} tw_i^*(\hat{\boldsymbol{\mu}}(t), \mathcal{A}) - N_i^X(t) & \text{o.w.} \end{cases} \tag{13}$$

It is noteworthy that D-tracking operates correctly only when $\hat{\boldsymbol{\mu}} \in \mathcal{I}(\mathcal{A})$, that is, a single optimal arm exists. If this condition is not satisfied at any step t, the next action is selected uniformly random. This scenario almost surely ceases to occur as the number of samples increases and $\hat{\boldsymbol{\mu}}(t)$ converges to $\boldsymbol{\mu} \in \mathcal{I}(\mathcal{A})$, which is assumed to imply a single best arm.

The pseudocode of our algorithm, called *Non-Separator Track and Stop (NSTS)*, is presented in Algorithm 2. The following theorem proves its optimality.

Algorithm 2 Non-Separator Track and Stop (NSTS)

1: **Input:** Context probability matrix $\mathcal{A}$.
2: **Initialization:** Pull arms until collecting at least one sample from $P(Y|X = i, Z = j)$ for each $i \in [n], j \in [k]$, then set t to the number of rounds played.
3: **while** $\hat{\Lambda}(t) \leq \hat{c}_t(\delta)$ **do**
4: Pull arm X_{t+1} based on D-tracking rule (13).
5: Update $\hat{\boldsymbol{\mu}}(t)$ and $t \leftarrow t + 1$.
6: **end while**
7: **Output:** $i^*(\hat{\boldsymbol{\mu}}(t))$.

Theorem B.3 (Non-Separator Upper Bound)**.** *Algorithm 2 applied to a bandit instance with a non-separator context and Gaussian reward distribution with unit variance, parameterized by matrices $\mathcal{A}$ and $\boldsymbol{\mu}$ attains*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu},\mathcal{A}}[\tau_\delta]}{\ln(1/\delta)} \leq T_{NS}^*(\boldsymbol{\mu}, \mathcal{A}),$$

where $T^*(\boldsymbol{\mu}, \mathcal{A})$ is defined in Equation (9)*.*

B.3 Non-Optimality of Ignoring Non-Separator Context

By comparing the non-separator lower bound presented in Equation (9) with the lower bound for the BAI problem with unit-variance Gaussian rewards in [KCG16], we observe that the optimization problems characterizing the lower bounds are identical. At first glance, this may suggest that the post-action context is unnecessary. However, the key distinction lies in the reward distribution: when Z is ignored, the resulting reward distribution becomes a mixture of k unit-variance Gaussians, rather than a single Gaussian.

Since mixtures of Gaussians do not belong to the exponential family of distributions, algorithms developed for BAI within this family, like track and stop, are not directly applicable, and there are no theoretical guarantees for these distributions. One possible way is to exploit the fact that a mixture of Gaussians is sub-Gaussian. More precisely, the reward distribution for each action $X = i$ follows a sub-Gaussian distribution with $\sigma^2 = 1 + \frac{(b-a)^2}{4}$, where $b - a$ denotes the range of the reward means for $X = i$ and different context values, i.e., $\mu_{j,i} \in [a, b]$ for all $j \in [k]$ (by Lemma 4 of [RKB$^+$21]). For BAI with sub-Gaussian rewards of parameter σ^2, the sample complexity can be upper bounded by the sample complexity of BAI with σ^2-variance Gaussian rewards.

C Proof of General Lower bound: proposition 2.1

The proof employs the transportation lemma as outlined in Lemma 1 of [KCG16]. The proof of the transportation lemma, which can be found in Appendix A and A.1 of [KCG16], relies on changes of distributions and the log-likelihood ratio expression. Extending this theorem to our setup is quite

straightforward; we simply need to incorporate the context variables into the log-likelihood ratio formula, which means using

$$L_t = \sum_{a=1}^{n} \sum_{s=1}^{N_a^X(t)} \log\left(\frac{f_a(Z_{a,s}, Y_{a,s})}{f'_a(Z_{a,s}, Y_{a,s})}\right),$$

instead of

$$L_t = \sum_{a=1}^{n} \sum_{s=1}^{N_i^X(t)} \log\left(\frac{f_a(Y_{a,s})}{f'_a(Y_{a,s})}\right),$$

in Appendix A of [KCG16]. All the proofs of the transportation lemma remain unchanged. Therefore, based on the transportation lemma for our setting, for each $\boldsymbol{\mu}' \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})$, we have

$$\sum_{i=1}^{n} \mathbb{E}[N_i^X(\tau_\delta)] d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}) \geq \sup_{\mathcal{E} \in \mathcal{F}_{\tau_\delta}} d(P_{\boldsymbol{\mu}}(\mathcal{E}), P_{\boldsymbol{\mu}'}(\mathcal{E})).$$

Define event $\mathcal{E}' = \{\tau_\delta < \infty, \hat{i}_{\tau_\delta} \neq i^*(\boldsymbol{\mu})\}$, then $P_{\boldsymbol{\mu}}(\mathcal{E}') \geq 1 - \delta$ and $P_{\boldsymbol{\mu}'}(\mathcal{E}') \leq \delta$ according to δ-correctness property of the given algorithm. Therefore, based on the property of KL divergence, we have

$$\sup_{\mathcal{E} \in \mathcal{F}_{\tau_\delta}} d(P_{\boldsymbol{\mu}}(\mathcal{E}), P_{\boldsymbol{\mu}'}(\mathcal{E})) \geq d(P_{\boldsymbol{\mu}}(\mathcal{E}'), P_{\boldsymbol{\mu}'}(\mathcal{E}')) \geq d_B(\delta, 1-\delta),$$

which implies that

$$\sum_{i=1}^{n} \mathbb{E}[N_i^X(\tau_\delta)] d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}) \geq d_B(\delta, 1-\delta).$$

Therefore,

$$\inf_{\boldsymbol{\mu}' \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{i=1}^{n} \mathbb{E}[N_i^X(\tau_\delta)] d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}) \geq d_B(\delta, 1-\delta).$$

Then, we have

$$\begin{aligned}
\inf_{\boldsymbol{\mu}' \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{i=1}^{n} \mathbb{E}[N_i^X(\tau_\delta)] d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}) &= \mathbb{E}[\tau_\delta] \inf_{\boldsymbol{\mu}' \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{i=1}^{n} \frac{\mathbb{E}[N_i^X(\tau_\delta)]}{\mathbb{E}[\tau_\delta]} d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}) \\
&\leq \mathbb{E}[\tau_\delta] \sup_{\mathbf{w} \in \Delta^{n-1}} \inf_{\boldsymbol{\mu}' \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{i \in [n]} w_i d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}).
\end{aligned}$$

Finally,

$$\mathbb{E}[\tau_\delta] \geq \left(\sup_{\mathbf{w} \in \Delta^{n-1}} \inf_{\boldsymbol{\mu}' \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{i=1}^{n} w_i d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'})\right)^{-1} d_B(\delta, 1-\delta)$$

which concludes the proof.

D Generalized Likelihood Ratio Test

The Generalized Likelihood Ratio (GLR) is used to measure our confidence in a hypothesis. Let Ω_0 and Ω_1 be two model spaces. We wish to know whether a model λ belongs to Ω_0 or Ω_1. To do this, we can define two hypotheses: (i) $H_0 : (\lambda \in \Omega_0)$ and (ii) $H_1 : (\lambda \in \Omega_1)$. The log-likelihood ratio of this test after t observations is defined as

$$\hat{\Lambda}(t) \triangleq \log \frac{\sup_{\lambda \in \Omega_0 \cup \Omega_1} \mathcal{L}(X_1, X_2, \ldots, X_t; \lambda)}{\sup_{\lambda \in \Omega_0} \mathcal{L}(X_1, X_2, \ldots, X_t; \lambda)}, \tag{14}$$

where $\mathcal{L}(X_1, X_2, \ldots, X_t; \lambda)$ is the likelihood function of observations (i.e., $X_1, X_2, \ldots, X_t$) associated to parameter λ. If $\hat{\Lambda}(t)$ has a high value, we could reject the hypothesis $\mathcal{H}_0$.

Our problem in both settings can be viewed as a *general identification* [KK21]. In this problem, we assume $\mathcal{I}$ is partitioned into n subsets $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_n$, and the goal is to determine which subset contains vector $\boldsymbol{\mu}$. At round t, we have an estimate $\hat{\boldsymbol{\mu}}(t)$ and we need to decide whether we can

confidently choose the correct subset with low risk. Refer to Section 4 of [KK21] for more details. In our problem, we can define the subsets

$$\mathcal{O}_i \triangleq \left\{ \boldsymbol{\mu} \in \mathcal{I} \,\middle|\, \mathcal{A}_i^\top \boldsymbol{\mu}_i > \mathcal{A}_j^\top \boldsymbol{\mu}_j,\ \forall j \neq i \right\}.$$

Suppose that at round t of our problem, $i^*(\hat{\boldsymbol{\mu}}(t))$ is not unique, meaning $\hat{\boldsymbol{\mu}}(t)$ is not in $\mathcal{I}$. In this case, we lack sufficient information to determine the best arm (or, find the correct subset in general identification problem), so we must continue sampling. If $i^*(\hat{\boldsymbol{\mu}}(t))$ indicates a unique arm, however, we need a measure to evaluate our confidence. We consider the following hypothesis test.

$$\mathcal{H}_0 : (i^*(\boldsymbol{\mu}) \neq i^*(\hat{\boldsymbol{\mu}}(t))),$$
$$\mathcal{H}_1 : (i^*(\boldsymbol{\mu}) = i^*(\hat{\boldsymbol{\mu}}(t))),$$

which is equivalent to

$$\mathcal{H}_0 : (\boldsymbol{\mu} \notin \mathcal{O}_{i^*(\hat{\boldsymbol{\mu}}(t))}) = (\boldsymbol{\mu} \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})),$$
$$\mathcal{H}_1 : (\boldsymbol{\mu} \in \mathcal{O}_{i^*(\hat{\boldsymbol{\mu}}(t))}).$$

Hence, if $\hat{\Lambda}(t)$ is high, we can reject the $\mathcal{H}_0$, indicating with high confidence that $i^*(\hat{\boldsymbol{\mu}}(t))$ is indeed the best arm

Note that at round t, the observed samples are $((Z_1, Y_1), \ldots, (Z_t, Y_t))$ for the chosen actions $(i_1, \ldots, i_t)$. Consequently,

$$\hat{\Lambda}(t) = \log \frac{\sup_{\boldsymbol{\mu} \in I} \mathcal{L}((Z_1, Y_1), \ldots, (Z_t, Y_t); \boldsymbol{\mu})}{\sup_{\boldsymbol{\mu} \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \mathcal{L}((Z_1, Y_1), \ldots, (Z_t, Y_t); \boldsymbol{\mu})},$$

which expands to

$$\begin{aligned}
&\log \frac{\sup_{\boldsymbol{\mu}' \in I} \prod_{s \in [t]} P(Z = Z_s \mid X = i_s) P_{\boldsymbol{\mu}'}(Y = Y_s \mid X = i_s, Z = Z_s)}{\sup_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \prod_{s \in [t]} P(Z = Z_s \mid X = i_s) P_{\boldsymbol{\mu}'}(Y = Y_s \mid X = i_s, Z = Z_s)} \\
=&\log \frac{\sup_{\boldsymbol{\mu}' \in I} \prod_{s \in [t]} \mathcal{A}_{Z_s, i_s} P_{\boldsymbol{\mu}'}(Y = Y_s \mid X = i_s, Z = Z_s)}{\sup_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \prod_{s \in [t]} \mathcal{A}_{Z_s, i_s} P_{\boldsymbol{\mu}'}(Y = Y_s \mid X = i_s, Z = Z_s)} \\
=&\log \frac{\sup_{\boldsymbol{\mu}' \in I} \prod_{s \in [t]} P_{\boldsymbol{\mu}'}(Y = Y_s \mid X = i_s, Z = Z_s)}{\sup_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \prod_{s \in [t]} P_{\boldsymbol{\mu}'}(Y = Y_s \mid X = i_s, Z = Z_s)}
\end{aligned} \tag{15}$$

In the following two parts, for each case of context, we derive a simpler expression for $\hat{\Lambda}(t)$ under the non-separator and separator context.

D.1 Non-separator context

Note that in this case, $P_{\boldsymbol{\mu}}(Y = Y_s \mid X = i_s, Z = Z_s)$ is not necessarily equal to $P_{\boldsymbol{\mu}}(Y = Y_s \mid Z = Z_s)$. By assumption, $Y \mid X = i_s, Z = Z_s$ follows a Gaussian distribution with unit variance, which is a one-parameter exponential family. For this case, we can simplify Equation (15) to

$$\begin{aligned}
\hat{\Lambda}(t) &= \inf_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \sum_{i \in [n], j \in [k]} N_{j,i}(t) d(\mathcal{N}(\hat{\boldsymbol{\mu}}_{j,i}(t), 1), \mathcal{N}(\boldsymbol{\mu}'_{j,i}, 1)) \\
&= \inf_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \sum_{i \in [n], j \in [k]} N_{j,i}(t) \frac{(\hat{\boldsymbol{\mu}}_{j,i}(t) - \boldsymbol{\mu}'_{j,i})^2}{2}
\end{aligned}$$

where $\mathcal{N}(x, 1)$ denotes Gaussian distribution with mean x and unit variance and we used the fact $d(\mathcal{N}(x, 1), \mathcal{N}(y, 1)) = \frac{(x-y)^2}{2}$.

Computation of $\hat{\Lambda}(t)$. Recall that $\text{Alt}(\boldsymbol{\mu}, \mathcal{A}) = \cup_{i \neq i^*(\boldsymbol{\mu})} \mathcal{C}_i$, where $\mathcal{C}_i = \{\boldsymbol{\mu}' \in \mathcal{I} \mid \mathcal{A}_i^\top \boldsymbol{\mu}'_i > \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'_{i^*}\}$. Thus, we can rewrite $\hat{\Lambda}(t)$ as

$$\hat{\Lambda}(t) = \min_{s \neq i^*(\hat{\boldsymbol{\mu}}(t))} \inf_{\boldsymbol{\mu}' \in \mathcal{C}_s} \sum_{i \in [n], j \in [k]} N_{j,i}(t) \frac{(\hat{\boldsymbol{\mu}}_{j,i}(t) - \boldsymbol{\mu}'_{j,i})^2}{2},$$

which is equal to

$$\inf_{\boldsymbol{\mu}' \in \mathcal{I}} \sum_{i \in [n], j \in [k]} N_{j,i}(t) \frac{(\hat{\boldsymbol{\mu}}_{j,i}(t) - \boldsymbol{\mu}'_{j,i})^2}{2}$$
$$\text{s.t.} \quad \mathcal{A}_s^\top \boldsymbol{\mu}'_s > \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'_{i^*}.$$

Note that $N_{j,i}(t) > 0$ for all $i \in [n]$ and $j \in [k]$ due to initialization phase. $\mathcal{I}$ is dense in $\mathbb{R}^{k \times n}$. Hence, we can replace the constraint $\boldsymbol{\mu}' \in \mathcal{I}$ with $\boldsymbol{\mu}' \in \mathbb{R}^{k \times n}$, thus making

$$\inf_{\boldsymbol{\mu}' \in \mathbb{R}^{k \times n}} \sum_{i \in [n], j \in [k]} N_{j,i}(t) \frac{(\hat{\boldsymbol{\mu}}_{j,i}(t) - \boldsymbol{\mu}'_{j,i})^2}{2}$$
$$\text{s.t.} \quad \mathcal{A}_s^\top \boldsymbol{\mu}'_s > \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'_{i^*}.$$

Because the objective function is continuous and we are taking infimum, we can replace the strict inequality constraint with a non-strict one

$$\inf_{\boldsymbol{\mu}' \in \mathbb{R}^{k \times n}} \sum_{i \in [n], j \in [k]} N_{j,i}(t) \frac{(\hat{\boldsymbol{\mu}}_{j,i}(t) - \boldsymbol{\mu}'_{j,i})^2}{2}$$
$$\text{s.t.} \quad \mathcal{A}_s^\top \boldsymbol{\mu}'_s \geq \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'_{i^*}.$$

