Pure Exploration for Constrained Best Mixed Arm Identification with a Fixed Budget

Dengwang Tang, Rahul Jain, Ashutosh Nayyar, Pierluigi Nuzzo

Keywords: Constrained mixed arm identification, constrained bandit problem.

Summary

We introduce the constrained best mixed arm identification (CBMAI) problem under unknown reward and costs wherein there are K arms, each of which is associated with a reward and multiple cost attributes. These are random, and come from distributions with unknown means. The best mixed arm is a probability distribution over a subset of the K arms that maximizes the expected reward while satisfying the expected cost constraints. We are specifically interested in a pure exploration problem under a fixed sampling budget with the goal of identifying the *support of the best mixed arm*. We propose a novel, parameter-free algorithm, called the Score Function-based Successive Reject (SFSR) algorithm, that combines the classical successive reject framework with a novel rejection criteria using a score function based on linear programming theory. We establish a performance guarantee for our algorithm by providing a theoretical upper bound on the probability of mis-identification of the support of the best mixed arm and show that it decays exponentially in the budget N and some constants that characterize the hardness of the problem instance. We also develop an information-theoretic lower bound on the error probability that shows that these constants appropriately characterize the problem difficulty. We validate this empirically on a number of problem instances.

Contribution(s)

1. This paper provides a novel, parameter-free algorithm that identifies the optimal support of the best mixed arm for a constrained best arm identification problem with a fixed sampling budget. We establish a performance guarantee for our algorithm in the form of an exponentially decaying, instance-dependent error upper bound.
   Context: Prior work has considered the best arm identification problem with *known costs* and/or with only *deterministic* arms allowed. However, we consider unknown costs and allow randomized (i.e., mixed) arms since deterministic arms may be suboptimal, or simply not meet all the constraints.

Pure Exploration for Constrained Best Mixed Arm Identification with a Fixed Budget

Dengwang Tang, Rahul Jain, Ashutosh Nayyar, Pierluigi Nuzzo
dwtang@umich.edu, {rahul.jain,ashutosh.nayyar}@usc.edu, nuzzo@eecs.berkeley.edu

[1]**eBay Inc.**
[2,3]**Electrical and Computer Engineering, University of Southern California**
[4]**Electrical Engineering and Computer Sciences, University of California, Berkeley**

Abstract

We introduce the constrained best mixed arm identification (CBMAI) problem under unknown reward and costs, wherein there are K arms, each of which is associated with a reward and multiple cost attributes. These are random, and come from distributions with unknown means. The best mixed arm is a probability distribution over a subset of the K arms that maximizes the expected reward while satisfying the expected cost constraints. We are specifically interested in a pure exploration problem under a fixed sampling budget with the goal of identifying the *support of the best mixed arm*. We propose a novel, parameter-free algorithm, called the Score Function-based Successive Reject (SFSR) algorithm, that combines the classical successive reject framework with a novel rejection criteria using a score function based on linear programming theory. We establish a performance guarantee for our algorithm by providing a theoretical upper bound on the probability of mis-identification of the support of the best mixed arm and show that it decays exponentially in the budget N and some constants that characterize the hardness of the problem instance. We also develop an information-theoretic lower bound on the error probability that shows that these constants appropriately characterize the problem difficulty. We validate this empirically on a number of problem instances.

1 Introduction

Bandit models are prototypical models of online learning, exploration, and decision making (Lattimore & Szepesvári, 2020). For example, recommender systems for online shopping and video streaming often use learning algorithms to make recommendations that maximize click-through-rates. Online learning can be formulated as a regret minimization problem, which leads to algorithms that trade exploration with exploitation (Lai & Robbins, 1985), where the regret of a learning algorithm is defined with respect to a policy optimizing a single (reward) objective. A number of Bayesian and non-Bayesian algorithms (Lattimore & Szepesvári, 2020) have been proposed for this setting (Auer et al., 2002; Russo & Van Roy, 2014). Alternatively, online learning can be formulated as a pure exploration problem, also referred to as the best arm identification problem (Audibert & Bubeck, 2010; Garivier & Kaufmann, 2016), wherein adaptive data collection can be performed to identify the optimal arm without considering the rewards/performance during learning.

Many online learning problems involve multiple objectives that cannot be aggregated into a single objective function. Such problems are better formulated in terms of maximizing one objective while constraining the others. In the recent literature, progress has been made on constrained bandit models, as well as online reinforcement learning with constraints. In such exploration vs. exploitation settings, novel algorithms have emerged that can surprisingly minimize regret while ensuring

bounded constraint violation (Kalagarla et al., 2023). However, some practical settings are more suitable for a pure exploration problem rather than a regret minimization problem. Constrained online learning within a pure exploration problem formulation, and specifically, the constrained best arm identification problem is largely unsolved.

In this paper, we introduce the *constrained best mixed arm identification* (CBMAI) problem, wherein there are K arms, each of which is associated with a reward and multiple cost attributes. These are random, and come from distributions with unknown means. The best mixed arm is a probability distribution over a subset of the K arms that maximizes the expected reward while satisfying the expected cost constraints. We are specifically interested in a pure exploration problem with the goal of identifying the *support of the best mixed arm*, i.e., identifying the subset of arms with non-zero probability in the best mixed arm. Since the true expected costs are unknown in our setting, the exact shape of the constraint polytope is not known in advance. Thus, there are uncountably many candidates for the best mixed arm. However, the support of the best mixed arm is still one of finitely many (albeit exponentially many) possibilities. Given a fixed sampling budget N, we can sample the arms in any way and then use the gathered data (rewards/costs) to declare which arms are in the support of the best mixed arm.

We provide a novel, parameter-free algorithm, called the Score Function-based Successive Reject (`SFSR`) Algorithm, that identifies the optimal support for the CBMAI problem under the fixed-budget setting. The algorithm combines the successful Successive Reject algorithm (Audibert & Bubeck, 2010) with a novel rejection criteria using a score function based on linear programming theory. Using different score functions results in an algorithm with different flavors, and we present two choices for them. We establish a performance guarantee of the proposed algorithm in the form of an instance-dependent error upper bound, which decays exponentially in N with an exponent characterized by a certain measure of hardness of the instance. We also present a lower bound on the error probability for a broad class of algorithms which helps validate how good our upper bound is. We provide empirical results to show that our proposed algorithm significantly outperforms baselines on typical instances.

2 Related Literature

We review related literature on best arm identification and constrained learning problems which is somewhat distinct from the vast literature on multi-arm bandits (Lattimore & Szepesvári, 2020).

Best Arm Identification. The literature on (unconstrained) best arm identification can be divided into two categories: (1) The fixed confidence setting (Kaufmann & Kalyanakrishnan, 2013; Jamieson et al., 2014; Garivier & Kaufmann, 2016; Russo, 2016; Qin et al., 2017), where the aim is to identify the best arm with a specified error probability δ using the smallest number of samples. Two major algorithm design philosophies in this setting are Top-2 algorithms (Russo, 2016) and Track-and-Stop (Kaufmann & Kalyanakrishnan, 2013). (2) The fixed budget setting (Audibert & Bubeck, 2010; Karnin et al., 2013; Carpentier & Locatelli, 2016; Yang & Tan, 2022; Barrier et al., 2023; Wang et al., 2024), where the aim is to minimize the identification error probability given a sampling budget N. Round-based elimination algorithms are the dominant algorithm philosophy in this setting.

Regret-focused Learning in Constrained Problems. There has been a lot of recent literature on designing algorithms to achieve small reward and/or constraint violation regret in constrained multi-armed bandits (Amani et al., 2019; Moradipari et al., 2021; Liu et al., 2021b; Zhou & Ji, 2022; Pacchiano et al., 2024) and constrained MDPs (Liu et al., 2021a; Bura et al., 2022; Kalagarla et al., 2023). In these settings, it is necessary to balance exploration and exploitation. In contrast, we are interested in efficiently using the learning budget for pure exploration.

Constrained Best Arm Identification. Recently, there has been considerable interest in best arm identification in a constrained setting. In Lindner et al. (2022); Camilleri et al. (2022); Wang et al. (2022); Faizal & Nair (2022); Shang et al. (2023), the authors considered the best *deterministic*

arm identification problem from a finite set of arms in either fixed-confidence or fixed-budget settings. In contrast, our work focuses on finding the best *mixed* arm since deterministic arms may be suboptimal, or simply not meet the constraints. There are very few works on best mixed arm identification: The CBMAI problem under the *fixed-confidence* setting when the *costs are known* was considered in Carlsson et al. (2023). In contrast, we assume that the costs are unknown. Nakamura & Sugiyama (2024) considered a fixed-budget best knapsack identification problem assuming that the best solution belongs to a known finite set and there's an offline oracle for finding the best knapsack under constraints given an input reward function. We do not assume access to such an oracle due to the costs being unknown in our setting. A fixed-budget optimal support identification problem with the same constraints imposed on both the exploration process and the final solution was considered in Li et al. (2023). In contrast, we do not impose any constraints on the exploration process. Furthermore, the theoretical error bounds therein are unfortunately not correct since the strong concentration results for optimal solutions of randomly perturbed linear programs in their Lemmas B.1 and C.2, which are a critical part of the theoretical analysis, are erroneous. Kone et al. (2023) considered the problem of Pareto front identification for arms with multiple attributes. There have also been numerous works on constrained Bayesian Optimization (Gardner et al., 2014; Gelbart et al., 2014; Letham et al., 2019; Eriksson & Poloczek, 2021) where the primary focus is on empirical performance instead of theoretical guarantees.

3 Preliminaries

Notation. For a positive integer M, $[M] := \{1, 2, \cdots, M\}$. For a vector $v \in \mathbb{R}^M$ and $\mathcal{I} \subset [M]$, $v_{\mathcal{I}}$ denotes the sub-vector of v formed by indices in $\mathcal{I}$. For a matrix A with M columns and $\mathcal{I} \subset [M]$, $A_{\mathcal{I}}$ denotes the sub-matrix of A formed by columns indexed by $\mathcal{I}$. $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. $\|\cdot\|_2$ represents the Euclidean 2-norm.

Problem Statement. We introduce the *constrained best mixed arm identification* (CBMAI) problem in the context of a bandit model: There are K arms, indexed by $[K] = \{1, 2, \cdots, K\}$, each associated with a reward function R_a and L cost functions $C_{l,a}$, $a \in [K], l \in [L]$. In this setting, since any deterministic arm may be suboptimal, or fail to meet the constraints, we would like to determine a mixed arm p^*, i.e., a probability distribution over the arms (in the probability simplex $\mathcal{P}_K := \{p \in \mathbb{R}_+^K : \mathbf{1}^T p = 1\}$) such that it achieves the following

$$\max_{p \in \mathcal{P}_K} \{\mathbf{R}^T p : \mathbf{C} p \leq \bar{c}\}, \tag{1}$$

where $\mathbf{R} = (R_1, \cdots, R_K)^T$, $(\mathbf{C})_{l,a} = C_{l,a}$ and the vector $\bar{c} \in \mathbb{R}^L$. This is a linear program, and therefore an optimal solution of (1) can be obtained at an extreme point of the constraint polytope (Luenberger et al., 1984). Thus, when the costs are known, the best mixed arm will lie in a known finite set (Carlsson et al., 2023; Nakamura & Sugiyama, 2024) since the constraint polytope has a finite number of vertices.

However, our motivation comes from the bandit setting, and typically both the rewards R and costs C for each arm are random, and come from an unknown distribution. In that case, there can be uncountably many candidates for the best mixed arm making identifying the exact best mixed arm virtually impossible. Thus, we focus on the *optimal support identification*, i.e., identifying the arms that have non-zero probability in the best mixed arm. We denote such a set of arms by $\mathcal{I}^*$ (we will define it precisely later). We will consider that when the learning agent chooses arm a, it receives a reward $R_a \sim \mathcal{N}(r_a, \sigma_r^2)$, and also incurs costs $C_{l,a} \sim \mathcal{N}(c_{l,a}, \sigma_c^2)$, $l = 1, 2, \cdots, L$. The random rewards and costs are assumed mutually independent. We will assume that for the first K_0 arms, the mean reward $(r_a)_{a \in [K_0]}$ and mean costs $(c_{l,a})_{a \in [K_0], l \in [L]}$ are unknown. The means of the reward and costs of an arm in the (possibly empty) subset $\{K_0 + 1, \cdots, K\}$ is assumed to be known. Note that the traditional setting of all arms being unknown simply corresponds to the case where $K_0 = K$. The variances are not needed to be known by the algorithms we design, but assuming them known will simplify our analysis. Thus, we would like to solve the following LP problem that optimizes

the expected reward subject to constraints on expected costs

$$\max_{p \in \mathcal{P}_K} \{\mathbf{r}^T p : \mathbf{c}p \leq \bar{c}\}, \tag{LP}$$

where the components of $\mathbf{r} = (r_a)_{a \in [K]}$ and $\mathbf{c} = (c_{l,a})_{l \in [L], a \in [K]}$ corresponding to the first K_0 arms are unknown and must be learnt.

To that end, we need samples of reward and costs for various arms. We consider a *fixed-budget* setting, i.e., we can only obtain *at most* N such samples. We assume an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space, $\mathcal{F}$ is the event space, and $\mathbb{P}$ is a probability measure, and would like to design a learning agent ϕ that minimizes the *misidentification probability* $\mathbb{P}_{\mathbf{r},\mathbf{c}}(\mathcal{X}^\phi \neq \mathcal{I}^*)$, i.e., the probability of misidentifying the optimal support $\mathcal{I}^*$, where $\mathcal{X}^\phi$ is the subset of arms output by the algorithm. Note that we use a *strict criteria*: we only consider the identification to be correct if the output support is *exactly* the set of arms in the optimal support.

Remark 1. We make a few observations. (i) Our algorithm does not need the Gaussian distributions assumption. Furthermore, our main result (Theorem 1) can be adapted to sub-Gaussian reward and cost distributions (through replacing the use of Gaussian concentration inequalities in the proof with sub-Gaussian ones). However, assuming Gaussian distributions allows us to focus on presenting key ideas while avoiding unnecessary technical details. (ii) The reasoning for explicitly formulating known arms in our setting is as follows: First, for best *deterministic arm* identification problem with both unknown and known arms, constrained or not, one can always run any algorithm on the subset of unknown arms, obtain the best arm in this subset, and compare it with the best known arm, *making the formulation of known arms a bit of a distraction*. However, this is not the case for constrained best *mixed* arm identification, as the addition of a known arm could introduce new unknown arms into the optimal mixed arm. (iii) If the optimal support is known, one can easily finetune the mixing probabilities with online data and quickly converge to the best mixed arm without the need to explore arms outside of the support. However, finding the optimal support can be a challenging problem due to its combinatorial nature.

4 The Score Function-based Successive Reject (`SFSR`) Algorithm

We first derive our main algorithm, the *Score Function-based Successive Reject* (`SFSR`) algorithm, that uses a novel elimination rule we designed based on the *intersection value* (IV) score. We will later show that substituting this score function with another results in a different flavor of the algorithm that can also have good empirical performance.

Consider the standard form of (LP) where we add the slack vector $s \in \mathbb{R}_+^L$:

$$\max_{p \in \mathbb{R}_+^K, s \in \mathbb{R}_+^L} \{\mathbf{r}^T p : \mathbf{c}p + s = \bar{c}, \mathbf{1}^T p = 1\}. \tag{2}$$

Let $\mathbf{0}_m$ (resp. $\mathbf{1}_m$) denote the all-0 vector (resp. all-one vector) of length m. Let $\mathbf{I}_{L \times L}$ denote the L-by-L identity matrix. Set

$$\mathbf{x} = \begin{pmatrix} p \\ s \end{pmatrix}, \quad \mu = \begin{pmatrix} \mathbf{r} \\ \mathbf{0}_L \end{pmatrix}, \tag{3}$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{c} & \mathbf{I}_{L \times L} \\ \mathbf{1}_K^T & \mathbf{0}_L^T \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \bar{c} \\ 1 \end{pmatrix}, \tag{4}$$

then (2) can be simply written as

$$\max\{\mu^T \mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\} \tag{SFLP}$$

An optimal solution of the linear program in (SFLP) can be obtained at a *basic feasible solution* (BFS) (Luenberger et al., 1984) of $(\mathbf{A}, \mathbf{b})$ which is determined by a *basis* $\mathcal{I}^* \subset [K + L]$.

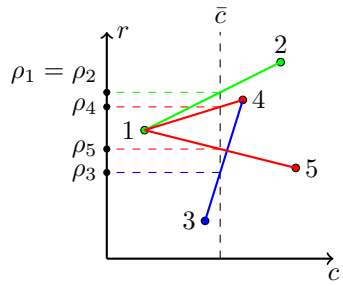

Figure 1: Intersection value scores for a 5-arm, 1-constraint instance

Definition 1 (BFS)**.** Let $\mathbf{A}$ be an $M \times R$ matrix and $\mathbf{b} \in \mathbb{R}^M$. A subset $\mathcal{I}$ of $[R]$ is said to be a *basis* of $\mathbf{A}$, if $|\mathcal{I}| = M$. A non-negative vector $\mathbf{x}^* \in \mathbb{R}_+^N$ is said to be a *basic feasible solution* (BFS) of $(\mathbf{A}, \mathbf{b})$ corresponding to the *basis* $\mathcal{I}$, if (i) $\mathbf{x}_i^* = 0$ for $i \notin \mathcal{I}$; (ii) the square sub-matrix $\mathbf{A}_\mathcal{I}$ is invertible; and (iii) $\mathbf{x}_\mathcal{I}^* = \mathbf{A}_\mathcal{I}^{-1}\mathbf{b} \geq 0$. In this case, $\mathcal{I}$ is called a *feasible basis* of $(\mathbf{A}, \mathbf{b})$.

For the ease of describing the algorithm, we will refer to the slack variable s_l corresponding to the l-th constraint as "arm $K + l$." With this convention, arms $K + 1$ to $K + L$ can be thought of as "virtual arms" corresponding to slack variables. The proposed algorithm starts with the set of all arms $[K + L]$ and successively rejects one arm in each round. The algorithm returns when either it believes the problem is infeasible, or when there are only $L + 1$ arms remaining.

The choice of which arm to eliminate at each round is determined by a *score function* (like an index in bandit learning algorithms) for each arm computed from the empirical estimates of their rewards and costs: Let $\mathcal{X} \subset [K + L]$ to the subset of remaining arms, Let $\hat{\mu}_\mathcal{X} \in \mathbb{R}^\mathcal{X}$ be the empirical mean vector defined by

$$\hat{\mu}_\mathcal{X} = (\hat{\mu}_a)_{a \in \mathcal{X}}, \qquad \hat{\mu}_a = \begin{cases} \hat{r}_a & a \leq K_0 \\ \mu_a & \text{otherwise} \end{cases}, \tag{5}$$

where $\hat{r}_a$ is the empirical mean reward of arm a.

Let the empirical constraint sub-matrix $\hat{\mathbf{A}}_\mathcal{X} \in \mathbb{R}^{[L+1] \times \mathcal{X}}$ be defined by

$$\begin{aligned} \hat{\mathbf{A}}_\mathcal{X} &= (\hat{\mathbf{A}}_{l,a})_{l \in [L+1], a \in \mathcal{X}}, \\ \hat{\mathbf{A}}_{l,a} &= \begin{cases} \hat{c}_{l,a} & l \leq L, a \leq K_0 \\ \mathbf{A}_{l,a} & \text{otherwise} \end{cases}. \end{aligned} \tag{6}$$

The *intersection value (IV) score function* for arm $a \in \mathcal{X}$ is then defined as

$$\begin{aligned} f_a^{\mathrm{IV}}(\hat{\mathbf{r}}, \hat{\mathbf{c}}) = \max\{&\hat{\mu}_\mathcal{X}^T \tilde{\mathbf{x}} : \tilde{\mathbf{x}} \text{ is a BFS of } (\hat{\mathbf{A}}_\mathcal{X}, \mathbf{b}) \\ &\text{corresponding to some basis } \mathcal{J} \text{ containing } a\} \end{aligned} \tag{7}$$

with the convention that the maximum over an empty set is $-\infty$. In Figure 1, we provide a visual presentation of the intersection value scores for the case $L = 1$. The example also explains the name *intersection value*, as it represents the intersections of line segments (for $L > 1$, simplices) with the cost boundary (for $L > 1$, faces of the constraint polytope).