We solve this using the Lagrangian

$$L(\boldsymbol{\mu}', \lambda) = \sum_{i \in [n], j \in [k]} N_{j,i}(t) \frac{(\hat{\boldsymbol{\mu}}_{j,i}(t) - \boldsymbol{\mu}'_{j,i})^2}{2} + \lambda(\mathcal{A}_{i^*}^\top \boldsymbol{\mu}'_{i^*} - \mathcal{A}_s^\top \boldsymbol{\mu}'_s). \tag{16}$$

Taking derivatives and simplifying, we obtain

$$\boldsymbol{\mu}'_{j.i} = \begin{cases} \hat{\boldsymbol{\mu}}_{j,i^*}(t) - \frac{\lambda \mathcal{A}_{j,i^*}}{N_{j,i^*}(t)} & i = i^*(\boldsymbol{\mu}), \\ \hat{\boldsymbol{\mu}}_{j,s}(t) + \frac{\lambda \mathcal{A}_{j,s}}{N_{j,s}(t)} & i = s, \\ \hat{\boldsymbol{\mu}}_{j,i}(t) & \text{o.w.} \end{cases}$$

Substituting $\boldsymbol{\mu}'_{j,i}$ back into (16) yields

$$\lambda^* = \frac{\hat{\Delta}_s}{\left(\sum_{j \in [k]} \frac{\mathcal{A}^2_{j,i^*(\hat{\boldsymbol{\mu}}(t))}}{N_{j,i^*(\hat{\boldsymbol{\mu}}(t))}(t)} + \frac{\mathcal{A}^2_{j,s}}{N_{j,s}(t)}\right)}.$$

Hence,

$$\inf_{\boldsymbol{\mu}' \in \mathcal{C}_s} \sum_{i \in [n], j \in [k]} N_{j,i}(t) \frac{(\hat{\boldsymbol{\mu}}_{j,i}(t) - \boldsymbol{\mu}'_{j,i})^2}{2} = \frac{\hat{\Delta}_s^2}{2\left(\sum_{j \in [k]} \frac{\mathcal{A}^2_{j,i^*(\hat{\boldsymbol{\mu}}(t))}}{N_{j,i^*(\hat{\boldsymbol{\mu}}(t))}(t)} + \frac{\mathcal{A}^2_{j,s}}{N_{j,s}(t)}\right)},$$

Finally, the simplified form of $\hat{\Lambda}(t)$ becomes

$$\hat{\Lambda}(t) = \min_{s \neq i^*(\hat{\boldsymbol{\mu}}(t))} \frac{\hat{\Delta}_s^2}{2\left(\sum_{j \in [k]} \frac{\mathcal{A}^2_{j,i^*(\hat{\boldsymbol{\mu}}(t))}}{N_{j,i^*(\hat{\boldsymbol{\mu}}(t))}(t)} + \frac{\mathcal{A}^2_{j,s}}{N_{j,s}(t)}\right)},$$

D.2 Separator context

In the separator case, $P_{\boldsymbol{\mu}}(Y = Y_s \mid X = i_s, Z = Z_s) = P_{\boldsymbol{\mu}}(Y = Y_s \mid Z = Z_s)$ holds. Thus,

$$\hat{\Lambda}(t) = \log \frac{\sup_{\boldsymbol{\mu}' \in I} \prod_{s \in [t]} P_{\boldsymbol{\mu}'}(Y = Y_s \mid Z = Z_s)}{\sup_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \prod_{s \in [t]} P_{\boldsymbol{\mu}'}(Y = Y_s \mid Z = Z_s)}$$

In Section 3, we assume $\boldsymbol{\mu}$ is a vector in $\mathbb{R}^k$, where $\boldsymbol{\mu}_i = \mathbb{E}[Y \mid Z = i]$. Similar to the non-separator setting, because $Y \mid Z = i_s$ is a Gaussian variable with unit variance, then obtain

$$\begin{aligned}\hat{\Lambda}(t) &= \inf_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \sum_{j \in [k]} N_j^Z(t) d(\mathcal{N}(\hat{\boldsymbol{\mu}}_j(t), 1), \mathcal{N}(\boldsymbol{\mu}'_j, 1)) \\ &= \inf_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A})} \sum_{j \in [k]} N_j^Z(t) \frac{(\hat{\boldsymbol{\mu}}_j(t) - \boldsymbol{\mu}'_j)^2}{2}.\end{aligned}$$

Computation of $\hat{\Lambda}(t)$. Let $\mathcal{I}^s(\mathcal{A})$ be the set of vectors in $\mathbb{R}^k$ that ensure a unique best arm in the separator setting with context probability matrix $\mathcal{A}$. The computation here is similar to that in the non-separator setting. Another expression for $\hat{\Lambda}(t)$ is

$$\hat{\Lambda}(t) = \min_{s \neq i^*(\hat{\boldsymbol{\mu}}(t))} \inf_{\boldsymbol{\mu}' \in \mathcal{C}_s} \sum_{j \in [k]} N_j^Z(t) \frac{(\hat{\boldsymbol{\mu}}_j(t) - \boldsymbol{\mu}'_j)^2}{2}.$$

where, in this case, $\mathcal{C}_s = \{\boldsymbol{\mu}' \in \mathcal{I}^s \mid \mathcal{A}_s^\top \boldsymbol{\mu}' > \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'\}$. Since $\mathcal{I}^s$ is dense in $\mathbb{R}^k$ and the objective function is continuous, the inner term is equivalent to

$$\begin{aligned}&\inf_{\boldsymbol{\mu}' \in \mathbb{R}^k} \sum_{j \in [k]} N_j^Z(t) \frac{(\hat{\boldsymbol{\mu}}_j(t) - \boldsymbol{\mu}'_j)^2}{2} \\ &\text{s.t.} \quad \mathcal{A}_s^\top \boldsymbol{\mu}' > \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'.\end{aligned}$$

Because of the initialization phase, $N_j^Z(t) > 0$ for all $j \in [k]$. We solve this using the Lagrangian method

$$L(\boldsymbol{\mu}', \lambda) = \sum_{j \in [k]} N_j^Z(t) \frac{(\hat{\boldsymbol{\mu}}_j(t) - \boldsymbol{\mu}'_j)^2}{2} + \lambda(\mathcal{A}_{i^*}^\top \boldsymbol{\mu}' - \mathcal{A}_s^\top \boldsymbol{\mu}').$$

After differentiation and some algebra, we obtain

$$\begin{aligned}\boldsymbol{\mu}_j^* &= \hat{\boldsymbol{\mu}}_j(t) - \frac{\lambda^*(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,s})}{N_j(t)} \\ \lambda^* &= \frac{\hat{\Delta}_s}{\sum_{j \in [k]} \frac{(\mathcal{A}_{j,i^*(\hat{\boldsymbol{\mu}}(t))} - \mathcal{A}_{j,s})^2}{N_j^Z(t)}}.\end{aligned}$$

Thus,

$$\hat{\Lambda}(t) = \min_{s \neq i^*(\hat{\boldsymbol{\mu}}(t))} \frac{\hat{\Delta}_s^2}{2 \sum_{j \in [k]} \frac{(\mathcal{A}_{j,i^*(\hat{\boldsymbol{\mu}}(t))} - \mathcal{A}_{j,s})^2}{N_j^Z(t)}}.$$

E Proofs of Section B

The definition of C^g from [KK21] is as follows.

Definition E.1. *The function C^g is defined as*

$$C^g(x) := \min_{\lambda \in (\frac{1}{2}, 1]} \frac{g(\lambda) + x}{\lambda},$$

where function $g : \left(\frac{1}{2}, 1\right] \to \mathbb{R}$ is given by

$$g(\lambda) = 2\lambda - 2\lambda \ln(4\lambda) + \ln(\zeta(2\lambda)) - \frac{1}{2}\ln(1 - \lambda),$$

and ζ denotes the Riemann zeta function, defined as $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$.

E.1 Proof of Theorem B.1

For a given vector $\mathbf{w} \in \mathbb{R}^n$, define

$$f_j(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) \triangleq \inf_{\boldsymbol{\mu}' \in \mathcal{C}_j} \sum_{i \in [n]} w_i d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}). \tag{17}$$

Equation (1) can then be reformulated as

$$T^*(\boldsymbol{\mu}, \mathcal{A})^{-1} = \sup_{\mathbf{w} \in \Delta^{n-1}} \min_{i \neq i^*(\boldsymbol{\mu})} f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}). \tag{18}$$

We prove Theorem B.1 by explicitly solving the optimization problem (17) for the non-separator setting and incorporating the solution into the general lower bound formula (18).

Lemma E.2. *Consider a bandit instance with a non-separator context and Gaussian reward distribution with unit variance, parameterized by matrices $\boldsymbol{\mu}$ and $\mathcal{A}$. For any weight vector $\mathbf{w}$ such that $\forall s \in [n] : w_s > 0$, the value of $f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) = \inf_{\boldsymbol{\mu}' \in \mathcal{C}_i} \sum_{s \in [n]} w_s d(P_s^{\boldsymbol{\mu}}, P_s^{\boldsymbol{\mu}'})$ can be explicitly determined as*

$$f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) = \frac{\Delta_i^2}{2} \left(\frac{w_{i^*(\boldsymbol{\mu})} w_i}{w_{i^*(\boldsymbol{\mu})} + w_i} \right).$$

Proof. First note that $P_s^{\boldsymbol{\mu}}$ and $P_s^{\boldsymbol{\mu}'}$ are the distribution of random vectors (Z, Y) and (Z', Y') corresponding to bandit instances with context probability matrix $\mathcal{A}$ and means matrix $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, respectively. Since reward distributions are Gaussian with unit variance, the KL divergence can be written as

$$\begin{aligned} d(P_s^{\boldsymbol{\mu}}, P_s^{\boldsymbol{\mu}'}) &= \sum_{j \in [k]} P_s^{\boldsymbol{\mu}}(Z = j) d(P_s^{\boldsymbol{\mu}}(Y \mid Z = j), P_s^{\boldsymbol{\mu}'}(Y \mid Z = j)) \\ &= \sum_{j \in [k]} \mathcal{A}_{j,s} d(\mathcal{N}(\mu_{j,s}, 1), \mathcal{N}(\mu'_{j,s}, 1)) = \sum_{j \in [k]} \mathcal{A}_{j,s} \frac{(\mu_{j,s} - \mu'_{j,s})^2}{2}. \end{aligned}$$

Recall the definition

$$\mathcal{C}_i = \left\{ \boldsymbol{\mu}' \in \mathcal{I} \middle| \mathcal{A}_i^\top \boldsymbol{\mu}'_i > \mathcal{A}_{i^*(\boldsymbol{\mu})} \boldsymbol{\mu}'_{i^*(\boldsymbol{\mu})} \right\}.$$

Note that $\sum_{s \in [n]} w_s d(P_s^{\boldsymbol{\mu}}, P_s^{\boldsymbol{\mu}'})$ is continuous with respect to $\boldsymbol{\mu}$ and $\mathcal{I}$ is dense in $\mathbb{R}^{k \times n}$. Therefore, we can express $f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A})$ as the solution to the following optimization problem

$$\begin{aligned} f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) = \inf_{\boldsymbol{\mu}' \in \mathbb{R}^{k \times n}} & \sum_{s \in [n]} w_s \sum_{j \in [k]} \mathcal{A}_{j,s} \frac{(\mu_{j,s} - \mu'_{j,s})^2}{2} \\ \text{s.t.} \quad & \mathcal{A}_i^\top \boldsymbol{\mu}'_i > \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'_{i^*}. \end{aligned}$$

The solution to this problem is equivalent to the solution to the following problem when $\alpha \to 0$,

$$\begin{aligned} \inf_{\boldsymbol{\mu}' \in \mathbb{R}^{k \times n}} & \sum_{s \in [n]} w_s \sum_{j \in [k]} \mathcal{A}_{j,s} \frac{(\mu_{j,s} - \mu'_{j,s})^2}{2} \\ \text{s.t.} \quad & \mathcal{A}_i^\top \boldsymbol{\mu}'_i \geq \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'_{i^*} + \alpha. \end{aligned}$$

To solve the second problem, we write the Lagrangian as

$$L(\boldsymbol{\mu}', \lambda) = \sum_{s \in [n]} w_s \sum_{j \in [k]} \mathcal{A}_{j,s} \frac{(\mu_{j,s} - \mu'_{j,s})^2}{2} + \lambda(\mathcal{A}_{i^*}^\top \boldsymbol{\mu}'_{i^*} - \mathcal{A}_i^\top \boldsymbol{\mu}'_i + \alpha).$$

By calculating the derivative of L with respect to each variable and performing some algebra, we derive the optimal values for $\boldsymbol{\mu}'^*, \lambda^*$:

$$\boldsymbol{\mu}'^*_{j,t} = \begin{cases} \mu_{j,i^*} - \frac{\lambda}{w_{i^*}} & t = i^*(\boldsymbol{\mu}), \\ \mu_{j,i} + \frac{\lambda}{w_i} & t = i, \\ \mu_{j,t} & o.w. \end{cases}$$

$$\lambda^* = \frac{\Delta_i + \alpha}{\left(\frac{w_{i^*} + w_i}{w_{i^*} w_i}\right)}.$$

By inserting these values into the objective function, we have

$$f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) = \lim_{\alpha \to 0} \frac{(\Delta_i + \alpha)^2}{2} \left(\frac{w_{i^*} w_i}{w_{i^*} + w_i}\right) = \frac{\Delta_i^2}{2} \left(\frac{w_{i^*} w_i}{w_{i^*} + w_i}\right),$$

which completes the proof. □

For a given $\mathbf{w} \in \Delta^{n-1}$, if $w_i = 0$ for some $i \in [n]$, then we can find $\boldsymbol{\mu}'$ such that $P(P_s^{\boldsymbol{\mu}'}) = P(P_s^{\boldsymbol{\mu}})$ for all $s \in [k]$ and also $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\mu})$. Therefore, the term $\inf_{\boldsymbol{\mu}' \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{i \in [n]} w_i d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'})$ is equal to zero. Since we want to find $\mathbf{w} \in \Delta^{n-1}$ that maximizes

$$\inf_{\boldsymbol{\mu}' \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{i \in [n]} w_i d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}),$$

we must restrict ourselves to $\mathbf{w}$ with strictly positive entries. By combining Lemma E.2 and Equation (18), we conclude Theorem B.1.

E.2 Proof of Lemma B.2

Lemma B.2. *Consider a bandit instance with a non-separator context and Gaussian reward distribution with unit variance, parameterized by matrices $\mathcal{A}$ and $\boldsymbol{\mu}$. Any algorithm with the stopping rule* (12) *is δ-correct, that is, $P_{\boldsymbol{\mu}, \mathcal{A}}(\tau_\delta < \infty, \hat{i}_\tau \neq i^*(\boldsymbol{\mu})) \leq \delta$.*

Proof. As discussed in Appendix D, our problem can be viewed as a general identification, in which $\mathcal{I}$ is partitioned into M subsets $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_M$. The goal is to determine which subset contains vector $\boldsymbol{\mu}$. [KK21] provide a δ-correct stopping rule for general identification based on the *rank* of problem, defined as follows.

Definition E.3 (Rank)**.** *Consider a general identification problem specified by a partition $\mathcal{O} = \bigcup_{i=1}^M \mathcal{O}_i$. We say this problem has rank R if for every $i \in \{1, \ldots, M\}$, we can write*

$$\mathcal{O} \setminus \mathcal{O}_i = \bigcup_{q \in [Q]} \left\{ \lambda \in \mathcal{I} \;\middle|\; (\lambda_{k_1^{i,q}}, \ldots, \lambda_{k_R^{i,q}}) \in \mathcal{L}_{i,q} \right\},$$

for a family of arm indices $k_r^{i,q} \in [K]$ and open sets $\mathcal{L}_{i,q}$ indexed by $r \in [R]$, $q \in [Q]$ and $i \in [M]$. In other words, the rank is R if every set $\mathcal{O} \setminus \mathcal{O}_i$ is a finite union of sets that are each defined in terms of only R arms.

Based on the rank of a general identification problem, [KK21] provide the following theorem, which shows that the GLR stopping rule is δ-correct.