The SFSR algorithm will only pull unknown arms $a \in [K_0]$. The number of times T_k to pull each remaining unknown arm in round k is defined as follows: Set $n_0 := 0$, and for $k \in [K - 1]$,

$$n_k := \left\lceil \frac{1}{\Psi(K_0, L)} \frac{N - K_0}{K + 1 - k} \right\rceil, \quad T_k := n_k - n_{k-1}, \tag{8}$$

where $\Psi(K_0, L) := \sum_{j=1}^{K_0} \frac{1}{\max(2, j-L)}$. The SFSR algorithm is formally presented in Algorithm 1.

Algorithm 1: Score Function-based Successive Reject (SFSR)

Input: Means of reward and costs of known arms $(r_a)_{a\in(K_0,K]}, (c_{l,a})_{a\in(K_0,K],l\in[L]}$; Cost bound $(\bar{c}_l)_{l\in[L]}$; Pulling budget N

Output: Support of the best mixed arm and slack variables (or the symbol $\varnothing$ representing infeasibility)

Compute $(T_k)_{k=1}^{K-1}$ with (8);
Set $\mathcal{X} = [K+L]$;
for $k = 1$ *to* $K-1$ **do**
  Pull each arm $a \in \mathcal{X} \cap [K_0]$ for T_k times;
  Update empirical means of reward $\hat{r}_a$ and costs $(\hat{c}_{l,a})_{l=1}^{L}$ for arms in $\mathcal{X} \cap [K_0]$;
  Compute the score $\hat{\rho}_a^k = f_a^{\mathrm{IV}}(\hat{\mathbf{r}}, \hat{\mathbf{c}})$ for each $a \in \mathcal{X}$ through (5)(6)(7) ;
  if $\max_{a\in\mathcal{X}} \hat{\rho}_a^k = -\infty$ **then**
    return $\varnothing$;
  Eliminate the arm with the lowest score from $\mathcal{X}$ (with arbitrary tie-breaking);
return $\mathcal{X}$;

Remark 2. To use Algorithm 1 to estimate the best mixed arm (i.e., the support *and* the associated probabilities), one can simply construct the empirical constraint sub-matrix $\hat{\mathbf{A}}_{\mathcal{X}}$ according to (6) and compute $\hat{x}^*_{\mathcal{X}} = \hat{\mathbf{A}}_{\mathcal{X}}^{-1}\mathbf{b}$. Note that $\hat{x}^*_{\mathcal{X}}$ may include slack variables as well.

Note that at the beginning of round k, the set $\mathcal{X}$ of remaining arms contains some true arms and may contain some *virtual arms* (corresponding to the slack variables). Whether a true or a virtual arm will be eliminated in round k depends on the random realizations of the rewards and costs. Thus, unlike the classical Successive Reject algorithm Audibert & Bubeck (2010), the number of true arms remaining after each round in our algorithm is a random variable. While the total number of (true) arm pulls by our algorithm is random, we can show that we will always meet the pulling budget N.

Proposition 1. *Under Algorithm 1, the number of total arm pulls never exceeds N.*

The proof can be found in the Supplementary Materials Section D.

Using a different score function. Instead of the intersection value score function f^{IV}, we can also use another function f^{L}, called the Lagrangian function, that comes from linear programming duality theory. The Lagrangian score function is defined as follows: Let $\hat{\mu}_{\mathcal{X}}$ and $\hat{\mathbf{A}}_{\mathcal{X}}$ follow the definitions in (5)(6). Let $\hat{\lambda}^* \in \mathbb{R}^{L+1}$ be an optimal solution to the empirical dual linear program (EDLP) $\min_{\lambda\in\mathbb{R}^{L+1}}\{\mathbf{b}^T\lambda : \hat{\mathbf{A}}_{\mathcal{X}}^T\lambda \geq \hat{\mu}_{\mathcal{X}}\}$. We define

$$f_a^{\mathrm{L}}(\hat{\mathbf{r}}, \hat{\mathbf{c}}) = \begin{cases} \left(\hat{\mu}_{\mathcal{X}} - \hat{\mathbf{A}}_{\mathcal{X}}^T\hat{\lambda}^*\right)_a & \text{EDLP is bounded} \\ -\infty & \text{otherwise} \end{cases} \tag{9}$$

This yields another flavor of the SFSR algorithm that we call SFSR-L. In Appendix H, we will see that the SFSR-L algorithm on some problem instances can perform better than the SFSR algorithm.

5 Analysis

We first introduce the following mild assumption that we will use for our analysis.

Assumption 1. The linear program (SFLP) has a unique optimal solution $\mathbf{x}^*$ with exactly $L+1$ non-zero coordinates.

Remark 3. Assumption 1 does not restrict the best mixed arm to be a strict mix of $L+1$ arms: Note that $\mathbf{x}^*$ contains both the mixing probabilities and the slack variables for the constraints. Assumption 1 requires that if the optimal mixed arm is a mix of m arms, then there need to be exactly $L+1-m$ non-binding constraints under this mixed arm.

The uniqueness of optimal solution is a standard assumption in best arm identification problems (e.g. Audibert & Bubeck (2010); Kaufmann et al. (2016); Faizal & Nair (2022)). The further assumption on the size of the support is necessary for CBMAI problems since it ensures the stability of the optimal solution: Without this assumption, an infinitesimal change of the cost matrix $\mathbf{c}$ could result in a change of the support of the best mixed arm. In this case, identifying the support of best mixed arm beyond a certain probability would be impossible under any budget N, since it requires estimating $\mathbf{c}$ with infinite precision.

Next, given an instance satisfying Assumption 1 we formally define the gaps $\Delta_0, (\Delta_{(i)})_{i\in[K+L]}$ that characterize the hardness of the instance. These gaps appear in both the upper bound (Theorem 1) and lower bound (Theorem 2) on the error probability.

Let $\mathcal{I}^* = \mathrm{supp}(\mathbf{x}^*)$ denote the support of the optimal solution (or the *optimal basis*). For each basis set $\mathcal{J} \subset [K+L], |\mathcal{J}| = L+1$, define the *basis value gap* of $\mathcal{J}$ by

$$\begin{aligned}\Delta^2_{\mathcal{J}} = \inf_{\tilde{\mathbf{r}}\in\mathbb{R}^{K_0}, \tilde{\mathbf{c}}\in\mathbb{R}^{L\times K_0}} \Big\{ & d^2_{\sigma,r}(\mathbf{r},\tilde{\mathbf{r}}) + d^2_{\sigma,c}(\mathbf{c},\tilde{\mathbf{c}}) : \\ & \tilde{\mathbf{A}}^{-1}_{\mathcal{I}^*}\mathbf{b} \geq 0, \tilde{\mathbf{A}}^{-1}_{\mathcal{J}}\mathbf{b} \geq 0,\ \tilde{\mu}^T_{\mathcal{J}}\tilde{\mathbf{A}}^{-1}_{\mathcal{J}}\mathbf{b} \geq \tilde{\mu}^T_{\mathcal{I}^*}\tilde{\mathbf{A}}^{-1}_{\mathcal{I}^*}\mathbf{b}\Big\}\end{aligned} \tag{10}$$

where, $d^2_{\sigma,r}(\mathbf{r},\tilde{\mathbf{r}}) := \sigma_r^{-2}\sum_{a=1}^{K_0}(r_a - \tilde{r}_a)^2$, $d^2_{\sigma,c}(\mathbf{c},\tilde{\mathbf{c}}) := \sigma_c^{-2}\sum_{l=1}^{L}\sum_{a=1}^{K_0}(c_{l,a} - \tilde{c}_{l,a})^2$, and $\tilde{\mathbf{A}}, \tilde{\mu}$ are defined through $(\tilde{r}, \tilde{c})$ in the same way as how $\hat{\mathbf{A}}, \hat{\mu}$ are defined through $(\hat{\mathbf{r}}, \hat{\mathbf{c}})$ in (5)(6). We follow the convention that the infimum of an empty set is $+\infty$. The basis value gap represents the minimum distance one needs to move $(\mathbf{r}, \mathbf{c})$ to an alternative instance $(\tilde{\mathbf{r}}, \tilde{\mathbf{c}})$ where the expected reward under the originally optimal basis $\mathcal{I}^*$ is overtaken by that of another basis $\mathcal{J}$ while preserving the feasibility of $\mathcal{I}^*$.

Note that in (10), the infimum can be attained by moving only the rewards and costs associated with arms in $(\mathcal{J}\cup\mathcal{I}^*)\cap[K_0]$. We write (10) as an infimum over all reward-and-cost vectors for the sake of consistency and ease of notations.

Proposition 2. *Under Assumption 1, $\mathcal{J} \neq \mathcal{I}^*$ if and only if $\Delta^2_{\mathcal{J}} > 0$.*

We relegate the proof to Supplementary Materials Section E.

Furthermore, for each $a \in [K+L]$, we define the *arm value gap* of a as

$$\Delta^2_a = \min\{\Delta^2_{\mathcal{J}} : a \in \mathcal{J} \subset [K+L],\ |\mathcal{J}| = L+1\}. \tag{11}$$

Following Audibert & Bubeck (2010), let (k) denote the arm (including virtual arms) with the k-th smallest Δ_a among all arms $a \in [K+L]$. Under Assumption 1, it follows from Proposition 2 that $0 = \Delta_{(1)} = \cdots = \Delta_{(L+1)} < \Delta_{(L+2)} \leq \cdots \leq \Delta_{(K+L)}$.