Theorem E.4 ([KK21])**.** *For any identification problem of rank R with M partitions and Gaussian reward distributions with unit variance, the GLR stopping rule 12 is δ-correct with threshold*

$$\hat{c}_t(\delta) = 2R \ln\left(4 + \ln\left(\frac{t}{R}\right)\right) + RC^g\left(\frac{\ln\left(\frac{M-1}{\delta}\right)}{R}\right).$$

where C^g is defined in Definition E.1.

Next, we show that the rank of the best arm identification problem in a bandit with a non-separator context is $2k$. The partitions in this problem are given by

$$\mathcal{O}_i = \left\{ \boldsymbol{\mu} \in \mathcal{I} \,\middle|\, \mathcal{A}_i^\top \boldsymbol{\mu}_i > \mathcal{A}_j^\top \boldsymbol{\mu}_j,\ \forall j \neq i \right\},$$

for all $i \in [n]$. It follows that

$$\mathcal{I} \setminus \mathcal{O}_i = \bigcup_{j \neq i} \{ \boldsymbol{\mu} \in \mathcal{I} \big| (\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) \in \mathcal{L}_{i,j} \},$$

where $\mathcal{L}_{i,j} = \{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) \in \mathbb{R}^{2k} \big| \mathcal{A}_i^\top \boldsymbol{\mu}_1 < \mathcal{A}_j^\top \boldsymbol{\mu}_2\}$. This shows that our problem has rank $R = 2k$. By applying Theorem E.4 with $R = 2k$ and $M = n$ to our problem, we complete the proof. □

E.3 Proof of Theorem B.3

Proof. At the beginning of the proof, we provide some important lemmas that we need.

Lemma E.5. *There exist constants D, E, such that the sequential thresholds $\hat{c}_t(\delta)$ defined in* (11)*, have the following property*

$$\forall t > E, \forall \delta \in (0, 1] : \hat{c}_t(\delta) \leq \ln\left(\frac{Dt}{\delta}\right)$$

Proof. Note that, based on the Definition of E.1, we have

$$C^g(x) = \min_{\lambda \in (0.5,1]} \frac{x + g(\lambda)}{\lambda}$$

Substituting the definition of g gives

$$\begin{aligned} C^g(x) &= \min_{\lambda \in (0.5,1]} \frac{x + 2\lambda - 2\lambda \ln(4\lambda) + \ln(\zeta(2\lambda)) - \frac{1}{2}\ln(1-\lambda)}{\lambda} \\ &= \min_{\lambda \in (0.5,1]} \frac{x - \frac{1}{2}\ln(1-\lambda)}{\lambda} + \frac{2\lambda - 2\lambda\ln(4\lambda) + \ln(\zeta(2\lambda))}{\lambda}. \end{aligned}$$

Now, for large value of x, if we set $\lambda = 1 - \frac{1}{x}$, then $\zeta(2\lambda)$ is upper bounded by a constant and therefore, we have

$$C^g(x) \leq c' + \frac{x + \frac{1}{2}\ln(x)}{1 - \frac{1}{x}}.$$

where c' is a constant value. Since x is large, we have

$$\frac{x + \frac{1}{2}\ln(x)}{1 - \frac{1}{x}} \leq x + \ln(x).$$

Consequently, $C^g(x)$ can be upper bounded by

$$C^g(x) \leq x + \ln(x) + c'$$

for sufficiently large values of x.

For small values of x, we can bound by setting $\lambda = 0.75$,

$$C^g(x) \leq \frac{4}{3}(x + g(0.75)) = \frac{4}{3}x + c''.$$

Therefore, by choosing an appropriate constant c, we obtain

$$C^g(x) \leq x + \ln(x) + c.$$

Returning to the expression for $\hat{c}_t(\delta)$, we have

$$\begin{aligned} \hat{c}_t(\delta) &= (4k)\ln\left(4 + \ln\left(\frac{t}{2k}\right)\right) + (2k)C^g\left(\frac{\ln(\frac{n-1}{\delta})}{2k}\right) \\ &\leq (4k)\ln\left(4 + \ln\left(\frac{t}{2k}\right)\right) + 2k\left(\frac{\ln(\frac{n-1}{\delta})}{2k} + \ln\left(\frac{\ln(\frac{n-1}{\delta})}{2k}\right) + c\right) \\ &= \ln\left(e^{2kc}(n-1)\frac{(4 + \ln\left(\frac{t}{2k}\right))^{4k}}{\delta}\right) + 2k\left(\ln\left(\frac{\ln(\frac{n-1}{\delta})}{2k}\right)\right) \\ &= \ln\left(e^{2kc}(n-1)\frac{(4 + \ln\left(\frac{t}{2k}\right))^{4k}}{\delta}\left(\frac{\ln(\frac{n-1}{\delta})}{2k}\right)^{2k}\right) \end{aligned}$$

And consequently, for sufficiently large t, there exists a constant D such that

$$\hat{c}_t(\delta) \leq \ln\left(\frac{Dt}{\delta}\right).$$

□

Lemma E.6. *For each $\boldsymbol{\mu} \in \mathcal{I}$ and context probability matrix $\mathcal{A}$, the mapping $\boldsymbol{\mu} \to \mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})$ is continuous.*

Proof. Let $\mathcal{M}$ be the interior of Δ^{n-1}, i.e., all $\mathbf{w} \in \Delta^{n-1}$ with strictly positive entries. The mapping $(\boldsymbol{\mu}', \mathbf{w}') \to \min_{i \neq i^*(\boldsymbol{\mu})} \frac{\Delta_i^2}{2} \left(\frac{w_{i^*(\boldsymbol{\mu})} w_i}{w_{i^*(\boldsymbol{\mu})} + w_i} \right)$ is jointly continuous on $(\mathcal{I} \times \mathcal{M})$, as it is the minimum of finitely many continuous functions. Because constraint set Δ^{n-1} does not depend on $\boldsymbol{\mu}'$, and $\mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})$ is unique, Berge's Maximum Theorem implies the continuity of the mapping $\boldsymbol{\mu} \to \mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})$. □

Since we analyze the expected sample complexity asymptotically, the algorithm's initialization phase contributes only a constant term with respect to δ. Note that the waiting time to collect a sample $(X = i, Z = j)$ follows a geometric distribution with mean $\frac{1}{\mathcal{A}_{j,i}}$, which is at most $\frac{1}{a_{\min}}$. Consequently, the expected time for initialization is at most $\frac{nk}{a_{\min}}$, which is a constant with respect to δ and therefore $\limsup_{\delta \to 0} \frac{\frac{nk}{a_{\min}}}{\ln(1/\delta)} = 0$. For simplicity of the proof, we assume $t = 0$ at the end of the initialization phase.

Note that the set $\mathcal{O}_{i^*}$ is open. By combining this fact and Lemma E.6, for every $\epsilon > 0$, there exists a constant $\xi = \xi(\epsilon, \boldsymbol{\mu}, \mathcal{A})$ such that

$$\text{if } \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty \leq \xi \Rightarrow \boldsymbol{\mu}' \in \mathcal{I},\ i^*(\boldsymbol{\mu}') = i^*(\boldsymbol{\mu}),\ \|\mathbf{w}^*(\boldsymbol{\mu}', \mathcal{A}) - \mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})\|_\infty \leq \epsilon.$$

Let $\mathcal{I}_\epsilon = \{\boldsymbol{\mu}' \in \mathcal{I} \big| \ \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty \leq \xi\}$. We now define the following event and show that it occurs with high probability.

$$\mathcal{E}_T(\epsilon) \triangleq \bigcap_{t=T^{1/4}}^{T} (\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}_\epsilon).$$

Note that on the event $\mathcal{E}_T(\epsilon)$, Algorithm 2 does not play randomly for $t > T^{\frac{1}{4}}$, and uses the D-tracking rule.

Lemma E.7. *There exist two constants B and C such that*

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

We provide the proof in Appendix E.4.

We define another event

$$\mathcal{E}_T'(\epsilon) \triangleq \bigcap_{t=T^{1/4}}^{T} \bigcap_{i \in [n], j \in [k]} \left(\left| N_{j,i}(t) - \mathcal{A}_{j,i} N_i^X(t) \right| \leq t\epsilon \right),$$

which occurs with high probability.

Lemma E.8. *There exist two constants B' and C' such that*

$$P(\mathcal{E}_T'^c(\epsilon)) \leq B'T \exp\left(-C'T^{3/8}\right).$$

We provide the proof of this lemma in Appendix E.5.

Here, we review an important lemma from [GK16], which demonstrates the effectiveness of the D-tracking rule.

Lemma E.9 ([GK16], Lemma 20)**.** *For each $\epsilon > 0$, there exists a constant T_ϵ, such that for each $T \geq T_\epsilon$, on event $\mathcal{E}_T(\epsilon)$, we have*

$$\forall t \geq \sqrt{T} : \left\| \frac{\mathbf{N}^X(t)}{t} - \mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A}) \right\|_\infty \leq 3(n-1)\epsilon.$$

Define $w^*_{j,i} \triangleq \mathcal{A}_{j,i}\mathbf{w}^*_i(\boldsymbol{\mu}, \mathcal{A})$. By the triangle inequality, we have

$$\begin{aligned}\left|\frac{N_{j,i}(t)}{t} - w^*_{j,i}\right| &\leq \left|\frac{N_{j,i}(t)}{t} - \mathcal{A}_{j,i}\frac{N^X_i(t)}{t}\right| + \left|\mathcal{A}_{j,i}\frac{N^X_i(t)}{t} - w^*_{j,i}\right| \\ &= \left|\frac{N_{j,i}(t)}{t} - \mathcal{A}_{j,i}\frac{N^X_i(t)}{t}\right| + \left|\mathcal{A}_{j,i}\frac{N^X_i(t)}{t} - \mathcal{A}_{j,i}\mathbf{w}^*_i(\boldsymbol{\mu}, \mathcal{A})\right| \\ &\leq \left|\frac{N_{j,i}(t)}{t} - \mathcal{A}_{j,i}\frac{N^X_i(t)}{t}\right| + \left|\frac{N^X_i(t)}{t} - \mathbf{w}^*_i(\boldsymbol{\mu}, \mathcal{A})\right|.\end{aligned}$$

where the last inequality holds because $\mathcal{A}_{j,i} < 1$.

According to Lemma E.9, on the event $\mathcal{E}_T(\epsilon) \cap \mathcal{E}'_T(\epsilon)$, we have

$$\left|\frac{N_{j,i}(t)}{t} - w^*_{j,i}\right| \leq \left|\frac{N_{j,i}(t)}{t} - \mathcal{A}_{j,i}\frac{N^X_i(t)}{t}\right| + \left|\frac{N^X_i(t)}{t} - \mathbf{w}^*_i(\boldsymbol{\mu}, \mathcal{A})\right| \leq (3n-2)\epsilon$$

Finally, On the event $\mathcal{E}_T(\epsilon) \cap \mathcal{E}'_T(\epsilon)$, for each $t \geq T^{\frac{1}{4}}$

1. $\hat{\boldsymbol{\mu}}(t)$ stays close to $\boldsymbol{\mu}$.
2. $i^*(\hat{\boldsymbol{\mu}}(t)) = i^*(\boldsymbol{\mu})$, hence $\text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A}) = \text{Alt}(\boldsymbol{\mu}, \mathcal{A})$.
3. $\frac{N_{j,i}(t)}{t}$ stays close to $\mathbf{w}^*_{j,i}$ for all $i \in [n]$ and $j \in [k]$.

Combining these properties allows us to derive a lower bound on $\hat{\Lambda}(t)$. Recall the set $\mathcal{O}_i = \left\{\boldsymbol{\mu} \in \mathcal{I} \,\middle|\, \mathcal{A}_i^\top \boldsymbol{\mu}_i > \mathcal{A}_j^\top \boldsymbol{\mu}_j,\ \forall j \neq i\right\}$. We define the function $g : \mathcal{O}_{i^*} \times [0,1]^{k\times n} \to \mathbb{R}$ as

$$g(\boldsymbol{\mu}', \mathbf{w}') = \inf_{\lambda \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{i\in[n], j\in[k]} w'_{j,i}\frac{(\mu'_{j,i} - \lambda_{j,i})^2}{2},$$

and the constant

$$C^*_\epsilon(\boldsymbol{\mu}, \mathcal{A}) = \inf_{\substack{\boldsymbol{\mu}' \in \mathcal{I}_\epsilon \\ \mathbf{w}' : |w'_{j,i} - w^*_{j,i}| \leq (3n-2)\epsilon}} g(\boldsymbol{\mu}', \mathbf{w}').$$

Note that the mapping $(\lambda, \boldsymbol{\mu}', \mathbf{w}') \to \sum_{i\in[n], j\in[k]} w'_{j,i}\frac{(\mu'_{j,i} - \lambda_{j,i})^2}{2}$ is jointly continuous and the constraints set $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ is independent of $(\boldsymbol{\mu}', \mathbf{w}')$. Therefore, applying Berge's maximum theorem [Ber63] implies that g is continuous and the constant $C^*_\epsilon(\boldsymbol{\mu}, \mathcal{A})$ exists.

With these definitions, let $T \geq T_\epsilon$ and event $\mathcal{E}_T(\epsilon) \cap \mathcal{E}'_T(\epsilon)$ holds, then for $t \geq \sqrt{T}$, $\hat{\Lambda}(t)$ is equal to $g(\hat{\boldsymbol{\mu}}(t), \frac{N(t)}{t})$, implying that $\hat{\Lambda}(t) \geq tC^*_\epsilon(\boldsymbol{\mu}, \mathcal{A})$. Using this inequality, for $T \geq T_\epsilon$ and on event $\mathcal{E}_T(\epsilon) \cap \mathcal{E}'_T(\epsilon)$

$$\begin{aligned}\min(\tau_\delta, T) &\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{(\tau_\delta > t)} \leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{(\hat{\Lambda}(t) \leq \hat{c}_t(\delta))} \\ &\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{(tC^*_\epsilon(\boldsymbol{\mu}, \mathcal{A}) \leq \hat{c}_t(\delta))} \leq \sqrt{T} + \frac{\hat{c}_T(\delta)}{C^*_\epsilon(\boldsymbol{\mu}, \mathcal{A})}.\end{aligned}$$

Defining

$$T_\epsilon(\delta) = \inf\left\{T \in \mathbb{N} \,\middle|\, \sqrt{T} + \frac{\hat{c}_T(\delta)}{C^*_\epsilon(\boldsymbol{\mu}, \mathcal{A})} \leq T\right\},$$

for $T \geq \max(T_\epsilon, T_\epsilon(\delta))$, event $\mathcal{E}_T(\epsilon) \cap \mathcal{E}'_T(\epsilon)$ implies $\tau_\delta \leq T$ which shows

$$P_{\boldsymbol{\mu}, \mathcal{A}}(\tau_\delta > T) \leq P(\mathcal{E}^c_T(\epsilon) \cup \mathcal{E}'^c_T(\epsilon)) \leq BT\exp(-CT^{1/8}) + B'T\exp(-C'T^{3/8}).$$

Now using this inequality and the fact that the initialization phase of our algorithm has a sample complexity of at most $\frac{nk}{a_{\min}}$, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu},\mathcal{A}}[\tau_\delta] &\leq T_\epsilon + T_\epsilon(\delta) + \sum_{T\geq \max(T_\epsilon, T_\epsilon(\delta))} P(\tau_\delta > T) \\ &\leq T_\epsilon + T_\epsilon(\delta) + \sum_{T\geq \max(T_\epsilon, T_\epsilon(\delta))} P(\mathcal{E}_T^c(\epsilon) \cup \mathcal{E}_T'^c(\epsilon)) \\ &\leq T_\epsilon + T_\epsilon(\delta) + \sum_{T=1}^{\infty} BT\exp(-CT^{1/8}) + \sum_{T=1}^{\infty} B'T\exp(-C'T^{3/8}). \end{aligned} \tag{19}$$

In this inequality, $\sum_{T=1}^{\infty} BT\exp(-CT^{1/8}) + \sum_{T=1}^{\infty} B'T\exp(-C'T^{3/8})$ is a constant value, T_ϵ is independent of δ and we provide an upper bound on the value of $T_\epsilon(\delta)$. We define a new constant $C(\alpha)$ for $\alpha > 0$ as

$$C(\alpha) = \inf\left\{T \in \mathbb{N} \middle| T - \sqrt{T} \geq \frac{T}{1+\alpha}\right\}.$$