In addition to the above, define the *optimal support infeasibility gap* as

$$\Delta^2_0 = \inf_{\tilde{\mathbf{c}}\in\mathbb{R}^{L\times K_0}} \Big\{ d^2_{\sigma,c}(\mathbf{c},\tilde{\mathbf{c}}) : \det(\tilde{\mathbf{A}}_{\mathcal{I}^*}) = 0 \text{ or } \tilde{\mathbf{A}}^{-1}_{\mathcal{I}^*}\mathbf{b} \ngeq 0\Big\} \tag{12}$$

The fact that $\Delta^2_0 > 0$ under Assumption 1 can be established via the continuity and strict positiveness of the mappings $\tilde{c} \mapsto \det(\tilde{\mathbf{A}}_{\mathcal{I}^*})$ and $\tilde{c} \mapsto \tilde{\mathbf{A}}^{-1}_{\mathcal{I}^*}\mathbf{b}$ at $\tilde{c} = \mathbf{c}_{[K_0]}$.

5.1 Upper Bound on the Error Probability of the SFSR Algorithm

We now provide an upper bound on the mis-identification probability of the SFSR algorithm.

Theorem 1. *Let $\mathcal{X}^{SFSR}$ denote the output of Algorithm 1. Then, under Assumption 1, we have*

$$\mathbb{P}(\mathcal{X}^{SFSR} \neq \mathcal{I}^*) \leq \mathcal{O}_L(K)\cdot\exp\left(-\frac{\tilde{N}\Delta^2_0}{K}\right) + \mathcal{O}_L(K^{L+2})\cdot\exp\left(-\min_{2\leq i\leq K}\frac{\tilde{N}\Delta^2_{(i+L)}}{i}\right) \tag{13}$$

where $\tilde{N} := \frac{N-K_0}{3(\frac{L+1}{2}+\log K_0)}$ *and* $\mathcal{O}_L$ *is the standard big-O notation where* L *is treated as a constant.*

Proof. See Appendix A for the proof. □

Remark 4. Theorem 1 shows that the error is dominated by two components: The quantity Δ_0 describes how close the optimal mixed arm is to *infeasibility*, while $\Delta_{(i+L)}$ describes how close the optimal mixed arm is to *other candidates* for the best mixed arm. This echoes the result of Faizal & Nair (2022) for constrained best deterministic arm identification problems.

5.2 Lower Bound

As we stated above, our objective is to design a pure-exploration algorithm ϕ that will minimize the mis-identification probability. But how would we know that our upper bound on it is tight, or not? To characterize that, we introduce a lower bound with which the upper bound can then be compared.

To meaningfully derive a lower bound on the performance of any CBMAI algorithm, it is essential to specify the class of instances to be considered. (It's not so difficult to design algorithms that achieve uniformly good performance on a small class of instances, since the algorithm only needs to distinguish between them.) We consider a class of instances with Gaussian rewards and costs such that the variances σ_c^2, σ_r^2 are fixed and known, so that an instance is parameterized by $\theta = (\theta_a)_{a\in[K]} = (r_a, c_{1,a}, \cdots, c_{L,a})_{a\in[K]}$. For simplicity, we consider $K_0 = K$, i.e., all arms are unknown. We define Θ to be a class of instances θ that either (i) has no feasible solution, or (ii) satisfies Assumption 1. For $\theta \in \Theta$, define $\mathcal{I}^\theta = \varnothing$ if θ is an instance with no feasible solution. Otherwise, define $\mathcal{I}^\theta$ to be the optimal basis for this instance.

We consider a class of algorithms that satisfy the following *consistency* requirement.

Definition 2 (Consistency (Barrier et al., 2023))**.** Let ϕ_N be an algorithm for CBMAI with budget N. A sequence of algorithms $(\phi_N)_{N=N_0}^{\infty}$ is said to be *consistent* if for any instance $\theta \in \Theta$, $\mathbb{P}_\theta(\mathcal{X}^{\phi_N} \neq \mathcal{I}^\theta) \to 0$ as $N \to +\infty$.

The consistency condition means that given sufficient amount of budget, an algorithm can eventually: (i) identify the optimal basis when it is possible to do so, and (ii) output $\varnothing$ whenever the instance has no feasible solution.

It can be shown that the naive Uniform Sampling and Linear Program (`USLP`) algorithm (i.e., pull each arm $\lfloor N/K \rfloor$ times, compute the empirical means of rewards and costs of all arms, and then solve the empirical version of (SFLP)) is a consistent algorithm. We note that the `SFSR` algorithm is a consistent algorithm.

Theorem 2. *For any consistent algorithm* π_N*, under any instance* θ *satisfying Assumption 1, the mis-identification probability satisfies*

$$\limsup_{N\to\infty} -\frac{1}{N}\log \mathbb{P}_\theta(\mathcal{X}^{\pi_N} \neq \mathcal{I}^\theta) \leq \frac{1}{2}\min\{\Delta_0^2, \Delta_{(L+2)}^2\}. \tag{14}$$

Furthermore, for the `USLP` *algorithm, we have*

$$\limsup_{N\to\infty} -\frac{1}{N}\log \mathbb{P}_\theta(\mathcal{X}^{\mathrm{USLP}_N} \neq \mathcal{I}^\theta) \leq \frac{1}{2K}\min\{\Delta_0^2, \Delta_{(L+2)}^2\}. \tag{15}$$

The proof can be found in Appendix B.

For Algorithm 1, Theorem 1 yields that

$$\liminf_{N\to\infty} -\frac{1}{N}\log \mathbb{P}_\theta(\mathcal{X}^{\mathrm{SFSR}_N} \neq \mathcal{I}^\theta) \geq \frac{\min\left\{\frac{\Delta_0^2}{K}, \frac{\Delta_{(L+2)}^2}{2}, \frac{\Delta_{(L+3)}^2}{3}, \cdots, \frac{\Delta_{(L+K)}^2}{K}\right\}}{3\left(\frac{L+1}{2}+\log K\right)}. \tag{16}$$

We can now compare the upper bound for the SFSR algorithm above to the lower bound in (14) and observe the presence of several common terms. Note that *tight* instance-dependent lower bounds for fixed budget identification problems are not known even for unconstrained BAI problems (Degenne, 2023; Qin, 2022). So, while the lower bound we provide in Theorem 2 may not be tight, it does show that the gaps Δ_0 and $(\Delta_{(L+i)})_{i=2}^K$ are appropriate indicators of the hardness of a CBMAI problem: If one of them is very small, then the CBMAI problem is difficult for any consistent algorithm to handle. Instance independent min-max lower bound of the type in Carpentier & Locatelli (2016) or Yang & Tan (2022) can also be derived but are not very meaningful for this problem since one can always construct hard CBMAI instances by translating hard unconstrained BAI problems.

6 Empirical Performance

In this section, we compare the empirical performance of the two flavors of SFSR with the naive USLP algorithm. We only presented instances with $L = 1$ here. Empirical results for instances with more than one constraints are included in Supplementary Materials Section H.

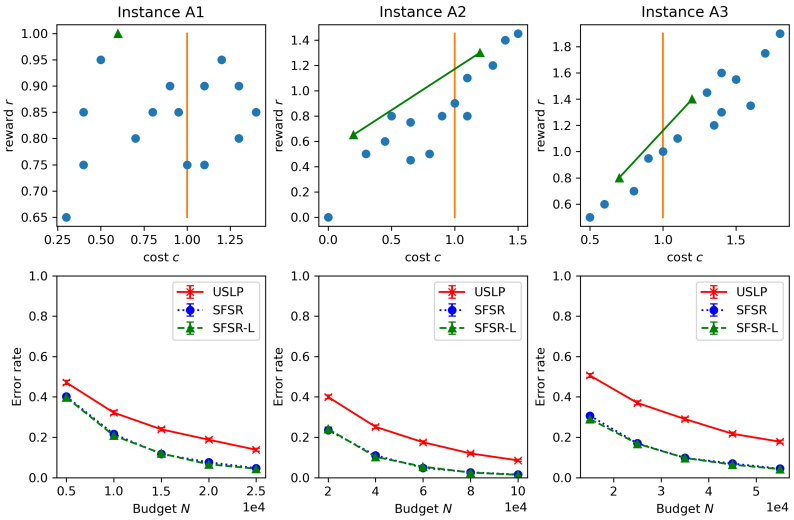

Figure 2: Top: Three 16-arm instances. Arms in the optimal support are labeled with green triangles. Bottom: Empirical results for three algorithms under varying budgets. 95% confidence intervals are indicated and tight.

In all of the experiments, we set $K_0 = K = 16, \sigma_r = 1, \sigma_c = 0.5$, and $\bar{c}_1 = 1$. For each combination of instance-algorithm-budget, we conduct 10,000 independent runs to obtain the error rate as the proportion of times the algorithm produces the wrong support. In every figure, we added (tiny) error bars to represent a 95% confidence interval for the error rate. We implemented the experiments in Python and conducted the experiments on an Apple M1 MacBook Air. Each figure takes about 15 minutes to generate.

We first consider three arbitrary instances $A1, A2$ and $A3$ in Figure 2, where certain arms are clearly sub-optimal while others are not. The instances and their corresponding results are shown in Figure 2. This includes one instance where the optimal arm is deterministic, and two instances where the optimal arm is a strict mix of two arms. The baseline is the Uniform Sampling and Linear Program (USLP) algorithm that pulls each arm $\lfloor N/K \rfloor$ times, computes the empirical means of rewards and costs of all arms, and then solves the empirical version of (SFLP)) (there is no other known algorithm in the literature or otherwise). The results show that both SFSR and SFSR-L clearly outperform USLP, and the two flavors of SFSR have nearly no discernible difference in performance. Furthermore, we also observe that the error rate decreases exponentially in N. Additional comparison of SFSR and SFSR-L is provided in Supplementary Materials Section G.

7 Conclusions

In this paper, we introduced the constrained best mixed arm identification (CBMAI) problem. While prior work has considered such a constrained problem with deterministic arms, it is well known that one can do better allowing for mixed arms. Unfortunately, the mixed arm problem is much more challenging due to existence of uncountably many candidate solutions. We have proposed the first algorithm for the CBMAI problem, which we are able to show theoretically and empirically has very good performance in terms of the error probability decreasing exponentially in N. The problem is of wide interest in many practical settings that often have multiple objectives but with unknown reward and cost models.