Then using the upper bound on thresholds in Lemma E.5, one obtains

$$T_\epsilon(\delta) \leq E + C(\alpha) + \inf\left\{T \in \mathbb{N} \middle| \frac{\ln\left(\frac{DT}{\delta}\right)}{C_\epsilon^*(\boldsymbol{\mu},\mathcal{A})} \leq \frac{T}{1+\alpha}\right\}.$$

Using Proposition 8 in [KK21], we have the following

$$\begin{aligned}&\inf\left\{T \in \mathbb{N} \middle| \frac{\ln\left(\frac{DT}{\delta}\right)}{C_\epsilon^*(\boldsymbol{\mu},\mathcal{A})} \leq \frac{T}{1+\alpha}\right\} \\ &\leq \frac{1+\alpha}{C_\epsilon^*(\boldsymbol{\mu},\mathcal{A})}\left[\ln\left(\frac{(1+\alpha)D}{C_\epsilon^*(\boldsymbol{\mu},\mathcal{A})\delta}\right) + \ln\left(\ln\left(\frac{(1+\alpha)D}{C_\epsilon^*(\boldsymbol{\mu},\mathcal{A})\delta}\right) + \sqrt{2\ln\left(\frac{(1+\alpha)D}{C_\epsilon^*(\boldsymbol{\mu},\mathcal{A})\delta}\right) - 2}\right)\right].\end{aligned}$$

Inserting this inequality in (19), we obtain the following for every $\alpha, \epsilon > 0$

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu},\mathcal{A}}[\tau_\delta]}{\ln(1/\delta)} \leq \frac{1+\alpha}{C_\epsilon^*(\boldsymbol{\mu},\mathcal{A})}.$$

Letting both of α and ϵ go to zero, by continuity of g,

$$\lim_{\delta\to 0} C_\epsilon^*(\boldsymbol{\mu},\mathcal{A}) = T_{NS}^*(\boldsymbol{\mu},\mathcal{A})^{-1}$$

holds which implies

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu},\mathcal{A}}[\tau_\delta]}{\ln(1/\delta)} \leq T_{NS}^*(\boldsymbol{\mu},\mathcal{A}).$$

□

E.4 Proof of Lemma E.7

As a consequence of the forced exploration phase in the D-tracking algorithm, for any round t with $t > n^2$, each arm has been pulled at least $\sqrt{t} - n$ times (refer to Lemma 7 in [GK16]). Applying the union bound over t, we obtain

$$P(\mathcal{E}_T^c(\epsilon)) \leq \sum_{t=T^{1/4}}^{T} P(\hat{\boldsymbol{\mu}}(t) \notin I_\epsilon)$$

Similarly, we can apply the union bound for the event $\{\hat{\boldsymbol{\mu}}(t) \notin I_\epsilon\}$ over each arm-context pair

$$P(\hat{\boldsymbol{\mu}}(t) \notin I_\epsilon) \leq \sum_{\substack{i\in[n]\\ j\in[k]}} P(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi).$$

Hence,

$$P(\mathcal{E}_T^c(\epsilon)) \leq \sum_{t=T^{1/4}}^{T} \sum_{\substack{i\in[n]\\ j\in[k]}} P(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi).$$

Recall that $N_{j,i}(t)$ is the number of samples collected from $P(Y|X=i, Z=j)$ until round t. Then,

$$\begin{aligned}
&P\left(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi\right) \\
&= P\left(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi \middle| N_{j,i}(t) \geq \frac{a_{\min}}{2}(\sqrt{t} - n)\right) P\left(N_{j,i}(t) \geq \frac{a_{\min}}{2}(\sqrt{t} - n)\right) \\
&+ P\left(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi \middle| N_{j,i}(t) < \frac{a_{\min}}{2}(\sqrt{t} - n)\right) P\left(N_{j,i}(t) < \frac{a_{\min}}{2}(\sqrt{t} - n)\right)
\end{aligned}$$

By upper bounding $P\left(N_{j,i}(t) \geq \frac{a_{\min}}{2}(\sqrt{t} - n)\right) \leq 1$ and $P\left(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi \middle| N_{j,i}(t) < \frac{a_{\min}}{2}(\sqrt{t} - n)\right) \leq 1$, we have

$$\begin{aligned}
P\left(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi\right) \leq P\left(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi \middle| N_{j,i}(t) \geq \frac{a_{\min}}{2}(\sqrt{t} - n)\right) \\
+ P\left(N_{j,i}(t) < \frac{a_{\min}}{2}(\sqrt{t} - n)\right).
\end{aligned}$$

Both terms can be bounded using Hoeffding's inequality. Since rewards are 1-sub-Gaussian random variables. First,

$$P\left(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi \middle| N_{j,i}(t) \geq \frac{a_{\min}}{2}(\sqrt{t} - n)\right) \leq 2\exp\left(-\frac{a_{\min}(\sqrt{t} - n)\xi^2}{4}\right).$$

For the second term, note that each context j occurs with probability at least $a_{\min}$ for each action i. We can write $N_{j,i}(t)$ as the sum of $N_i^X(t)$ Bernoulli random variables with parameter $\mathcal{A}_{j,i}$, denoted by S_i. Then,

$$\begin{aligned}
&P\left(N_{j,i}(t) < \frac{a_{\min}}{2}(\sqrt{t} - n)\right) \\
&= P\left(\sum_{m=1}^{N_i^X(t)} S_m < \frac{a_{\min}}{2}(\sqrt{t} - n)\right) \overset{S_m \geq 0}{\leq} P\left(\sum_{m=1}^{\sqrt{t}-n} S_m < \frac{a_{\min}}{2}(\sqrt{t} - n)\right) \\
&= P\left(\sum_{m=1}^{\sqrt{t}-n} S_m - (\sqrt{t} - n)\mathcal{A}_{j,i} < \frac{a_{\min}}{2}(\sqrt{t} - n) - (\sqrt{t} - n)\mathcal{A}_{j,i}\right) \\
&\overset{\mathcal{A}_{j,i} \geq a_{\min}}{\leq} P\left(\sum_{m=1}^{\sqrt{t}-n} S_m - (\sqrt{t} - n)\mathcal{A}_{j,i} < -\frac{a_{\min}}{2}(\sqrt{t} - n)\right)
\end{aligned}$$

Because Bernoulli variables are $\frac{1}{2}$-sub-gaussian, Hoeffding's inequality implies

$$P\left(\sum_{m=1}^{(\sqrt{t}-n)} S_m - (\sqrt{t} - n)\mathcal{A}_{j,i} < -\frac{a_{\min}}{2}(\sqrt{t} - n)\right) \leq \exp\left(-\frac{a_{\min}^2(\sqrt{t} - n)}{2}\right)$$

Combining both bounds, we get

$$\begin{aligned}
P(\mathcal{E}_T^c(\epsilon)) &\leq \sum_{t=T^{1/4}}^{T} \sum_{\substack{i\in[n]\\ j\in[k]}} P(|\hat{\mu}_{j,i}(t) - \mu_{j,i}| > \xi) \\
&\leq \sum_{t=T^{1/4}}^{T} \sum_{\substack{i\in[n]\\ j\in[k]}} 2\exp\left(-\frac{a_{\min}(\sqrt{t} - n)\xi^2}{4}\right) + \exp\left(-\frac{a_{\min}^2(\sqrt{t} - n)}{2}\right)
\end{aligned}$$

Define

$$C = \min\left(\frac{a_{\min}\xi^2}{4}, \frac{a_{\min}^2}{2}\right), B = nk\left(2\exp\left(\frac{na_{\min}\xi^2}{4}\right) + \exp\left(\frac{na_{\min}^2}{2}\right)\right).$$

Then, we have

$$P(\mathcal{E}_T^c(\epsilon)) \leq \sum_{t=T^{1/4}}^{T} B\exp(-C\sqrt{t}) \leq \sum_{t=T^{1/4}}^{T} B\exp(-CT^{1/8}) \leq BT\exp(-CT^{1/8}),$$

which concludes the proof.

E.5 Proof of Lemma E.8

Note that

$$N_{j,i}(t) = \sum_{s=1}^{t} \mathbb{1}_{(X_s=i, Z_s=j)} = \sum_{m=1}^{N_i^X(t)} \mathbb{1}_{(Z_{t_m}=j)},$$

where t_ms denotes the rounds in which the action i is pulled. Indeed, $\mathbb{1}_{(Z_{t_m}=j)}$ is a Bernoulli random variable with parameter $\mathcal{A}_{j,i}$. Since Bernoulli variables are $\frac{1}{2}$-sub-Gaussian, after applying Hoeffding's inequality, we have

$$\begin{aligned} P\left(\left|N_{j,i}(t) - \mathcal{A}_{j,i}N_i^X(t)\right| \geq t\epsilon\right) &= P\left(\left|\sum_{m=1}^{N_i^X(t)} \mathbb{1}_{(Z_{t_m}=j))} - \mathcal{A}_{j,i}N_i^X(t)\right| \geq t\epsilon\right) \leq 2\exp\left(-2\frac{t^2\epsilon^2}{N_i^X(t)}\right) \\ &\leq 2\exp\left(-4\frac{t^2\epsilon^2}{\sqrt{t}}\right) = 2\exp\left(-4t^{3/8}\epsilon^2\right) \end{aligned}$$

where the last inequality is achieved by the fact $N_i^X(t) \geq \sqrt{t} - \frac{n}{2} \geq \frac{\sqrt{t}}{2}$. Therefore, applying the union bound on $\mathcal{E}_T'^c(\epsilon)$ leads to

$$\begin{aligned} P(\mathcal{E}_T'^c(\epsilon)) &\leq \sum_{t=T^{1/4}}^{T} \sum_{i\in[n], j\in[k]} P\left(\left|N_{j,i}(t) - \mathcal{A}_{j,i}N_i^X(t)\right| \geq t\epsilon\right) \\ &\leq \sum_{t=T^{1/4}}^{T} \sum_{i\in[n], j\in[k]} 2\exp\left(-4t^{1.5}\epsilon^2\right) \\ &\leq 2nkT\exp\left(-4T^{3/8}\epsilon^2\right). \end{aligned}$$

Setting $B' = 2nk$ and $C' = 4\epsilon^2$, completes the proof.

F Proofs of Section 3

We define $\mathcal{I}^s(\mathcal{A})$ as the set of vectors in $\mathbb{R}^k$ that ensure a unique best arm in the separator setting with context probability matrix $\mathcal{A}$.

F.1 Mathematical Formulation of G-tracking

At round t, the G-tracking rule selects the next target point as the intersection of the ray from $\overline{N}(t) = \frac{N^Z(t)}{t}$ toward $w^*(t)$ with the boundary of the convex hull of the context probability vectors $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$.

This ray is given by the set of points of the form:

$$\overline{N}(t) + \alpha(w^*(t) - \overline{N}(t)), \quad \text{for } \alpha \geq 1.$$

We want the largest $\alpha \geq 1$ such that this point still lies within $\text{conv}(\mathcal{A}_1, \ldots, \mathcal{A}_n)$, which is the set $\left\{\sum_{i=1}^n \pi_i\mathcal{A}_i \,\middle|\, \pi \in \Delta^{n-1}\right\}$.

This gives rise to the following linear program:

$$\begin{aligned}
&\max_{\alpha,\pi} \ \alpha \\
&\text{s.t. } \ \overline{N}(t) + \alpha(w^*(t) - \overline{N}(t)) = \sum_{i=1}^{n} \pi_i \mathcal{A}_i, \\
&\sum_{i=1}^{n} \pi_i = 1, \quad \pi \geq 0, \quad \alpha \geq 1.
\end{aligned}$$

To implement the G-tracking rule, we solved the above linear program, and π_i is the desired policy.

F.2 Proof of Theorem 3.1

Similar to the non-separator case, we prove this theorem by explicitly solving the optimization problem (17) for the separator setting and inserting this solution into the general lower bound formula (18). Note that in contrast to the non-separator case, an action can obtain zero proportion in optimal weights.

Lemma F.1. *Consider a bandit instance with a separator context and Gaussian reward distribution with unit variance, parameterized by matrices $\boldsymbol{\mu}$ and $\mathcal{A}$. For any weight vector $\mathbf{w}$ with non-zero indices, the value of $f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) = \inf_{\boldsymbol{\mu}' \in \mathcal{C}_i} \sum w_i d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'})$ can be explicitly determined as*

$$f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) = \frac{\Delta_i^2}{2\sum_{j\in[k]} \frac{(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,i})^2}{\sum_{l\in[n]} w_l \mathcal{A}_{j,l}}}.$$

Proof. First we calculate $d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'})$ for two rewards mean vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ using the assumption of Gaussian rewards with unit variance:

$$d(P_i^{\boldsymbol{\mu}}, P_i^{\boldsymbol{\mu}'}) = \sum_{j\in[k]} \mathcal{A}_{j,i} d(\mu_j, \mu'_j) = \sum_{j\in[k]} \mathcal{A}_{j,i} \frac{(\mu_j - \mu'_j)^2}{2}.$$

In the separator setting, $\mathcal{I}^s$ indeed is a subset of vectors $\boldsymbol{\mu} \in \mathbb{R}^k$ that imply a unique best arm. Consequently, the sets $\mathcal{C}_i$ is defined as $\{\boldsymbol{\mu}' \in \mathcal{I}^s \mid \mathcal{A}_i^\top \boldsymbol{\mu}' > \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'\}$. Then using the continuity of f_i in $\boldsymbol{\mu}$ and the fact that $\mathcal{I}^s$ is dense in $\mathbb{R}^k$, we can write $f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A})$ as the solution of the following optimization problem:

$$\begin{aligned}
f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) = &\operatorname*{argmin}_{\boldsymbol{\mu}' \in \mathbb{R}^k} \sum_{s\in[n]} w_s \sum_{j\in[k]} \mathcal{A}_{j,s} \frac{(\mu_j - \mu'_j)^2}{2} \\
&\text{s.t.} \quad \mathcal{A}_i^\top \boldsymbol{\mu}' > \mathcal{A}_{i^*}^\top \boldsymbol{\mu}'.
\end{aligned}$$

The solution to this problem is equivalent to the solution to the following problem when $\alpha \to 0$:

$$\begin{aligned}
&\operatorname*{argmin}_{\boldsymbol{\mu}' \in \mathbb{R}^k} \sum_{s\in[n]} w_s \sum_{j\in[k]} \mathcal{A}_{j,s} \frac{(\mu_j - \mu'_j)^2}{2} \\
&\text{s.t.} \quad \mathcal{A}_i^\top \boldsymbol{\mu}' \geq \mathcal{A}_{i^*}^\top \boldsymbol{\mu}' + \alpha.
\end{aligned}$$

To solve this problem, we define the Lagrangian as

$$L(\boldsymbol{\mu}', \lambda) \ = \ \sum_{s\in[n]} w_s \sum_{j\in[k]} \mathcal{A}_{j,s} \frac{(\mu_j - \mu'_j)^2}{2} \ + \ \lambda \Big((\mathcal{A}_{i^*} - \mathcal{A}_i)^\top \boldsymbol{\mu}' + \alpha \Big).$$

Taking the derivative with respect to μ'_j and λ and setting it to zero with some basic algebra gives the following optimal values

$$\mu'_j \;=\; \mu_j \;-\; \lambda \frac{(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,i})}{\sum\limits_{l\in[n]} w_l A_{j,l}},$$

$$\lambda^* = \frac{\Delta_i + \alpha}{\sum_{j\in[k]} \frac{(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,i})^2}{\sum_{l\in[n]} w_l \mathcal{A}_{j,l}}}.$$

Inserting these values into the objective function proves the lemma.

$$f_i(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) = \lim_{\alpha\to 0} \frac{(\Delta_i + \alpha)^2}{2\sum_{j\in[k]} \frac{(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,i})^2}{\sum_{l\in[n]} w_l \mathcal{A}_{j,l}}} = \frac{\Delta_i^2}{2\sum_{j\in[k]} \frac{(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,i})^2}{\sum_{l\in[n]} w_l \mathcal{A}_{j,l}}}$$

□

Combining the aforementioned lemma with Equation (18) completes the proof.

F.3 Proof of Lemma 3.2

Lemma 3.2. *Consider a bandit instance with a separator context and Gaussian reward distribution with unit variance, parameterized by the matrix $\mathcal{A}$ and the vector $\boldsymbol{\mu}$. Any algorithm with the stopping rule of* (6) *is δ-correct, that is, $P_{\boldsymbol{\mu},\mathcal{A}}(\tau_\delta < \infty, \hat{i}_\tau \neq i^*(\boldsymbol{\mu})) \leq \delta$.*

Proof. As discussed in Appendix D, our problem can be viewed as a general Identification problem. In separator case, $\boldsymbol{\mu}$ are k-dimensional vectors in $\mathcal{I}^s$. Thus, we can redefine $\mathcal{O}_i$

$$\mathcal{O}_i = \left\{\boldsymbol{\mu} \in \mathcal{I}^s \,\middle|\, \mathcal{A}_i^\top \boldsymbol{\mu} > \mathcal{A}_j^\top \boldsymbol{\mu},\ \forall j \neq i\right\}$$

The proof then follows directly from Proposition 15 of [KK21], which implies that the GLR stopping rule with the sequential threshold in (5) is δ-correct. Note that in Proposition 15 of [KK21], the threshold is given for one-parameter exponential families. Since our reward distributions are Gaussian, we employ a tighter threshold that the authors introduce in their paper. □

F.4 Proof of Theorem 3.3

Recall that $\mathcal{I}^s(\mathcal{A})$ denotes the set of vectors in $\mathbb{R}^k$ which implies a unique best arm in the separator setting with context probability matrix $\mathcal{A}$. For a point $p \in \mathbb{R}^k$ and a set $C \in \mathbb{R}^k$, we use $dist_{L^2}(p, C)$ to denote the distance of p to C with respect to L^2 norm, more precisely

$$dist_{L^2}(p, C) = \inf_{c\in C} \|p - c\|_2 .$$

Similar to the non-separator case, we need to present a few helper lemmas.

Lemma F.2. *For each $\boldsymbol{\mu} \in \mathcal{I}^s$ and context probability matrix $\mathcal{A}$, the set $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ is a convex set and the set-valued mapping $\boldsymbol{\mu} \to \mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ is upper hemicontinuous.*

Proof. Recall that $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ is the set of solutions that maximize

$$F(\mathbf{w}, \boldsymbol{\mu}) \triangleq \inf_{\boldsymbol{\mu}'\in \mathrm{Alt}(\boldsymbol{\mu},\mathcal{A})} \sum_{j\in[k]} w_j \frac{(\boldsymbol{\mu}_j - \boldsymbol{\mu}'_j)^2}{2}$$

where $\mathbf{w} \in \mathrm{conv}(\mathcal{A})$. Since this function is the infimum over linear functions in $\mathbf{w}$, thus, F is concave with respect to $\mathbf{w}$. Consequently, if $\mathbf{w}_1$ and $\mathbf{w}_2$ are both maximizers of F, then any convex combination of them is also a maximizer, implying that $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ is a convex set.

As shown in Lemma F.1, $F(\mathbf{w})$ can be written as

$$\min_{i\neq i^*(\boldsymbol{\mu})} \frac{\Delta_i^2}{2\sum_{j\in[k]} \frac{(\mathcal{A}_{j,i^*} - \mathcal{A}_{j,i})^2}{w_{z,j}}},$$

which is jointly continuous on $\mathrm{conv}(\mathcal{A}) \times I^s$. By applying Theorem 22 in [DK19], we conclude that the mapping $\boldsymbol{\mu} \to \mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ is upper hemicontinuous. □

As we analyze the sample complexity asymptotically, the initialization phase of the algorithm contributes only a constant term. This is because, for each context i, the expected waiting time to collect a sample with $Z = i$ is at most $\frac{1}{a_{\min}}$, as the probability of observing context i at each round is at least $a_{\min}$. Thus, the expected number of initialization steps is at most $\frac{k}{a_{\min}}$, which is a constant, and $\limsup_{\delta \to 0} \frac{\frac{k}{a_{\min}}}{\ln(1/\delta)} = 0$. For simplicity, we set $t = 0$ at the end of the initialization phase.

For a non-negative real number α, we define the $L^2(\alpha)$-approximation set of $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ as

$$\mathbf{w}_\alpha^*(\boldsymbol{\mu}, \mathcal{A}) = \left\{ p \in \Delta^{k-1} \mid dist_{L^2}\left(p, \mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})\right) \leq \alpha \right\}.$$

Based on Lemma F.2, the set $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ is convex which shows that for each α, the set $\mathbf{w}_\alpha^*(\boldsymbol{\mu}, \mathcal{A})$ is also convex. [2]

By Lemma F.2 and the fact that the set of vectors $\boldsymbol{\mu}'$ which has the same best arm as $\boldsymbol{\mu}$ is open, for each real number $\epsilon > 0$, there exist a constant $\xi = \xi(\epsilon, \boldsymbol{\mu}, \mathcal{A})$ such that

$$\text{if } \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty \leq \xi \Rightarrow \boldsymbol{\mu}' \in \mathcal{I}^s(\mathcal{A}),\ i^*(\boldsymbol{\mu}') = i^*(\boldsymbol{\mu}), \forall \mathbf{w}' \in \mathbf{w}_z^*(\boldsymbol{\mu}', \mathcal{A}) : \mathbf{w}' \in \mathbf{w}_\epsilon^*(\boldsymbol{\mu}, \mathcal{A}).$$

Let $\mathcal{I}_\epsilon = \{\boldsymbol{\mu}' \in \mathcal{I}(\mathcal{A}) \big| \ \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty \leq \xi\}$. For each T, we define two events that are likely to happen as

$$\mathcal{E}_T(\epsilon) \triangleq \bigcap_{t=T^{1/4}}^{T} (\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}_\epsilon),$$
$$\mathcal{E}'_T(\epsilon) \triangleq \bigcap_{t=T^{1/4}}^{T} \left(\left\| \mathbf{N}^Z(t) - \sum_{i \in [t]} \mathbf{p}(i) \right\|_2 < t^{\frac{3}{4}} \right).$$

Recall that $\mathbf{N}^Z(t)$ shows the number of samples collected from contexts until round t and $\mathbf{p}(i)$ shows the policy on the context values that is chosen by G-tracking at round i. The following lemma shows these events occur with high probability.

Lemma F.3. *The exist two constants B and C such that*

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

We defer the proof of Lemma F.3 to Appendix F.5. For now, we present a lemma that establishes the convergence of the G-tracking rule.

Lemma F.4. *For each $\epsilon > 0$, there exist a constant T_ϵ such that for each $T \geq T_\epsilon$, on event $\mathcal{E}_T(\epsilon) \cap \mathcal{E}'_T(\epsilon)$, we have*

$$\forall t \geq \sqrt{T} : dist_{L^2}\left(\frac{\mathbf{N}^Z(t)}{t}, \mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A}) \right) \leq 5\epsilon.$$

Proof. First, we introduce some geometric notations. For points $A, B, C, D \in \mathbb{R}^k$, let $\overline{AB}$ show the length of the segment connecting A to B, AB show the line passing through A and B, and $AB \parallel CD$ show that vectors $\overrightarrow{AB}$ and $\overrightarrow{CD}$ are parallel. At each round t, let $\overline{\mathbf{p}}(t) = \frac{\sum_{i \in [t]} \mathbf{p}(i)}{t}, \overline{\mathbf{N}}(t) = \frac{\mathbf{N}^Z(t)}{t}$, and

$$Y_t = dist_{L^2}\left(\overline{\mathbf{N}}(t), \mathbf{w}_{2\epsilon}^*(\boldsymbol{\mu}, \mathcal{A})\right), \quad X_t = dist_{L^2}\left(\overline{\mathbf{p}}(t), \mathbf{w}_{2\epsilon}^*(\boldsymbol{\mu}, \mathcal{A})\right).$$

[2] This follows directly from the definition of convexity, as any convex combination of two points in $\mathbf{w}_\alpha^*(\boldsymbol{\mu}, \mathcal{A})$ remains within $\mathbf{w}_\alpha^*(\boldsymbol{\mu}, \mathcal{A})$.

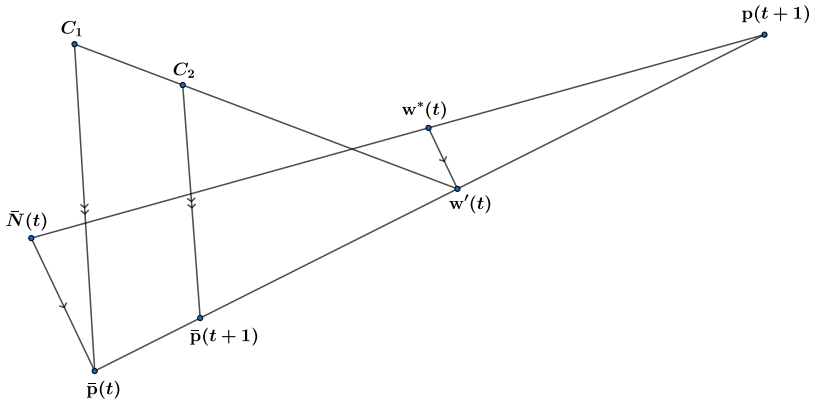


Figure 5: Illustration of the relative positions of points in the proof of Lemma F.4, where all points lie in Δ^{k-1}.

Moreover, let $\mathbf{w}'(t)$ be the unique point on the line $\overline{\mathbf{p}}(t)\mathbf{p}(t+1)$ such that $\overline{\mathbf{N}}(t)\overline{\mathbf{p}}(t) \parallel \mathbf{w}^*(t)\mathbf{w}'(t)$. C_1 denotes the closest point in $\mathbf{w}^*_{2\epsilon}(\boldsymbol{\mu}, \mathcal{A})$ to $\overline{\mathbf{p}}(t)$ (note that $\mathbf{w}^*_{2\epsilon}(\boldsymbol{\mu}, \mathcal{A})$ is closed) and C_2 denotes the unique point on $C_1\mathbf{w}'(t)$ such that $C_2\overline{\mathbf{p}}(t+1) \parallel C_1\overline{\mathbf{p}}(t)$. Figure 5 provides a complete illustration of all these points.

We set $T_\epsilon = \epsilon^{-16}$. By this choice of T_ϵ, if $T \geq T_\epsilon$, for each $t \geq T^{\frac{1}{4}}$ we have

$$\overline{\overline{\mathbf{p}}(t)\overline{\mathbf{p}}(t+1)} \leq \epsilon \quad \text{and} \quad \overline{\overline{\mathbf{N}}(t)\overline{\mathbf{p}}(t)} \leq \epsilon. \tag{20}$$

The reason is that

$$\overline{\overline{\mathbf{p}}(t)\overline{\mathbf{p}}(t+1)} = \frac{1}{t+1}\overline{\overline{\mathbf{p}}(t)\mathbf{p}(t+1)} \leq \frac{\sqrt{2}}{t+1} \leq \frac{\sqrt{2}}{T^{1/4}} \leq \epsilon.$$

The first inequality holds because the points $\overline{\mathbf{p}}(t)$ and $\mathbf{p}(t+1)$ are on Δ^{k-1} and the distance of any pair of points of the simplex is at most $\sqrt{2}$. For the second property, note that as $\mathcal{E}'_T(\epsilon)$ holds, $\overline{\overline{\mathbf{N}}(t)\overline{\mathbf{p}}(t)} < t^{-1/4} < T^{-1/16} \leq \epsilon$.

Based on the definition, $C_1 \in \mathbf{w}^*_{2\epsilon}(\boldsymbol{\mu}, \mathcal{A})$ and as $\mathcal{E}_T(\epsilon)$ holds, $\mathbf{w}^*(t) \in \mathbf{w}^*_\epsilon(\boldsymbol{\mu}, \mathcal{A})$. Using Thales's theorem in triangle $\overline{\mathbf{N}}(t)\overline{\mathbf{p}}(t)\mathbf{p}(t+1)$, we obtain $\overline{\mathbf{w}^*(t)\mathbf{w}'(t)} \leq \overline{\overline{\mathbf{N}}(t)\overline{\mathbf{p}}(t)} \leq \epsilon$ which means that $\mathbf{w}'(t) \in \mathbf{w}^*_{2\epsilon}(\boldsymbol{\mu}, \mathcal{A})$. By convexity of the set $\mathbf{w}^*_{2\epsilon}(\boldsymbol{\mu}, \mathcal{A})$, we have $C_2 \in \mathbf{w}^*_{2\epsilon}(\boldsymbol{\mu}, \mathcal{A})$.

We prove the following two statements for each $t \geq T^{1/4}$:

$$\text{if } X_t > \epsilon \Longrightarrow X_{t+1} \leq \frac{t}{t+1}X_t, \tag{21}$$

$$\text{if } X_t \leq \epsilon \Longrightarrow \forall t' > t : X_{t'} \leq 2\epsilon. \tag{22}$$

To prove (21), note that as $\mathbf{w}'(t) \in \mathbf{w}^*_{2\epsilon}(\boldsymbol{\mu}, \mathcal{A})$, if $X_t > \epsilon$ then $\overline{\overline{\mathbf{p}}(t)\mathbf{w}'(t)} > \epsilon$. This shows that based on (20), $\overline{\mathbf{p}}(t+1)$ lines on the segment between $\overline{\mathbf{p}}(t)\mathbf{w}'(t)$. Then using the Thales's theorem in triangle $C_1\overline{\mathbf{p}}(t)\mathbf{w}'(t)$, we have

$$\begin{aligned} X_{t+1} = dist_{L^2}\left(\overline{\mathbf{p}}(t+1), \mathbf{w}^*_{2\epsilon}(\boldsymbol{\mu}, \mathcal{A})\right) &\leq \overline{C_2\overline{\mathbf{p}}(t+1)} = \frac{\overline{\mathbf{w}'(t)\overline{\mathbf{p}}(t+1)}}{\overline{\mathbf{w}'(t)\overline{\mathbf{p}}(t)}}\overline{C_1\overline{\mathbf{p}}(t)} \\ &\leq \frac{\overline{\mathbf{p}(t+1)\overline{\mathbf{p}}(t+1)}}{\overline{\mathbf{p}(t+1)\overline{\mathbf{p}}(t)}}\overline{C_1\overline{\mathbf{p}}(t)} = \frac{t}{t+1}\, dist_{L^2}\left(\overline{\mathbf{p}}(t), \mathbf{w}^*_{2\epsilon}(\boldsymbol{\mu}, \mathcal{A})\right) = \frac{t}{t+1}X_t. \end{aligned}$$

For the second inequality, we used the property that for positive numbers a, b, c with $a < b$, we have $\frac{a}{b} < \frac{a+c}{b+c}$.

We prove (22) by contradiction. Assume $X_t \leq \epsilon$ and there exist a $t' > t$ such that $X_{t'} > 2\epsilon$ and t' is the smallest number with this property. As $X_{t'} > 2\epsilon$, then (20) implies $X_{t'-1} > \epsilon$. Using (21), we have $X_{t'} \leq \frac{t'}{t'+1} X_{t'-1}$ which implies $X_{t'-1} > 2\epsilon$ which is a contradiction and proves (22).