Our work provides a basis for further extensions in a number of directions. One could consider contextual bandit models, e.g., linear bandit models that have wide applicability in recommendation systems. A fixed confidence version of CBMAI problem is also interesting, which has only been solved under the known constraints case (Carlsson et al., 2023). This work could also be extended to a constrained MDP setting to find the best policy that also obeys average constraints. There is probably also scope to design an even better algorithm by combining various score functions.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under awards 2025732, 2514683 and 2514748, and ONR MURI N00014-24-1-274.

A Proof of Theorem 1

The proof follows the general strategy first introduced in Audibert & Bubeck (2010) for analyzing fixed-budget BAI algorithms that reject arms successively. As with Audibert & Bubeck (2010), we assume that an infinite reward and cost sequence for each unknown arm $a \in [K_0]$ is drawn before the algorithm started. In this way, the empirical mean reward or cost of arm a after m draws is always well-defined.

Let $\mathcal{X}_k$ denote the set of remaining arms after $k-1$ arms (including virtual arms) are eliminated. Recall that (k) denotes the arm (including virtual arms) with the k-th smallest Δ_a among all arms. At least one of the arms $a \in \{(K+L-k+1), \cdots, (K+L)\}$ is in $\mathcal{X}_k$. If one of the arms in $\mathcal{I}^*$ is eliminated at the end of round k for the first time, it implies that the following event $\mathcal{E}_k$ happened:

$$\mathcal{I}^* \subset \mathcal{X}_k, \quad \exists a \in \{(K+L-k+1), \cdots, (K+L)\} \cap \mathcal{X}_k, \quad \hat{\rho}_a^k \geq \min_{i \in \mathcal{I}^*} \hat{\rho}_i^k \tag{17}$$

where $\hat{\rho}_i^k$ is the intersection value score for arm i at the end of round k.

Next, fix k. Let $\hat{\mathbf{r}} \in \mathbb{R}^{K_0}, \hat{\mathbf{c}} \in \mathbb{R}^{L \times K_0}$ be the empirical means of rewards and costs respectively after each unknown arm has been drawn n_k times. Let $\hat{\mathbf{A}} \in \mathbb{R}^{(L+1)\times(K+L)}$ denote the empirical version of the $\mathbf{A}$ matrix where $\mathbf{A}_{l,i} = c_{l,i}$ is replaced by $\hat{c}_{l,i}$ for $l \in [L], i \in [K_0]$.

If (17) happens, at least one of the following events $\mathcal{E}_{k,0}, (\mathcal{E}_{k,a})_{a \in \{(K+L-k+1), \cdots, (K+L)\}}$ must happen:

$$\mathcal{E}_{k,0} := \{\det(\hat{\mathbf{A}}_{\mathcal{I}^*}) = 0, \text{ or } \hat{\mathbf{A}}_{\mathcal{I}^*}^{-1}\mathbf{b} \not\geq 0\} \tag{18}$$

$$\mathcal{E}_{k,a} := \{\mathcal{I}^* \subset \mathcal{X}_k, a \in \mathcal{X}_k, \hat{\mathbf{A}}_{\mathcal{I}^*}^{-1}\mathbf{b} \geq 0, \ \hat{\rho}_a^k \geq \min_{i \in \mathcal{I}^*} \hat{\rho}_i^k\} \tag{19}$$

On event $\mathcal{E}_{k,0}$, by definition of Δ_0 in (12), we have

$$\xi_0 := \sigma_c^{-2} \sum_{i \in \mathcal{I}^* \cap [K_0]} \sum_{l=1}^{L} (c_{l,i} - \hat{c}_{l,i})^2 \geq \Delta_0^2 \tag{20}$$

For each $(l,i) \in [L] \times [K_0]$, the random variables $\sqrt{n_k}\sigma_c^{-1}(\hat{c}_{l,i} - c_{l,i})$ are i.i.d. standard normal random variables. Therefore, $n_k\xi_0$ is a chi-square random variable with degree $m = L \cdot |\mathcal{I}^* \cap [K_0]|$ (written as χ^2_m). We have

$$\mathbb{P}(\mathcal{E}_{k,0}) \leq \mathbb{P}(\chi^2_m \geq n_k\Delta_0^2) \leq 3^{m/2}\exp\left(-\frac{n_k\Delta_0^2}{3}\right) \leq 3^{L(L+1)/2}\exp\left(-\frac{n_k\Delta_0^2}{3}\right) \tag{21}$$

where in the second inequality we applied Lemma 1 (see Supp. Materials), and in the third inequality we used the fact that $m \leq L|\mathcal{I}^*| = L(L+1)$.

Now consider the event $\mathcal{E}_{k,a}$ for some $a \in \{(K+L-k+1), \cdots, (K+L)\}$. On this event, $\mathcal{I}^*$ corresponds to a BFS of $(\hat{\mathbf{A}}_{\mathcal{X}_k}, \mathbf{b})$. Then, by definition of the IV scoring function (7), for all $i \in \mathcal{I}^*$, we have $\hat{\rho}_i^k \geq \hat{\mu}_{\mathcal{I}^*}^T\hat{\mathbf{A}}_{\mathcal{I}^*}^{-1}\mathbf{b}$. Therefore, on event $\mathcal{E}_{k,a}$, we have $\hat{\rho}_a^k \geq \hat{\mu}_{\mathcal{I}^*}^T\hat{\mathbf{A}}_{\mathcal{I}^*}^{-1}\mathbf{b}$. Again, by definition of the scoring function, this means that there exists a basis $\mathcal{J} \subset [K+L], |\mathcal{J}| = L+1$ such that $a \in \mathcal{J}$ and

$$\hat{\mathbf{A}}_{\mathcal{J}}^{-1}\mathbf{b} \geq 0, \quad \hat{\mu}_{\mathcal{J}}^T\hat{\mathbf{A}}_{\mathcal{J}}^{-1}\mathbf{b} \geq \hat{\mu}_{\mathcal{I}^*}^T\hat{\mathbf{A}}_{\mathcal{I}^*}^{-1}\mathbf{b}. \tag{22}$$

Therefore, by the definition of Δ_a in (11), the above implies that

$$\xi_{\mathcal{J}} := \sum_{i \in (\mathcal{I}^* \cup \mathcal{J}) \cap [K_0]} \left[\sigma_r^{-2}(\hat{r}_i - r_i)^2 + \sigma_c^{-2}\sum_{l=1}^{L}(\hat{c}_{l,i} - c_{l,i})^2\right] \geq \Delta_a^2. \tag{23}$$

The random variables $\sqrt{n_k}\sigma_r^{-1}(\hat{r}_i - r_i), i \in [K_0]$ along with the random variables $\sqrt{n_k}\sigma_c^{-1}(\hat{c}_{l,i} - c_{l,i}), (l,i) \in [L] \times [K_0]$ are i.i.d. standard normal random variables. Therefore, $n_k\xi_{\mathcal{J}}$ is a chi-square random variable with degree $m_{\mathcal{J}} = (L+1) \cdot |(\mathcal{J} \cup \mathcal{I}^*) \cap [K_0]|$. We have $m_{\mathcal{J}} \leq (L+1)(|\mathcal{J}| + |\mathcal{I}^*|) = 2(L+1)^2$. Subsequently,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{k,a}) &\leq \sum_{\mathcal{J}:|\mathcal{J}|=L+1, a\in\mathcal{J}} \mathbb{P}(\xi_{\mathcal{J}} \geq \Delta_a^2) \leq \sum_{\mathcal{J}:|\mathcal{J}|=L+1, a\in\mathcal{J}} \mathbb{P}(\chi^2_{m_{\mathcal{J}}} \geq n_k\Delta_a) && (24)\\ &\leq \binom{K+L-1}{L} 3^{(L+1)^2}\exp\left(-\frac{n_k\Delta_a^2}{3}\right) && \text{(Lemma 1, and } m_{\mathcal{J}} \leq 2(L+1)^2)\end{aligned}$$

Therefore,

$$\mathbb{P}(\mathcal{E}_k) \leq \mathbb{P}(\mathcal{E}_{k,0}) + \sum_{a \in \{(K+L-k+1), \cdots, (K+L)\}} \mathbb{P}(\mathcal{E}_{k,a}) \tag{25}$$

$$\leq 3^{L(L+1)/2}\exp\left(-\frac{1}{3}n_k\Delta_0^2\right) + k\binom{K+L-1}{L}3^{(L+1)^2}\exp\left(-\frac{1}{3}n_k\Delta^2_{(K+L+1-k)}\right) \tag{26}$$

Finally, taking union bound and using $n_k \geq \frac{1}{\Psi(K_0,L)}\frac{N-K_0}{K+1-k}$, we have

$$\mathbb{P}(\mathcal{X}^{\text{SFSR}} \neq \mathcal{I}^*) \leq \sum_{k=1}^{K-1}\mathbb{P}(\mathcal{E}_k) \tag{27}$$

$$\leq \sum_{k=1}^{K-1}\left[3^{L(L+1)/2}\exp\left(-\frac{1}{3}n_k\Delta_0^2\right) + k\binom{K+L-1}{L}3^{(L+1)^2}\exp\left(-\frac{1}{3}n_k\Delta^2_{(K+L+1-k)}\right)\right] \tag{28}$$

$$\leq (K-1)3^{L(L+1)/2}\exp\left(-\frac{(N-K_0)\Delta_0^2}{3(L/2+\log K_0)K}\right) \tag{29}$$

$$+ \frac{K(K+1)}{2}\binom{K+L-1}{L}3^{(L+1)^2}\exp\left(-\frac{N-K_0}{3(L/2+\log K_0)}\min_{2\leq i\leq K}\frac{\Delta^2_{(L+i)}}{i}\right) \tag{30}$$

where in the last inequality we used

$$n_k \geq \frac{N-K_0}{\Psi(K_0,L)} \frac{1}{K+1-k} \qquad \forall k \in [K-1], \tag{31}$$

$$\Psi(K_0,L) \leq \frac{L+1}{2} + \sum_{2 \leq j \leq K_0 - L} \frac{1}{j} \leq \frac{L+1}{2} + \int_1^{K_0} \frac{1}{t} \mathrm{d}t = \frac{L+1}{2} + \log K_0 \tag{32}$$

Remark 5. Notice that we only use the Gaussian reward assumptions in the above proof for obtaining the tail bounds on the sum of squares of independent Gaussian random variables (Lemma 1). To adapt our approach to sub-Gaussian reward settings, one can simply replace such concentration results with similar results on the sum of squares of independent sub-Gaussian random variables by using the fact that squares of sub-Gaussian random variables are sub-exponential variables.