Now we introduce the time τ as

$$\tau \triangleq \inf \left\{ t > T^{1/4} \mid X_t \leq \epsilon \right\}.$$

Note that as a result of 21, τ is finite. The aim is to prove $\tau \leq \sqrt{T}$. To show this assume $\tau > \sqrt{T}$. In this case, for all $t \in [\lceil T^{1/4} \rceil, \lfloor \sqrt{T} \rfloor]$, $X_t > \epsilon$ which means that $X_{t+1} \leq \frac{t}{t+1} X_t$. Then

$$\begin{aligned} X_{\lfloor \sqrt{T} \rfloor} &\leq X_{\lceil T^{1/4} \rceil} \times \frac{\lceil T^{1/4} \rceil}{\lceil T^{1/4} \rceil + 1} \times \frac{\lceil T^{1/4} \rceil + 1}{\lceil T^{1/4} \rceil + 2} \times \cdots \times \frac{\lfloor \sqrt{T} \rfloor - 1}{\lfloor \sqrt{T} \rfloor} \\ &= X_{\lceil T^{1/4} \rceil} \frac{\lceil T^{1/4} \rceil}{\lfloor \sqrt{T} \rfloor} \leq \sqrt{2} \frac{\lceil T^{1/4} \rceil}{\lfloor \sqrt{T} \rfloor} \leq \epsilon, \end{aligned}$$

where we used $X_{\lceil T^{1/4} \rceil} \leq \sqrt{2}$ and $T \geq T_\epsilon$. The above inequality shows a contradiction which implies $\tau \leq \sqrt{T}$. Then, we have

$$\tau \leq \sqrt{T} \Longrightarrow \forall t > \sqrt{T} : X_t \leq 2\epsilon \xRightarrow{(20)} Y_t \leq 3\epsilon \Longrightarrow dist_{L^2} \left(\frac{\mathbf{N}^Z(t)}{t}, \mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A}) \right) \leq 5\epsilon.$$

□

On the good event $\mathcal{E}_T(\epsilon) \cap \mathcal{E}'_T(\epsilon)$, for each $t \geq \sqrt{T}$, these three properties hold: (i) $\hat{\boldsymbol{\mu}}(t)$ is close to $\boldsymbol{\mu}$, (ii) $i^*(\hat{\boldsymbol{\mu}}(t)) = i^*(\boldsymbol{\mu})$ which implies $\text{Alt}(\hat{\boldsymbol{\mu}}(t), \mathcal{A}) = \text{Alt}(\boldsymbol{\mu}, \mathcal{A})$, and (iii) the points $\mathbf{N}^Z(t)/t$ to the set of solutions $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ stay close.

Similar to the non-separator setting, by combining these properties we can derive a lower bound on $\hat{\Lambda}(t)$. Recall the set $\mathcal{O}_i = \left\{ \boldsymbol{\mu} \in \mathcal{I}^s \,\middle|\, \mathcal{A}_i^\top \boldsymbol{\mu} > \mathcal{A}_j^\top \boldsymbol{\mu},\ \forall j \neq i \right\}$. We define the function $g : \mathcal{O}_{i^*} \times [0,1]^k \to \mathbb{R}$ as

$$g(\boldsymbol{\mu}', \mathbf{w}') = \inf_{\lambda \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{j \in [k]} \frac{w'_j}{2} (\mu'_j - \lambda_j)^2,$$

and the constant

$$C_\epsilon^*(\boldsymbol{\mu}, \mathcal{A}) = \inf_{\substack{\boldsymbol{\mu}' \in \mathcal{I}_\epsilon^s \\ \mathbf{w}' : dist_{L^2}\left(\mathbf{w}', \mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})\right) \leq 5\epsilon}} g(\boldsymbol{\mu}', \mathbf{w}').$$

Note that the mapping $(\lambda, \boldsymbol{\mu}', \mathbf{w}') \to \sum_{j \in [k]} \frac{w'_j}{2} (\mu'_j - \lambda_j)^2$ is jointly continuous and the constraints set $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ is independent of $(\boldsymbol{\mu}', \mathbf{w}')$. Then Berge's maximum theorem [Ber63] implies that g is continuous.

With these definitions, let $T \geq T_\epsilon$ and the event $\mathcal{E}_T(\epsilon) \cap \mathcal{E}'_T(\epsilon)$ holds. Then, for $t \geq \sqrt{T}$, $\hat{\Lambda}(t)$ is equal to $g(\hat{\boldsymbol{\mu}}(t), \frac{\mathbf{N}^Z(t)}{t})$, and the three mentioned properties imply that $\hat{\Lambda}(t) \geq t C_\epsilon^*(\boldsymbol{\mu}, \mathcal{A})$.

The rest of the proof follows exactly the same as the non-separator setting.

F.5 Proof of Lemma F.3

We first bound both $P(\mathcal{E}_T^c(\epsilon))$ and $P(\mathcal{E}_T'^c(\epsilon))$ separately, and then we apply the union bound on $P((\mathcal{E}_T(\epsilon) \cap \mathcal{E}'_T(\epsilon))^c)$.

Bounding $\mathcal{E}_T^c(\epsilon)$. Note that according to the union bound, we have

$$P(\mathcal{E}_T^c(\epsilon)) \leq \sum_{t=T^{1/4}}^{T} \sum_{j\in[k]} P(|\hat{\mu}_j(t) - \mu_j| > \xi).$$

Now, note that we have

$$\begin{aligned} P(|\hat{\mu}_j(t) - \mu_j| > \xi) &= P(|\hat{\mu}_j(t) - \mu_j| > \xi \mid N_j^Z(t) < a_{\min}\sqrt{t}) P(N_j^Z(t) < a_{\min}\sqrt{t}) \\ &\quad + P(|\hat{\mu}_j(t) - \mu_j| > \xi \mid N_j^Z(t) \geq a_{\min}\sqrt{t}) P(N_j^Z(t) \geq a_{\min}\sqrt{t}). \end{aligned}$$

Since $P(|\hat{\mu}_j(t) - \mu_j| > \xi \mid N_j^Z(t) < a_{\min}\sqrt{t})$ and $P(N_j^Z(t) \geq a_{\min}\sqrt{t})$ are less than 1, we have

$$P(|\hat{\mu}_j(t) - \mu_j| > \xi) \leq P(N_j^Z(t) < a_{\min}\sqrt{t}) + P(|\hat{\mu}_j(t) - \mu_j| > \xi \mid N_j^Z(t) \geq a_{\min}\sqrt{t})$$

Note that $N_j^Z(t)$ can be written as

$$N_j^Z(t) = \sum_{s=1}^{t} \mathbb{1}_{(z_s=j)}.$$

Therefore, if W_s denotes $\mathbb{1}_{(z_s=j)}$, it is a Bernoulli random variable with parameter greater than $a_{\min}$. By using Hoeffding's inequality, we have

$$\begin{aligned} &P\left(N_j^Z(t) < a_{\min}\sqrt{t}\right) = P\left(\sum_{s=1}^{t} W_s < a_{\min}\sqrt{t}\right) \\ &\leq P\left(\sum_{s=1}^{t} W_s - \mathbb{E}[N_j^Z(t)] < a_{\min}\sqrt{t} - t a_{\min}\right) \\ &\leq \exp\left(\frac{-2a_{\min}^2(t-\sqrt{t})^2}{t}\right) = \exp\left(-2a_{\min}^2(\sqrt{t}-1)^2\right) \\ &\leq \exp\left(-a_{\min}^2 t\right) \end{aligned}$$

The last inequality is correct for sufficiently large t.

Similarly, by applying Hoeffding's inequality, we have

$$P(|\hat{\mu}_j(t) - \mu_j| > \xi \mid N_j^Z(t) \geq a_{\min}\sqrt{t}) \leq 2\exp\left(-\frac{\xi^2 a_{\min}\sqrt{t}}{2}\right).$$

Therefore,

$$P(|\hat{\mu}_j(t) - \mu_j| > \xi) \leq \exp\left(-a_{\min}^2 t\right) + 2\exp\left(-\frac{\xi^2 a_{\min}\sqrt{t}}{2}\right),$$

and then,

$$P(\mathcal{E}_T^c(\epsilon)) \leq \sum_{t=T^{1/4}}^{T} \sum_{j\in[k]} P(|\hat{\mu}_j(t) - \mu_j| > \xi) \leq B_1 T \exp(-C_1 T^{1/8})$$

where B_1 and C_1 are constants and the functions of $\xi,\ a_{\min}$, and k.

Bounding $\mathcal{E}_T^{'c}(\epsilon)$. Similar to the previous part, we use the union bound over t, then we have

$$\begin{aligned} P(\mathcal{E}_T^{'c}(\epsilon)) &\leq \sum_{t=T^{\frac{1}{4}}}^{T} P\left(\left\|\mathbf{N}^Z(t) - \sum_{i\in[t]} \mathbf{p}(i)\right\|_2 > t^{\frac{3}{4}}\right) \\ &\leq \sum_{t=T^{\frac{1}{4}}}^{T} P\left(\sqrt{k}\left\|\mathbf{N}^Z(t) - \sum_{i\in[t]} \mathbf{p}(i)\right\|_\infty > t^{\frac{3}{4}}\right) \qquad (23) \\ &= \sum_{t=T^{\frac{1}{4}}}^{T} P\left(\left\|\mathbf{N}^Z(t) - \sum_{i\in[t]} \mathbf{p}(i)\right\|_\infty > \frac{t^{\frac{3}{4}}}{k^{\frac{1}{2}}}\right), \qquad (24) \end{aligned}$$

where we also used the inequality $\left\|\mathbf{N}^Z(t) - \sum_{i\in[t]} \mathbf{p}(i)\right\|_2 \leq \sqrt{k}\left\|\mathbf{N}^Z(t) - \sum_{i\in[t]} \mathbf{p}(i)\right\|_\infty$.

Again, after applying the union bound over $j \in [k]$, we have

$$P\left(\left\|\mathbf{N}^Z(t) - \sum_{i\in[t]} \mathbf{p}(i)\right\|_\infty > \frac{t^{\frac{3}{4}}}{k^{\frac{1}{2}}}\right) \leq \sum_{j=1}^{k} P\left(\left|N_j^Z(t) - \sum_{i\in[t]} p_j(i)\right| > \frac{t^{\frac{3}{4}}}{k^{\frac{1}{2}}}\right) \tag{25}$$

where $p_j(i)$ denotes the j-th element of the vector $\mathbf{p}(i)$. Now, we can write $N_j^Z(t)$ as the sum of t random variables as follows.

$$N_j^Z(t) = \sum_{s=1}^{t} \mathbb{1}_{(z_s=j)} = \sum_{i=1}^{t} W_i$$

Note that according to Algorithm 1, W_i is a Bernoulli random variable with parameter $\mathbf{p}(i)_i(s)$. Therefore, we have

$$P\left(\left|N_j^Z(t) - \sum_{i\in[t]} p_j(i)\right| > \frac{t^{\frac{3}{4}}}{k^{\frac{1}{2}}}\right) = P\left(\left|\sum_{i=1}^{t} W_i - \sum_{i\in[t]} p_j(i)\right| > \frac{t^{\frac{3}{4}}}{k^{\frac{1}{2}}}\right),$$

Then, by applying Hoeffding's inequality, we obtain

$$P\left(\left|\sum_{i=1}^{t} W_i - \sum_{i\in[t]} p_j(i)\right| > \frac{t^{\frac{3}{4}}}{k^{\frac{1}{2}}}\right) \leq 2\exp\left(-\frac{2t^{\frac{3}{2}}}{kt}\right) = 2\exp\left(-\frac{2t^{\frac{1}{2}}}{k}\right)$$

Finally, by combining the above inequality and Equations (23) and (25), we have

$$P(\mathcal{E}_T^{'c}(\epsilon)) \leq \sum_{t=T^{\frac{1}{4}}}^{T} 2k\exp\left(-\frac{2t^{\frac{1}{2}}}{k}\right) \leq TB_2\exp\left(-\frac{2T^{\frac{1}{8}}}{k}\right)$$

where $B_2 = 2k$ and $C_2 = \frac{2}{k}$.

All in all, by combining the derived upper bounds for $P(\mathcal{E}_T^c(\epsilon))$ and $P(\mathcal{E}_T^{'c}(\epsilon))$, we have

$$P((\mathcal{E}_T(\epsilon) \cap \mathcal{E}_T'(\epsilon))^c) \leq P(\mathcal{E}_T^c(\epsilon)) + P(\mathcal{E}_T^{'c}(\epsilon)) \leq BT\exp(-CT^{\frac{1}{8}}),$$

where $B = 2\max(B_1, B_2)$ and $C = \min(C_1, C_2)$.

G Discussion on Solving the Optimization Problem of Lower bound

The optimization problem (9) can be solved efficiently and with arbitrary precision. Notably, the same optimization problem arises in the classic BAI problem with Gaussian rewards [BGK22]. Although the definitions of Δ_i differ between these two settings—and in this paper, these gaps depend on the context probability matrix—the optimization problem itself remains similar, allowing the use of similar techniques to solve it.

Lemma G.1 ([BGK22])**.** *Consider the equation*

$$\sum_{\substack{i\in[n]\\ i\neq i^*(\boldsymbol{\mu})}} \frac{1}{(w\Delta_i^2 - 1)^2} = 1, \tag{26}$$

where $\Delta_i > 0$. This equation has a unique solution w^* in the interval $\left[\frac{2}{\Delta_{min}^2}, \frac{1+\sqrt{n-1}}{\Delta_{min}^2}\right]$. Define $\mathbf{u} = \left(\frac{w^*}{w^*\Delta_1^2-1}, \frac{w^*}{w^*\Delta_2^2-1}, \ldots, w^*, \ldots, \frac{w^*}{w^*\Delta_n^2-1}\right)$, where $i^*(\boldsymbol{\mu})$-th element is w^*. The normalized vector $\frac{\mathbf{u}}{\|\mathbf{u}\|_1} \in \Delta^{n-1}$ is the unique solution of Equation (9)*.*

As discussed in [BGK22], the left-hand side of Equation (26) is strictly decreasing over the interval $\left[\frac{2}{\Delta_{min}^2}, \frac{1+\sqrt{n-1}}{\Delta_{min}^2}\right]$. Consequently, simple numerical methods, such as binary search, can be employed to approximate the solution efficiently. Leveraging this property, we can efficiently compute the weights for the optimization problem in Equation (9). This capability enables the design of an efficient learning algorithm, as discussed later.

H Discussion on the Case of Unknown Context Probability Matrix

The assumption that the context probability matrix $\mathcal{A}$ is known does not always hold in practice. A natural question, then, is whether the problem becomes fundamentally harder, in terms of minimum sample complexity, when $\mathcal{A}$ is unknown. For instance, in the setting of contextual bandits, [NWRJ24] showed that the sample complexity of BAI with unknown context distributions remains of the same order as in the case with known distributions. This indicates that, in that setting, not knowing the context distribution does not significantly increase the difficulty of the problem.

However, the following lemma shows that this is not the case in the separator post-action context setting. In this setting, the absence of knowledge about $\mathcal{A}$ can arbitrarily increase the sample complexity lower bound, underscoring a fundamental difference from the contextual bandit setting.

Lemma H.1. *Consider the instance with $n = 3$, $k = 2$ and parameters*

$$\mathcal{A} = \begin{bmatrix} \epsilon & 2\epsilon & \frac{1}{2} \\ 1-\epsilon & 1-2\epsilon & \frac{1}{2} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{27}$$

where ϵ is a small positive real number. Then, we have

1. *On instance* (27) *with known $\mathcal{A}$, the characteristic $T_S^*(\boldsymbol{\mu}, \mathcal{A})$ defined in* (2) *is equal to* 8.
2. *For any δ-correct algorithm $(P_t, \tau, \hat{i})$ interacting with instance* (27) *with unknown $\mathcal{A}$, $\mathbb{E}_{\boldsymbol{\mu},\mathcal{A}}[\tau] \geq \frac{1}{4\epsilon} d_B(\delta, 1-\delta)$.*

Proof. To prove the first part, recall that

$$T_S^*(\boldsymbol{\mu}, \mathcal{A})^{-1} = \sup_{\mathbf{w} \in \Delta^2} \min_{i \neq i^*(\boldsymbol{\mu},\mathcal{A})} \frac{\Delta_i^2}{2 \sum_{j \in [k]} \frac{(\mathcal{A}_{j,i^*(\boldsymbol{\mu},\mathcal{A})} - \mathcal{A}_{j,i})^2}{\sum_{l \in [n]} w_l \mathcal{A}_{j,l}}}.$$

Note that $i^*(\boldsymbol{\mu}, \mathcal{A}) = 1$, $\Delta_2 = \epsilon$, and $\Delta_3 = \frac{1}{2} - \epsilon$. Moreover, for each $i \neq i^*(\boldsymbol{\mu}, \mathcal{A}), j \in [k]$, $\Delta_i^2 = (\mathcal{A}_{j,i^*(\boldsymbol{\mu},\mathcal{A})} - \mathcal{A}_{j,i})^2$. After inserting and simplification, we have

$$T_S^*(\boldsymbol{\mu}, \mathcal{A})^{-1} = \sup_{\mathbf{w} \in \Delta^2} \frac{1}{2\left(\frac{1}{w_1\epsilon + 2w_2\epsilon + \frac{w_3}{2}} + \frac{1}{w_1(1-\epsilon) + w_2(1-2\epsilon) + \frac{w_3}{2}}\right)} = \sup_{\mathbf{w} \in \Delta^2} \frac{w_{z,1} w_{z,2}}{2(w_{z,1} + w_{z,2})},$$

where $w_{z,1} = w_1\epsilon + 2w_2\epsilon + \frac{w_3}{2}$ and $w_{z,2} = w_1(1-\epsilon) + w_2(1-2\epsilon) + \frac{w_3}{2}$. Then setting $\mathbf{w} = (0, 0, 1)$ implies $w_{z,1} = w_{z,2} = \frac{1}{2}$, then

$$T_S^*(\boldsymbol{\mu}, \mathcal{A})^{-1} \geq \frac{1}{8}.$$

On the other hand, since $w_{z,1} + w_{z,2} = 1$, we have $w_{z,1} w_{z,2} \leq \frac{1}{4}$ and $\frac{w_{z,1} w_{z,2}}{2(w_{z,1} + w_{z,2})} \leq \frac{1}{8}$, which implies

$$T_S^*(\boldsymbol{\mu}, \mathcal{A})^{-1} \leq \frac{1}{8}.$$

Therefore, we conclude that $T_S^*(\boldsymbol{\mu}, \mathcal{A}) = 8$.

To prove the second part, note that Equation (1), by considering the instance $(\boldsymbol{\mu}', \mathcal{A}') \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ with parameters

$$\boldsymbol{\mu}' = \boldsymbol{\mu}, \quad \mathcal{A}' = \begin{bmatrix} 3\epsilon & 2\epsilon & \frac{1}{2} \\ 1-3\epsilon & 1-2\epsilon & \frac{1}{2} \end{bmatrix},$$

implies that for any δ-correct algorithm $(P_t, \tau, \hat{i})$, we have $\mathbb{E}_{\boldsymbol{\mu},\mathcal{A}}[\tau] \geq T^*(\mu, \mathcal{A}) d_B(\delta, 1-\delta)$, where

$$T^*(\boldsymbol{\mu}, \mathcal{A})^{-1} \leq \sup_{\mathbf{w} \in \Delta^2} \sum_{i \in \{1,2,3\}} w_i d\left(P_i^{\boldsymbol{\mu},\mathcal{A}}, P_i^{\boldsymbol{\mu}',\mathcal{A}'}\right).$$

Since these two instances only differ in the first column of their context probability matrix, then

$$\sup_{\mathbf{w} \in \Delta^2} \sum_{i \in \{1,2,3\}} w_i d\left(P_i^{\boldsymbol{\mu},\mathcal{A}}, P_i^{\boldsymbol{\mu}',\mathcal{A}'}\right) = \sup_{\mathbf{w} \in \Delta^2} w_1 d(\mathcal{A}_1, \mathcal{A}_1'),$$

where $d(\mathcal{A}_1, \mathcal{A}'_1)$ denotes the KL divergence between two binary variables and is equal to

$$d(\mathcal{A}_1, \mathcal{A}'_1) = \epsilon \ln\left(\frac{\epsilon}{3\epsilon}\right) + (1-\epsilon)\ln\left(\frac{1-\epsilon}{1-3\epsilon}\right) = -\epsilon\ln(3) + (1-\epsilon)\ln\left(\frac{1-\epsilon}{1-3\epsilon}\right).$$

We can upper bound $d(\mathcal{A}_1, \mathcal{A}'_1)$ as

$$\begin{aligned}
&\ln\left(\frac{1-\epsilon}{1-3\epsilon}\right) \leq \frac{1-\epsilon}{1-3\epsilon} - 1 = \frac{2\epsilon}{1-3\epsilon} \\
&\xrightarrow{1-\epsilon<2(1-3\epsilon)} d(\mathcal{A}_1, \mathcal{A}'_1) \leq -\epsilon\ln(3) + 4\epsilon < 4\epsilon \\
&\implies T^*(\boldsymbol{\mu}, \mathcal{A})^{-1} \leq \sup_{\mathbf{w}\in\Delta^2} w_1 d(\mathcal{A}_1, \mathcal{A}'_1) \leq 4\epsilon,
\end{aligned}$$

which completes the proof.

□

Plugging in an estimate of $\mathcal{A}$ is not sufficient. A natural idea is to estimate $\mathcal{A}$ online and to run our algorithms with the estimate $\hat{\mathcal{A}}_t$ in place of $\mathcal{A}$. This does not, in general, yield a correct or optimal algorithm. When $\mathcal{A}$ is unknown, alternative instances may change both the context distributions and the reward means, so the optimal allocation must balance the information collected about both $\mathcal{A}$ and $\boldsymbol{\mu}$. Optimizing as if $\hat{\mathcal{A}}_t$ were exact removes the KL cost of perturbing $\mathcal{A}$, and may assign zero samples to arms whose context distributions must still be verified. To see this, consider $n = k = 2$ with

$$\mathcal{A}_1 = (0.9, 0.1)^\top, \quad \mathcal{A}_2 = (0.6, 0.4)^\top, \quad \boldsymbol{\mu} = (0.9, 0.1)^\top.$$

Arm 1 is optimal, since $\mathcal{A}_1^\top\boldsymbol{\mu} = 0.82 > 0.58 = \mathcal{A}_2^\top\boldsymbol{\mu}$. If $\mathcal{A}$ is treated as known, the optimal allocation in Equation (2) samples only arm 2, because it produces the most balanced observations of the two contexts and is therefore the most informative about $\mu_1 - \mu_2$. In the unknown-$\mathcal{A}$ problem, however, consider the alternative instance with $\mathcal{A}'_1 = (0.5, 0.5)^\top$, $\mathcal{A}'_2 = \mathcal{A}_2$, and $\boldsymbol{\mu}' = \boldsymbol{\mu}$. In this instance, arm 2 is optimal, since $\mathcal{A}_1'^\top\boldsymbol{\mu}' = 0.5 < 0.58 = \mathcal{A}_2'^\top\boldsymbol{\mu}'$. Under the allocation $\mathbf{w} = (0, 1)$, this alternative has zero KL cost, because it only changes the context distribution of the arm that is never sampled. Hence, even if the estimate $\hat{\mathcal{A}}_t$ happens to equal the true $\mathcal{A}$, a valid algorithm for unknown $\mathcal{A}$ must still collect evidence against such alternatives, and the optimal allocations in the known-$\mathcal{A}$ and unknown-$\mathcal{A}$ problems can be far apart.

Now, we discuss the main challenge in designing optimal algorithms for the case of unknown $\mathcal{A}$. When $\mathcal{A}$ is unknown, the problem becomes more challenging because the set of alternative parameters (Alt) changes in structure. Recall that all tracking algorithms based on [GK16] must solve the optimization problem in Equation (1) to determine the optimal weights. Alternatively, one can use the formulation in Equation (17), but both approaches involve a minimization over some set. The main practical complexity of tracking algorithms lies in solving this optimization problem.

If an oracle were available to compute the optimal weights in Equation (1), one could still use tracking algorithms even when $\mathcal{A}$ is unknown and achieve the optimal sample complexity. Note that the GLR statistic and the stopping rule need to be adapted to this setting.

When $\mathcal{A}$ is unknown, both the matrices $(\boldsymbol{\mu}, \mathcal{A})$ should be learned by the agent during the learning process and if the error of estimation of each of them is high, the agent cannot stop. Let $\mathcal{I}^u$ denotes the set of parameter pairs $(\boldsymbol{\mu}, \mathcal{A})$ such that the best arm $i^*(\boldsymbol{\mu}, \mathcal{A})$ is unique. Then for a problem instance parameterized by $(\boldsymbol{\mu}, \mathcal{A})$, we define

$$\text{Alt}(\boldsymbol{\mu}, \mathcal{A}) = \{(\boldsymbol{\mu}', \mathcal{A}') \in \mathcal{I}^u \mid i^*(\boldsymbol{\mu}', \mathcal{A}') \neq i^*(\boldsymbol{\mu}, \mathcal{A})\}.$$

Similar to the case of known $\mathcal{A}$, $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ can be written as $\text{Alt}(\boldsymbol{\mu}, \mathcal{A}) = \cup_{i\neq i^*(\boldsymbol{\mu},\mathcal{A})}\mathcal{C}_i$, where

$$\mathcal{C}_i = \{(\boldsymbol{\mu}', \mathcal{A}') \in \mathcal{I}^u \mid \mathcal{A}_i'^\top\boldsymbol{\mu}'_i > \mathcal{A}'^\top_{i^*(\boldsymbol{\mu},\mathcal{A})}\boldsymbol{\mu}'_{i^*(\boldsymbol{\mu},\mathcal{A})}\}.$$

When $\mathcal{A}$ is known, the sets $\mathcal{C}_j$ or Alt are convex, so standard convex optimization tools can be applied. However, when $\mathcal{A}$ is unknown, the sets $\mathcal{C}_j$ or Alt, may not be convex, and thus finding optimal weights via convex optimization methods becomes infeasible. Similar issues have been noted in the reinforcement learning literature [AMP21].

As mentioned earlier, the case of separator context is a special case of non-separator setting, so it suffices to show non-convexity for the separator case. Note that if $\mathcal{A}$ is unknown, the set $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ changes because $\mathcal{A}$ itself can vary within the alternative set. Formally, the new definition of $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ for this setting is

$$\text{Alt}(\boldsymbol{\mu}, \mathcal{A}) = \{(\boldsymbol{\mu}', \mathcal{A}') \in \mathcal{I} \mid i^*(\boldsymbol{\mu}', \mathcal{A}') \neq i^*(\boldsymbol{\mu}, \mathcal{A})\},$$

where $\mathcal{I}$ denotes the set of all instances (parameterized by $(\boldsymbol{\mu}, \mathcal{A})$) that have a unique best arm. Similar to known setting, we have $\text{Alt}(\boldsymbol{\mu}, \mathcal{A}) = \cup_{i \neq i^*(\boldsymbol{\mu}, \mathcal{A})} \mathcal{C}_i$, and

$$\mathcal{C}_i = \{(\boldsymbol{\mu}', \mathcal{A}') \in \mathcal{I} \mid \mathcal{A}_i'^{\top} \boldsymbol{\mu}' > \mathcal{A}_{i^*(\boldsymbol{\mu}, \mathcal{A})}^{\top} \boldsymbol{\mu}\}.$$

As mentioned earlier, we need to solve

$$\inf_{(\boldsymbol{\mu}', \mathcal{A}') \in \text{Alt}(\boldsymbol{\mu}, \mathcal{A})} \sum_{i \in [n]} w_i d(P_i^{\boldsymbol{\mu}, \mathcal{A}}, P_i^{\boldsymbol{\mu}', \mathcal{A}'}),$$

or equivalently, using Equation (17),

$$f_j(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A}) = \inf_{(\boldsymbol{\mu}', \mathcal{A}') \in \mathcal{C}_j} \sum_{i \in [n]} w_i d(P_i^{\boldsymbol{\mu}, \mathcal{A}}, P_i^{\boldsymbol{\mu}', \mathcal{A}'}). \tag{28}$$

For the case of known $\mathcal{A}$, each set $\mathcal{C}_j$ is convex, so $f_j(\mathbf{w}, \boldsymbol{\mu}, \mathcal{A})$ reduces to a convex problem that can be solved via convex optimization methods. However, if $\mathcal{A}$ is unknown, that convexity property no longer holds. The following example shows that neither $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ nor $\mathcal{C}_j$ is necessarily convex.

Example. Consider $n = k = 2$. Suppose $i^*(\boldsymbol{\mu}, \mathcal{A}) = 1$, then $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ equals $\mathcal{C}_2$. Therefore, we only need to show that $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ is not convex. Let

$$\mathcal{A} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}, \ \boldsymbol{\mu} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \tag{29}$$

Clearly, $i^*(\boldsymbol{\mu}, \mathcal{A}) = 1$ because $\mathbb{E}[Y \mid X = 1] = 2.6$ and $\mathbb{E}[Y \mid X = 2] = 2.4$. Next, let

$$\mathcal{A}^{(1)} = \begin{bmatrix} 0.4 & 0.2 \\ 0.6 & 0.8 \end{bmatrix}, \ \boldsymbol{\mu}^{(1)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \tag{30}$$

$$\mathcal{A}^{(2)} = \begin{bmatrix} 0.1 & 0.2 \\ 0.9 & 0.8 \end{bmatrix}, \ \boldsymbol{\mu}^{(2)} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}. \tag{31}$$

We see $i^*(\boldsymbol{\mu}^{(1)}, \mathcal{A}^{(1)}) = i^*(\boldsymbol{\mu}^{(2)}, \mathcal{A}^{(2)}) = 2$. Therefore, $(\boldsymbol{\mu}^{(1)}, \mathcal{A}^{(1)})$ and $(\boldsymbol{\mu}^{(2)}, \mathcal{A}^{(2)})$ belong to $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$. Consider a convex combination of $(\boldsymbol{\mu}^{(1)}, \mathcal{A}^{(1)})$ and $(\boldsymbol{\mu}^{(2)}, \mathcal{A}^{(2)})$ as

$$\mathcal{A}^{(3)} = \frac{1}{2}(\mathcal{A}^{(1)} + \mathcal{A}^{(2)}) = \begin{bmatrix} 0.25 & 0.2 \\ 0.75 & 0.8 \end{bmatrix}, \boldsymbol{\mu}^{(3)} = \frac{1}{2}(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}) = \begin{bmatrix} 3 \\ 2 \end{bmatrix}. \tag{32}$$

Now, we have $i^*(\boldsymbol{\mu}^{(3)}, \mathcal{A}^{(3)}) = 1$. It indicates that $(\boldsymbol{\mu}^{(3)}, \mathcal{A}^{(3)})$ is *not* in $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$, which shows that $\text{Alt}(\boldsymbol{\mu}, \mathcal{A})$ is not a convex set.

I Further Experiments

This section contains additional details on the experimental setup and the results of further experiments for both non-separator and separator settings.

I.1 Real-World Data Experiment

We also evaluate our methods on instances constructed from the *KuaiSAR* recommendation logs [SSZ+23], a real-world short video recommendation dataset containing user–item interaction records together with feedback signals such as clicks, likes, and watch behavior. The goal of these experiments is not to provide a large-scale benchmark, but rather to verify on real interaction data that the empirical behavior is consistent with the theoretical predictions of our framework.

Table 2: Real-world data experiments based on KuaiSAR [SSZ+23]. Reported values are average stopping times over 20 independent runs. The separator rows correspond to the semi-synthetic instance described above.

Setting	Algorithm	Avg. stopping time
Non-separator (KuaiSAR, 7 actions)	TS	16355
Non-separator (KuaiSAR, 7 actions)	NSTS (ours)	**6973**
Separator (KuaiSAR semi-synthetic, 7 actions)	STS (ours)	**411**
Separator (KuaiSAR semi-synthetic, 7 actions)	TS	> 50000
Separator (KuaiSAR semi-synthetic, 7 actions)	LTS	> 50000

For the non-separator setting, we construct a BAI instance directly from the real data. We use first level content categories as actions, resulting in $n = 7$ actions with sufficient support. Concretely, each action corresponds to recommending an item from one of these content categories, so pulling an arm means selecting a category and then observing the feedback generated by an interaction record associated with that category. Operationally, whenever the learner selects an action, one row of the dataset corresponding to that action is sampled uniformly at random, and the observed post-action context and reward are taken from that sampled interaction record. As post-action context, we consider three discrete feedback outcomes observed after recommendation exposure: (i) no click, (ii) click only, and (iii) click+like. As reward, we use the clipped watch-ratio

$$R = \min\left(\frac{\text{playing time}}{\text{duration}}, 3\right),$$

where *playing time* is the total time the user spent watching the recommended video and *duration* is the nominal video length. This ratio is a standard normalized engagement measure; it can exceed 1 because users may replay or rewatch parts of a video, so the accumulated playing time can be longer than the video duration. We clip the ratio at 3 to control extreme values while preserving meaningful variation in engagement. To obtain a statistically meaningful BAI instance, we retain actions with sufficient support across all three context values and enforce a minimum gap between the best and second best arms. This yields a genuine non-separator instance: even after conditioning on the post-action context, the expected reward still depends on the action.