B Proof of Theorem 2

Let θ satisfy Assumption 1. Consider an alternative (not necessarily feasible) instance $\theta' \in \Theta$ with $\mathcal{I}^\theta \neq \mathcal{I}^{\theta'}$. By the consistency of π_N we have

$$q'_N := \mathbb{P}_{\theta'}(\mathcal{X}^{\pi_N} \neq \mathcal{I}^\theta) \xrightarrow{N\to\infty} 1, \text{ and, } q_N := \mathbb{P}_\theta(\mathcal{X}^{\pi_N} \neq \mathcal{I}^\theta) \xrightarrow{N\to\infty} 0. \tag{33}$$

Let $M_{a,N}$ denote the random number of times arm a is pulled under algorithm π_N. Through Lemma 1 of Kaufmann et al. (2016), we have

$$\sum_{a=1}^{K} \mathbb{E}_{\theta'}[M_{a,N}] \mathbf{D}_{\mathrm{KL}}(\theta'_a, \theta_a) \geq d_{\mathrm{KL}}(q'_N, q_N) \tag{34}$$

where $\mathbf{D}_{\mathrm{KL}}(\theta'_a, \theta_a)$ is the KL divergence between the distributions of reward-cost vectors of arm a under instances θ' and θ. The RHS of (34) satisfies

$$d_{\mathrm{KL}}(q'_N, q_N) = q'_N \log\left(\frac{q'_N}{q_N}\right) + (1-q'_N)\log\left(\frac{1-q'_N}{1-q_N}\right) \geq q'_N \log\left(\frac{1}{q_N}\right) - \log 2 \tag{35}$$

where we have used the fact that $-z\log z - (1-z)\log(1-z) \leq \log 2$ for $z \in (0,1)$.

Putting everything together and rearranging the terms, we have

$$\frac{1}{N}\log\left(\frac{1}{q_N}\right) \leq \frac{1}{q'_N}\left(\frac{\log 2}{N} + \sum_{a=1}^{K} \frac{\mathbb{E}_{\theta'}[M_{a,N}]}{N} \mathbf{D}_{\mathrm{KL}}(\theta'_a, \theta_a)\right) \tag{36}$$

Taking the limits on both sides, bounding $\frac{\mathbb{E}_{\theta'}[M_{a,N}]}{N}$ by 1, we have

$$\limsup_{N\to\infty} -\frac{1}{N}\log \mathbb{P}_\theta(\mathcal{X}^{\pi_N} \neq \mathcal{I}^\theta) \leq \sum_{a=1}^{K} \mathbf{D}_{\mathrm{KL}}(\theta'_a, \theta_a) \tag{37}$$

Using the formula for KL divergence between two multivariate Gaussian distributions, we have

$$\mathbf{D}_{\mathrm{KL}}(\theta'_a, \theta_a) = \frac{1}{2}\left[\sigma_r^{-2}(r'_a - r_a)^2 + \sigma_c^{-2}\sum_{l=1}^{L}(c'_{l,a} - c_{l,a})^2\right] \tag{38}$$

Combining the above together and taking infimum over $\theta' \in \Theta$, we have

$$\limsup_{N\to\infty} -\frac{1}{N}\log \mathbb{P}_\theta(\mathcal{X}^{\pi_N} \neq \mathcal{I}^\theta) \leq \frac{1}{2}\inf_{\theta' \in \Theta}\left\{\sigma_r^{-2}\|\mathbf{r}' - \mathbf{r}\|_2^2 + \sigma_c^{-2}\|\mathbf{c}' - \mathbf{c}\|_2^2 : \mathcal{I}^{\theta'} \neq \mathcal{I}^\theta\right\}, \tag{39}$$

Note that $\mathcal{I}^{\theta'} \neq \mathcal{I}^{\theta}$ is true if either (i) $\mathcal{I}^{\theta}$ is an infeasible basis under $(\mathbf{A}', \mathbf{b})$ (in this case, by definition of Δ_0, there exists such θ' whose distance from θ is smaller than $\Delta_0 + \epsilon$ for any small ϵ); or (ii) $\mathcal{I}^{\theta}$ is feasible under $(\mathbf{A}', \mathbf{b})$ and there exists a basis set $\mathcal{J} \neq \mathcal{I}^{\theta}$ such that $(\mu'_{\mathcal{I}^{\theta}})^T(\mathbf{A}'_{\mathcal{I}^{\theta}})^{-1}\mathbf{b} \leq (\mu'_{\mathcal{J}})^T(\mathbf{A}'_{\mathcal{J}})^{-1}\mathbf{b}$ (in this case, by definition of $\Delta_{\mathcal{J}}$, there exists such θ' whose distance from θ is smaller than $\Delta_{\mathcal{J}} + \epsilon$ for any small ϵ). Therefore,

$$\inf_{\theta' \in \Theta} \left\{ \sigma_r^{-2}\|\mathbf{r}' - \mathbf{r}\|_2^2 + \sigma_c^{-2}\|\mathbf{c}' - \mathbf{c}\|_2^2 : \mathcal{I}^{\theta'} \neq \mathcal{I}^{\theta} \right\} \leq \min \left\{ \Delta_0^2, \min_{\mathcal{J} \neq \mathcal{I}^{\theta}} \Delta_{\mathcal{J}}^2 \right\} = \min\{\Delta_0^2, \Delta_{(L+2)}^2\}, \tag{40}$$

which concludes the proof of (14). The proof of (15) can be obtained using the same steps as above along with the fact that for the USLP algorithm $\frac{\mathbb{E}_{\theta'}[M_{a,N}]}{N}$ is bounded by $\frac{1}{K}$.

In (40), there is a small caveat: the gaps $\Delta_0, (\Delta_{\mathcal{J}})_{\mathcal{J} \neq \mathcal{I}^{\theta}}$ were originally defined as infimums over all $\theta' \in \mathbb{R}^{(L+1)\times K}$ while the infimum in the RHS of (39) is taken over Θ, a proper subset of $\mathbb{R}^{(L+1)\times K}$. However, this is not a problem since Θ is dense in $\mathbb{R}^{(L+1)\times K}$ (see Supplementary Materials Section F for a brief explanation of this fact).

References

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Antoine Barrier, Aurélien Garivier, and Gilles Stoltz. On best-arm identification with a fixed budget in non-parametric multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pp. 136–181. PMLR, 2023. Available at https://proceedings.mlr.press/v201/barrier23a/barrier23a.pdf.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in neural information processing systems*, 35:1047–1059, 2022.

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. Available at https://arxiv.org/pdf/2206.11183.pdf.

Emil Carlsson, Debabrota Basu, Fredrik D Johansson, and Devdatt Dubhashi. Pure exploration in bandits with linear constraints. *arXiv preprint arXiv:2306.12774*, 2023. Available at https://arxiv.org/pdf/2306.12774.pdf.

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

Rémy Degenne. On the existence of a complexity in fixed budget bandit identification. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1131–1154. PMLR, 2023.

David Eriksson and Matthias Poloczek. Scalable constrained bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 730–738. PMLR, 2021. Available at https://arxiv.org/pdf/2002.08526.pdf.

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

Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John P Cunningham. Bayesian optimization with inequality constraints. In *ICML*, volume 2014, pp. 937–945, 2014.

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

Michael A Gelbart, Jasper Snoek, and Ryan P Adams. Bayesian optimization with unknown constraints. In *30th Conference on Uncertainty in Artificial Intelligence, UAI 2014*, pp. 250–259. AUAI Press, 2014.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pp. 423–439. PMLR, 2014.

Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. Safe posterior sampling for constrained MDPs with bounded constraint violation. *arXiv preprint arXiv:2301.11547*, 2023.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 1238–1246. PMLR, 2013. Available at http://proceedings.mlr.press/v28/karnin13.pdf.

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

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

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit pareto set identification: the fixed budget setting. *arXiv preprint arXiv:2311.03992*, 2023. Available at https://arxiv.org/pdf/2311.03992.pdf.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

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

Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2), 2019.

Shaoang Li, Lan Zhang, Yingqi Yu, and Xiangyang Li. Optimal arms identification with knapsacks. In *International Conference on Machine Learning*, pp. 20529–20555. PMLR, 2023. Available at https://proceedings.mlr.press/v202/li23aw/li23aw.pdf.

David Lindner, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause. Interactively learning preference constraints in linear bandits. In *International Conference on Machine Learning*, pp. 13505–13527. PMLR, 2022. Available at https://arxiv.org/pdf/2206.05255.pdf.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021a.

Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021b.

David G Luenberger, Yinyu Ye, et al. *Linear and nonlinear programming*, volume 2. Springer, 1984.

Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767, 2021.

Shintaro Nakamura and Masashi Sugiyama. Fixed-budget real-valued combinatorial pure exploration of multi-armed bandit. In *International Conference on Artificial Intelligence and Statistics*, pp. 1225–1233. PMLR, 2024. Available at https://proceedings.mlr.press/v238/nakamura24a/nakamura24a.pdf.

Aldo Pacchiano, Mohammad Ghavamzadeh, and Peter Bartlett. Contextual bandits with stage-wise constraints. *arXiv preprint arXiv:2401.08016*, 2024.

Chao Qin. Open problem: Optimal best arm identification with fixed-budget. In *Conference on Learning Theory*, pp. 5650–5654. PMLR, 2022.

Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems*, 30, 2017.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pp. 1417–1418. PMLR, 2016.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27, 2014.

Xuedong Shang, Igor Colin, Merwan Barlier, and Hamza Cherkaoui. Price of safety in linear best arm identification. *arXiv preprint arXiv:2309.08709*, 2023. Available at https://arxiv.org/pdf/2309.08709.pdf.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Best arm identification with fixed budget: A large deviation perspective. *Advances in Neural Information Processing Systems*, 36, 2024. Available at https://arxiv.org/pdf/2312.12137.pdf.

Zhenlin Wang, Andrew J Wagenmaker, and Kevin Jamieson. Best arm identification with safety constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 9114–9146. PMLR, 2022.

Junwen Yang and Vincent Tan. Minimax optimal fixed-budget best arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 35:12253–12266, 2022. Available at https://arxiv.org/pdf/2105.13017.

Xingyu Zhou and Bo Ji. On kernelized multi-armed bandits with constraints. *Advances in Neural Information Processing Systems*, 35:14–26, 2022.

Supplementary Materials

The following content was not necessarily subject to peer review.

C Auxiliary Results

Lemma 1. *Let χ_m^2 be a chi-squared variable with degree m and $t > 0$, then $\mathbb{P}(\chi_m^2 \geq t) \leq 3^{m/2}\exp(-t/3)$.*

Proof. The moment generating function of χ_m^2 is given by $\mathbb{E}[\exp(\zeta\chi_m^2)] = (1-2\zeta)^{-m/2}$ for $\zeta < 1/2$. Through Markov Inequality we have

$$\mathbb{P}(\chi_m^2 \geq t) \leq \mathbb{E}[\exp(\zeta\chi_m^2)]e^{-\zeta t} = (1-2\zeta)^{-m/2}e^{-\zeta t} \tag{41}$$

The proof is completed by picking $\zeta = 1/3$. □

D Proof of Proposition 1

Define $n_k = n_{K-1}$ for $k > K-1$. Imagine that in each episode, the algorithm pulls all arms still remaining in $\mathcal{X}$ (including virtual arms and known arms). Then the total number of arm pulls is $\sum_{k=1}^{K+L} n_k$.

If an "arm" i is the k-th rejected arm, then it is pulled exactly n_k times. If it is not rejected, then it is pulled n_{K-1} times. Hence to obtain the actual total number of arm pulls, we only need to subtract those n_k's corresponding to virtual arms and known arms from the summation. We conclude that the total number of arm pulls is at most

$$\max_{\substack{\mathcal{J}\subset[K+L]\\ |\mathcal{J}|=K-K_0+L}} \left(\sum_{k=1}^{K+L} n_k - \sum_{k\in\mathcal{J}} n_k\right) = \sum_{k=K-K_0+L+1}^{K+L} n_k \tag{42}$$

$$\leq \sum_{k=K-K_0+L+1}^{K+L} \left(1 + \frac{1}{\Psi(K_0,L)}\frac{N-K_0}{\max(2,K+1-k)}\right) \tag{43}$$

$$= K_0 + (N-K_0)\cdot\frac{1}{\Psi(K_0,L)}\sum_{j=L}^{K_0+L-1}\frac{1}{\max(2,K_0-j)} = N \tag{44}$$

E Proof of Proposition 2

The "if" part is clear by the definition of $\mathcal{I}^*$. We establish the "only if" part as follows.

Consider a basis $\mathcal{J} \neq \mathcal{I}^*$. Suppose that $\Delta_{\mathcal{J}}^2 = 0$. Then, there exists a sequence of reward-cost vectors $(\tilde{r}^{(n)}, \tilde{c}^{(n)})_{n=1}^{\infty}$ such that both of the following hold: (i) $(\tilde{r}^{(n)}, \tilde{c}^{(n)}) \to (r, c)$; (ii) $(\tilde{\mathbf{A}}_{\mathcal{I}^*}^{(n)})^{-1}\mathbf{b} \geq 0, (\tilde{\mathbf{A}}_{\mathcal{J}}^{(n)})^{-1}\mathbf{b} \geq 0,\ (\tilde{\mu}_{\mathcal{J}}^{(n)})^T(\tilde{\mathbf{A}}_{\mathcal{J}}^{(n)})^{-1}\mathbf{b} \geq (\tilde{\mu}_{\mathcal{I}^*}^{(n)})^T(\tilde{\mathbf{A}}_{\mathcal{I}^*}^{(n)})^{-1}\mathbf{b}$.

Set $\mathbf{x}^{(n)} \in \mathbb{R}^{K+L}$ to be the basic feasible solution of $(\tilde{\mathbf{A}}^{(n)}, \mathbf{b})$ corresponding to basis $\mathcal{J}$, i.e., $\mathbf{x}_{\mathcal{J}}^{(n)} = (\tilde{\mathbf{A}}_{\mathcal{J}}^{(n)})^{-1}\mathbf{b}$ and $\mathbf{x}_i^{(n)} = 0$ for $i \notin \mathcal{J}$. Note that $(\mathbf{x}^{(n)})_{n=1}^{\infty}$ is a uniformly bounded sequence of finite dimensional vectors. By taking subsequences, without lost of generality, assume that $\mathbf{x}^{(n)} \to \mathbf{x}^{(\infty)}$. We have

$$\mathbf{x}^{(\infty)} \geq 0, \quad \mathbf{A}\mathbf{x}^{(\infty)} = \lim_{n\to\infty}(\tilde{\mathbf{A}}^{(n)})\mathbf{x}^{(n)} = \lim_{n\to\infty}(\tilde{\mathbf{A}}_{\mathcal{J}}^{(n)})\mathbf{x}_{\mathcal{J}}^{(n)} = \mathbf{b}, \tag{45}$$

meaning that $\mathbf{x}^{(\infty)}$ is a feasible solution of (SFLP). By taking the limit of the last inequality in (ii) we have

$$\mu^T\mathbf{x}^{(\infty)} = \lim_{n\to\infty}(\tilde{\mu}_{\mathcal{J}}^{(n)})^T(\tilde{\mathbf{A}}_{\mathcal{J}}^{(n)})^{-1}\mathbf{b} \geq \limsup_{n\to\infty}(\tilde{\mu}_{\mathcal{I}^*}^{(n)})^T(\tilde{\mathbf{A}}_{\mathcal{I}^*}^{(n)})^{-1}\mathbf{b}. \tag{46}$$

Under Assumption 1, $\mathbf{A}_{\mathcal{I}^*}$ is invertible and hence the mapping $\tilde{c} \to \tilde{\mathbf{A}}_{\mathcal{I}^*}^{-1}$ is continuous at $\tilde{\mathbf{c}} = \mathbf{c}$. Therefore, through (i) we conclude that $\limsup_{n\to\infty}(\tilde{\mu}_{\mathcal{I}^*}^{(n)})^T(\tilde{\mathbf{A}}_{\mathcal{I}^*}^{(n)})^{-1}\mathbf{b} = \mu_{\mathcal{I}^*}^T\mathbf{A}_{\mathcal{I}^*}^{-1}\mathbf{b}$, i.e., the optimal value of (SFLP). Therefore, (46) means that $\mathbf{x}^{(\infty)}$ is also an optimal solution of (SFLP), which contradicts with the uniqueness assumption. (Let $i \in \mathcal{I}^*\backslash\mathcal{J}$, we have $\mathbf{x}_i^{(\infty)} = 0 \neq \mathbf{x}_i^*$ and hence $\mathbf{x}^{(\infty)} \neq \mathbf{x}^*$.)

F Proof of Θ being dense in $\mathbb{R}^{(L+1)\times K}$

If $\theta' \in \mathbb{R}^{(L+1)\times K}\backslash\Theta$, i.e., θ' is a feasible instance that violates Assumption 1, note that, through standard linear programming theory, there exists at least one optimal BFS of (SFLP). Moreover, it is necessary that one of the following statements be true: (i) The optimal solution is non-unique. In this case, there are two BFS achieving the same objective values, i.e., $(\mu'_{\mathcal{I}})^T(\mathbf{A}'_{\mathcal{I}})^{-1}\mathbf{b} - (\mu'_{\mathcal{J}})^T(\mathbf{A}'_{\mathcal{J}})^{-1}\mathbf{b} = 0$ for some bases $\mathcal{I} \neq \mathcal{J}$. (ii) The optional solution is unique but does not have $L+1$ strictly positive coordinates, i.e., certain coordinate of $(\mathbf{A}'_{\mathcal{I}})^{-1}\mathbf{b}$ is zero for some basis $\mathcal{I}$.

In either case, we have $h(\theta') = 0$ for some non-zero polynomial function $h : \mathbb{R}^{(L+1)\times K} \mapsto \mathbb{R}$. Therefore, we conclude that $\mathbb{R}^{(L+1)\times K}\backslash\Theta$ is a subset of the finite union of zero sets of such polynomials. Since the set of zeroes of any non-zero polynomial function cannot contain any open ball, we conclude that $\mathbb{R}^{(L+1)\times K}\backslash\Theta$ does not contain any open set, i.e., Θ is dense in $\mathbb{R}^{(L+1)\times K}$.

G Comparing `SFSR` and `SFSR-L` Algorithms

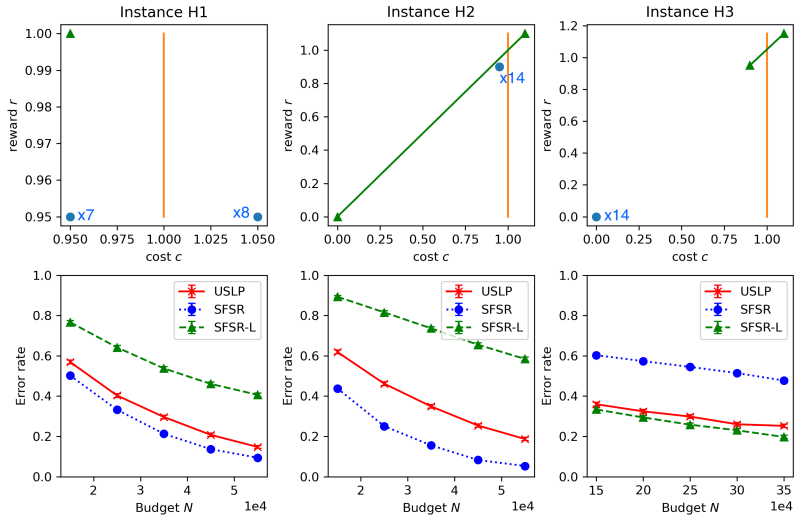

Figure 3: Top: Three hard 16-arm instances. Arms in the optimal support are labeled with green triangles. Bottom: Empirical results for three algorithms under varying budgets. 95% confidence intervals are indicated and tight.