For the separator setting, we start from the same real interaction data and use it to estimate the post-action context distribution of each action. We then construct rewards so that they depend only on the realized context. More precisely, for each context value z, we compute its empirical mean reward μ_z from the data and generate rewards as

$$Y = \mu_z + \mathcal{N}(0, 1).$$

Hence, the action still influences which context is observed, but conditional on the context, the reward is independent of the action. We refer to this separator instance as *semi-synthetic*: it is not fully raw real-world data, but it remains tightly grounded in the dataset because both the action-dependent context probabilities and the context level reward means are estimated from real user interactions.

For both settings, we report the average stopping time over 20 independent runs. For the TS baseline on the non-separator instance, since the reward is bounded in $[0, 3]$, we use the Gaussian stopping rule with variance proxy $\sigma^2 = (3 - 0)^2/4 = 9/4$, corresponding to the standard sub-Gaussian parameter guaranteed by Hoeffding's lemma for bounded random variables. The results are shown in Table 2. In the non-separator setting, NSTS significantly improves over TS. In the semi-synthetic separator setting, STS dramatically outperforms context-ignoring baselines, which fail to stop within the horizon of 50,000 rounds. These results are consistent with the main message of the paper: exploiting post-action context leads to substantial gains when used in a manner aligned with the underlying problem structure.

I.2 Non-Separator Context

We generate random instances for each combination of $n \in \{5, 10, 15\}$ and $k \in \{3, 5, 7\}$. Each instance has n arms and k context values, with rewards drawn from Gaussian distributions of unit variance. The mean matrix $\boldsymbol{\mu}$ lies in $[0, 10]^{k \times n}$, and $a_{\min} = \min_{i,j} \mathcal{A}_{j,i} \geq \frac{1}{4k}$. Without loss of generality, we assume the first arm is the best. We randomly generate the matrices $\boldsymbol{\mu}$ and $\mathcal{A}$, ensuring

that for each $i > 1$, $\Delta_i \in [\frac{1}{2n}, \frac{i+1}{2n}]$. This constraint prevents unchallenging instances in which all gaps are large. Throughout our experiments, we set the confidence parameter δ to 0.1.

We compare our algorithm (NSTS) with the classic track-and-stop (TS) algorithm, which disregards the context variable and instead operates solely on the rewards with D-tracking. Note that the TS does not have a theoretical guarantee in this setting because each arm's reward distribution is a mixture of Gaussians, which does not belong to the one-parameter exponential family required for theoretical guarantees of TS [GK16]. Note that as the expected values of the rewards are in $[0, 10]$, it can be shown that the rewards for each arm follow a sub-Gaussian distribution. We then use this property to design a stopping rule. The details are given below.

Table 3 reports the average number of rounds required by each algorithm before stopping, aggregated over 75 runs. The results indicate that leveraging post-action contexts can significantly improve performance.

Table 3: Average stopping times of NSTS and TS algorithms for different values of n and k. The Ratio column shows TS/NSTS.

n	$k=3$			$k=5$			$k=7$		
	NSTS	TS	Ratio	NSTS	TS	Ratio	NSTS	TS	Ratio
5	30635	427872	14.0	31515	308976	9.8	60960	534147	8.8
10	92034	1244462	13.5	224236	2090792	9.3	320224	2433472	7.6
15	272064	3667765	13.5	425685	3768745	8.9	428615	3573301	8.3

Before presenting further details on the results, we discuss the sequential thresholds used in the algorithms. For NSTS, we used the exact thresholds from (11), though they tend to be overly conservative, keeping the error probability well below δ.

As discussed earlier, there is no theoretical guarantee for TS when interacting with a bandit featuring post-action context. This is because, in such cases, the reward distribution of each arm becomes a mixture of Gaussians, which does not belong to the exponential family of distributions. Moreover, experimental results indicate that applying the same thresholds as in the classic bandit setting with unit-variance Gaussian rewards does not control the error rate below δ (refer to Appendix B.3 for a detailed discussion).

To design new thresholds for TS, we utilized Lemma of [RKB$^+$21]. Since $\boldsymbol{\mu} \in [0, 10]^{k \times n}$, the reward distribution of the arms is sub-Gaussian with parameter $\sigma = \sqrt{26}$. For employing the TS algorithm, we approximate the reward distribution with a Gaussian distribution with variance $\sigma^2 = 26$. Consequently, we can use the known threshold for classic bandit with Gaussian distribution [KK21].

Figure 6 shows the average L^2 distance between the vector $\frac{\mathbf{N}^X(t)}{t}$ and the optimal vector $\mathbf{w}^*(\boldsymbol{\mu}, \mathcal{A})$ over time for all nine instances introduced in the main text simulated over 75 runs. For each run, after its stopping time, the distance is considered as zero in the average. The results indicate that ignoring the post-action context leads to significant sub-optimality and misalignment in tracking the correct proportion of actions.

I.3 Separator Context

Here, we present additional results from running different algorithms on instances in the separator setting. For each $n \in \{5, 10, 15\}$ and $k \in \{3, 5, 8\}$, we randomly generated an instance with n arms and k context values. To avoid trivial instances, we repeatedly generated instances until finding one where $\Delta_i \in \left[\frac{1}{2n}, \frac{i+1}{2n}\right]$, with the first arm being the best arm.

As shown in the main text, the LTS algorithm does not perform competitively, and due to its long computation time, we do not report its results on these instances and only compare TS and Algorithm 1. Figure 7 presents a box plot of the stopping times for these algorithms across all instances. The results demonstrate that ignoring information about the post-action context leads to significant sub-optimality.

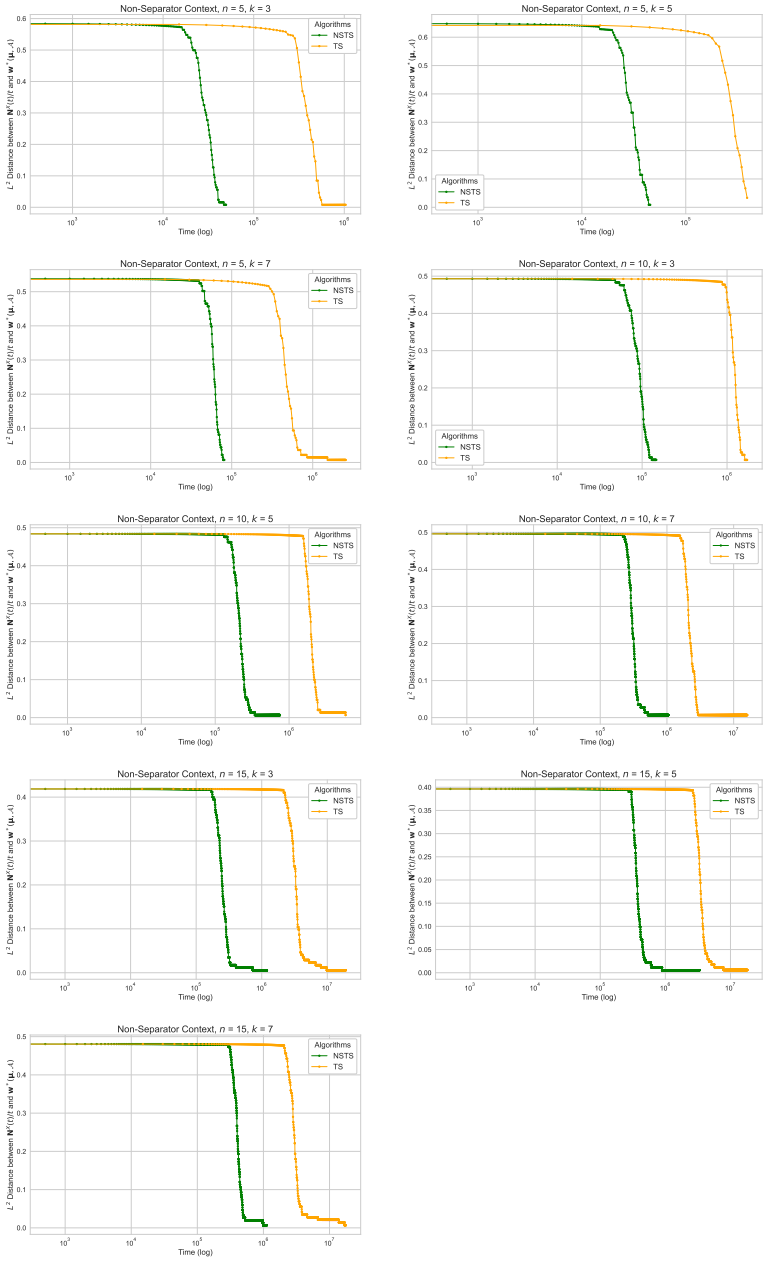


Figure 6: Comparison of the L^2 distance of the frequencies of pulled arms and the optimal frequency over time between two algorithms.

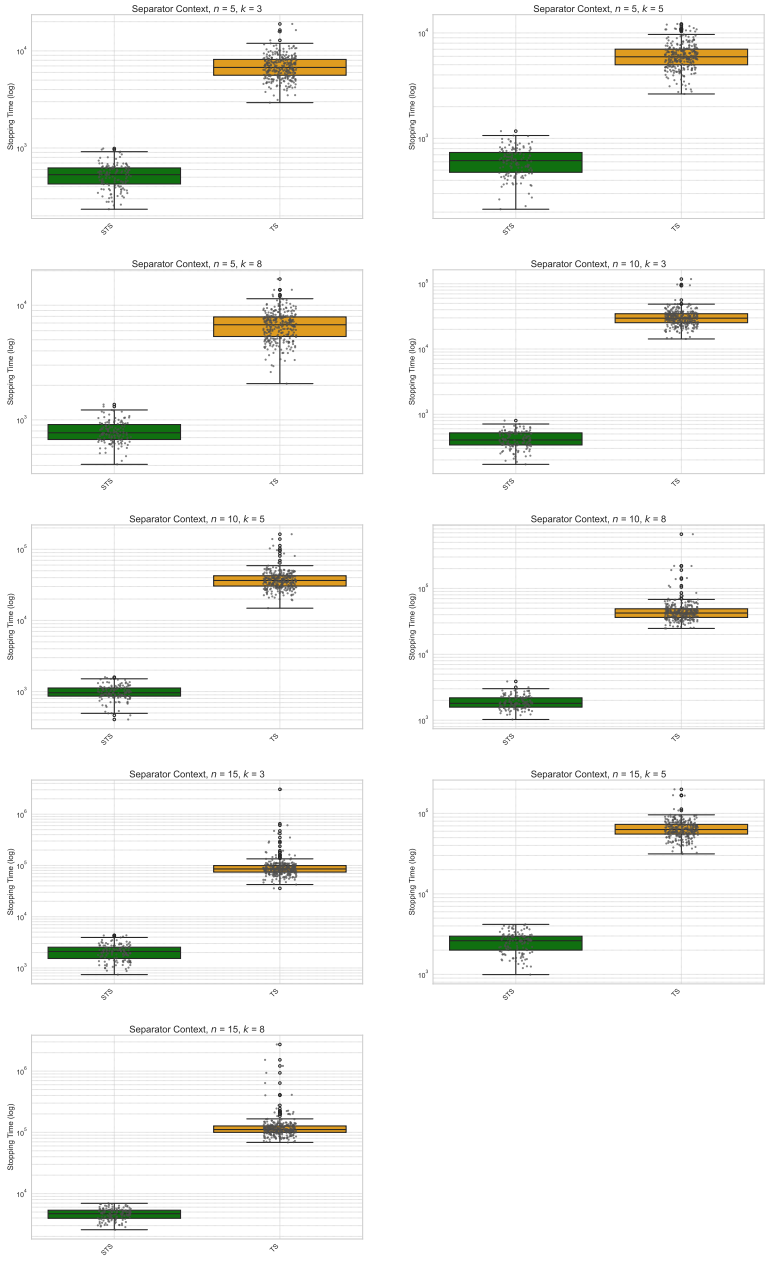


Figure 7: Comparison of the stopping times of different algorithms on randomly generated instances.

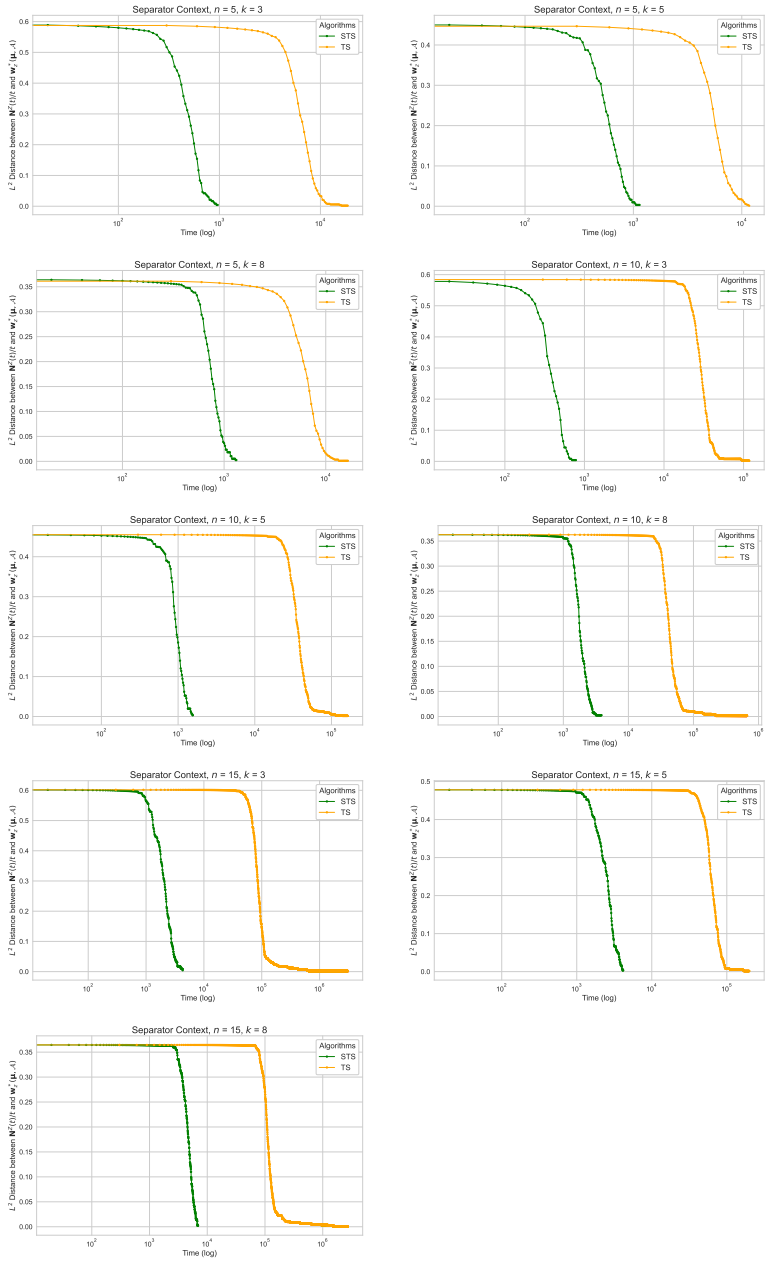


Figure 8: Comparison of the L^2 distance of the frequencies of observed contexts and the optimal frequency over time among different algorithms.

Figure 8 shows the average L^2 distance of the context frequencies vector $\frac{\mathbf{N}^z(t)}{t}$ and the optimal context frequencies $\mathbf{w}_z^*(\boldsymbol{\mu}, \mathcal{A})$ over time for all previously introduced instances and for different algorithms.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims in abstract and introduction have been discussed in Sections 3, 4, 5, and also appendices.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 2, we have provided all the assumptions we needed. We also discuss about the complexity without making some of the asssumptions in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Section 2 presents all the assumptions we need to prove our results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation of our algorithm is straightforward and is discussed in related sections. Also, for using other approaches, Appendix I contains the necessary information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will include the code in supplementary material. We will add a link to Github repository in the camera-ready version.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 5 and Appendix I, we discuss about experiment setups.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Refer to Figures 3 and 7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiments are simple to run and do not require any specific computational resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper adheres fully to the NeurIPS Code of

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: There is no societal impact as our work primarily focuses on theoretical analysis of a problem in online decision-making (bandits).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.