While both flavors of `SFSR` can achieve good performance on an average instance, in Figure 3, we see that in certain carefully constructed hard instances ($H1$-$H3$) while one of the two flavors, `SFSR` or `SFSR-L` performs well, the other does not (the guarantee in any case is probabilistic). In fact, the $H3$ instance in Figure 3 shows that it is hard enough that `SFSR-L` struggles to perform much better than the `USLP` algorithm. Below, we provide a detailed explanation on why `SFSR-L` fails on instance $H2$.

Consider a CBMAI instance with K arms and one type of cost. The mean reward and cost are shown as in the left of Figure 4: Arm 1 has low reward and low cost, arm 2 has high reward and

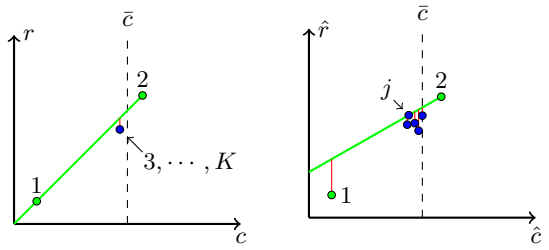

Figure 4: Illustration of SFSR−L in a 1-constraint instance. Left: True mean reward and cost. Right: Empirical means after the first episode. Despite that the empirical means do not deviate from the the true mean by too much, arm 1 (a member of the optimal support) ends up having the lowest empirical Lagrangian reward, and is eliminated as a result.

near feasible cost, and arm 3 to K all have the same mean reward and cost: The cost is feasible but close to cost bound $\bar{c}$, and the reward is chosen such that the best mixed arm is formed by a mixture of arm 1 and 2. The (negative) Lagrangian reward $f_a^{\mathrm{L}}(\hat{\mathbf{r}}, \hat{\mathbf{c}})$ of an arm $a \in [K]$ can be visualized in Figure 4 as the vertical distance between arm a to the "frontier" (i.e., the extended line formed by cost-reward vectors of two arms in the empirical optimal support).

Now, consider the end of episode 1 of SFSR−L, and the empirical means of rewards and costs are shown as on the right of Figure 4. Now, arm 2 and some arm $3 \leq j \leq K$ forms the empirical frontier. The empirical dual optimal solution (symbolized by the slope of the frontier) is very different from the true dual optimizer. More importantly, the empirical Lagrangian reward of arm 1 is now the lowest among all arms, and arm 1 is rejected by the SFSR−L algorithm in round 1 as a result. Note the identification error happens despite the fact that the arm 1 did not underperform (i.e., $\hat{r}_1 < r_1, \hat{c}_1 > c_1$) its mean.

While in elimination style algorithms there's always the possibility of erroneously rejecting optimal arms, we note that the type of event as shown on the right of Figure 4 is not unlikely: We only require *one of* the $K - 2$ arms to slightly outperform its true mean for arm 1 to be eliminated. In comparison, in unconstrained BAI problems, for the optimal arm (arm 1) to be rejected in episode 1 in an elimination-style algorithm (Audibert & Bubeck, 2010; Karnin et al., 2013), it requires *all of* the other arms (including the worst arm) to empirically outperform arm 1.

H Additional Empirical Results

In addition to the experiments in Section 6, we also applied the three algorithms (SFSR, SFSR−L, and USLP) to 6 instances with $L = 2$ constraints. We set $K = K_0 = 24, \sigma_r = 1, \sigma_c = 0.5$ and $\bar{c}_1 = \bar{c}_2 = 1.0$. In all of the three instances, we set the costs of the 24 arms to be the 24 combinations of $c_{1,i} \in \{0.4, 0.6, 0.8, 1.0, 1.2, 1.4\}$ and $c_{2,i} \in \{0.7, 0.9, 1.1, 1.3\}$. Then, to define the rewards for each instance, we first pick a noise vector $(W_i)_{i=1}^{24}$ (which we will describe later) independently for each instance. In instance $D1$, we set $r_i = 1.0 - W_i$. In instance $D2$, we set $r_i = c_{1,i} - W_i$. In instance $D3$, we set $r_i = c_{1,i} + c_{2,i} - W_i$. We run the randomizations for a few times until the optimal support of each instance Dj has exactly j arms. Finally, we increment the reward of each arm in the optimal support by 0.02 to ensure that the optimal support is unique and the instance is not overly difficult for any CBMAI algorithm.

We consider two ways of choosing the random vector $(W_i)_{i=1}^{24}$: (i) a random permutation of $\{0.0, 0.02, \cdots, 0.46\}$. (ii) i.i.d. uniform random choices from $\{0.0, 0.02, \cdots, 0.28\}$. For the former choice, we will refer to the instance as DjP. For the latter, we will use DjI. The specific instances we used are reported in Table 1.

For each combination of instance-algorithm-budget, we run the simulation for 5000 times independently and obtain the error rate as the proportion of times the algorithm output the wrong support. The results are provided in Figure 5. Each figure takes about 2 hours on an Apple M1 MacBook Air.

We can see that the `SFSR-L` algorithm on these two constraints instances either has the same performance as the `SFSR` algorithm or does a bit better in terms of having a lower error rate.

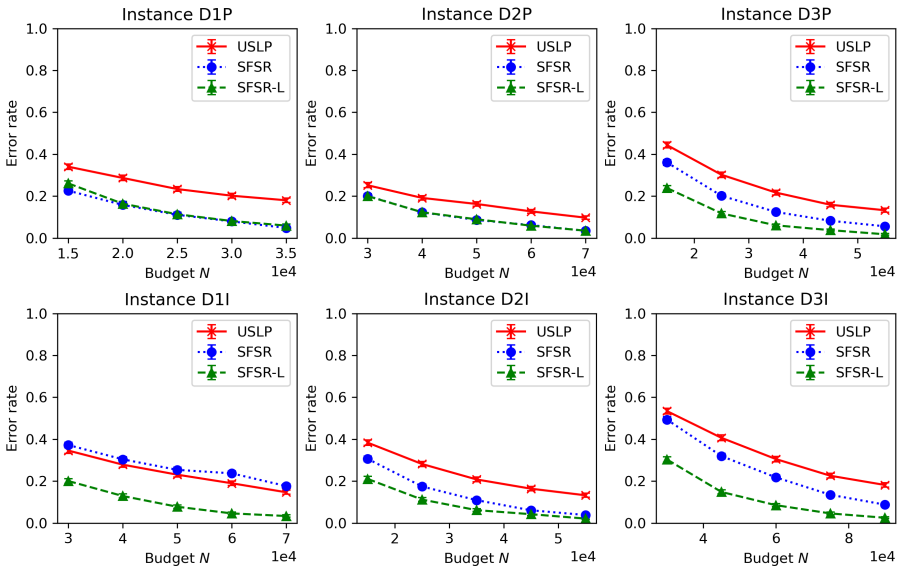

Figure 5: Simulation results for 6 instances with $L = 2$ under varying budget. 95% confidence intervals are indicated and tight.

Table 1: Description of the mean rewards and costs of instances. The rewards of arms in the optimal support is shown in bold. Top Row (from left to right): $D1P, D2P, D3P$. Bottom Row (from left to right): $D1I, D2I, D3I$.

$c_1 \backslash c_2$	0.7	0.9	1.1	1.3
0.4	0.88	0.80	0.82	0.66
0.6	0.72	**1.02**	0.70	0.54
0.8	0.92	0.74	0.94	0.84
1.0	0.76	0.60	0.56	0.86
1.2	0.98	0.64	0.68	0.78
1.4	0.62	0.96	0.90	0.58

$c_1 \backslash c_2$	0.7	0.9	1.1	1.3
0.4	0.08	0.28	−0.02	0.22
0.6	0.20	0.46	0.54	0.40
0.8	0.42	0.52	**0.80**	0.34
1.0	0.92	0.78	0.96	0.70
1.2	0.94	0.76	1.10	0.86
1.4	**1.42**	1.16	1.04	1.24

$c_1 \backslash c_2$	0.7	0.9	1.1	1.3
0.4	1.04	1.22	1.28	1.26
0.6	0.98	1.22	1.60	1.54
0.8	1.26	1.40	**1.88**	1.84
1.0	**1.72**	1.76	1.64	1.92
1.2	1.78	1.70	1.96	2.08
1.4	1.94	**2.30**	2.32	2.50

$c_1 \backslash c_2$	0.7	0.9	1.1	1.3
0.4	0.84	**1.02**	0.74	0.76
0.6	0.84	0.88	0.96	0.90
0.8	0.90	0.98	0.92	0.80
1.0	0.98	0.98	0.90	0.74
1.2	0.94	0.90	0.88	0.72
1.4	0.78	0.82	0.88	0.84

$c_1 \backslash c_2$	0.7	0.9	1.1	1.3
0.4	**0.40**	0.18	0.28	0.14
0.6	0.44	0.54	0.40	0.32
0.8	0.56	0.52	0.68	0.64
1.0	0.84	0.80	0.82	0.74
1.2	0.94	1.18	1.02	1.12
1.4	**1.42**	1.16	1.24	1.24

$c_1 \backslash c_2$	0.7	0.9	1.1	1.3
0.4	0.92	1.12	1.32	1.42
0.6	1.14	1.42	1.62	1.68
0.8	1.30	**1.70**	1.68	**2.06**
1.0	1.46	1.82	2.02	2.04
1.2	1.84	2.02	2.12	2.24
1.4	2.06	**2.28**	2.32	2.58