Optimal Arms Identification with Knapsacks

Shaoang Li [1] **Lan Zhang** [1] **Yingqi Yu** [1] **Xiang-Yang Li** [1]

Abstract

Best Arm Identification (BAI) is a general online pure exploration framework to identify optimal decisions among candidates via sequential interactions. We pioneer the Optimal Arms identification with Knapsacks (OAK) problem, which extends the BAI setting to model the resource consumption. We present a novel OAK algorithm and prove the upper bound of our algorithm by exploring the relationship between selecting optimal actions and the structure of the feasible region. Our analysis introduces a new complexity measure, which builds a bridge between the OAK setting and bandits with knapsacks problem. We establish the instance-dependent lower bound for the OAK problem based on the new complexity measure. Our results show that the proposed algorithm achieves a near-optimal probability bound for the OAK problem. In addition, we demonstrate that our algorithm recovers or improves the state-of-the-art upper bounds for several special cases, including the simple OAK setting and some classical pure exploration problems.

1. Introduction

Multi-armed bandits exemplify the exploration-exploitation trade-off framework in online decision-making problems. The decision-maker selects arms (actions, options, decisions) sequentially and learns from the rewards to maximize the expected cumulative rewards over a number of trials. Many applications need to identify the best action over the candidates, and the rewards or loss during the exploration is *ignored*, which is defined as the best arm identification (BAI) problem (Audibert et al., 2010). For example, in a medical trial problem with m candidate ingredients and T patients that can be admitted to the medical trial (the exploration phase is limited by a fixed budget T), the decision-maker would like to identify the best ingredient to minimize harm to the patients or maximize the medical therapeutic effect. A series of efforts have been made to solve variants of BAI problems (Xu et al., 2020; Katz-Samuels & Jamieson, 2020; Zhang & Ong, 2021; Zhong et al., 2021).

The bandits with knapsacks (BwK) framework was introduced by (Badanidiyuru et al., 2013) to deal with a more general and realistic setting that takes the resource consumption into consideration. In the BwK setting, the optimal fixed distribution over arms may outperform the arm with the highest expected reward. The support of optimal distribution is composed of the arms with optimal 'bang-per-buck,' i.e., reward per unit of resource consumption, thus there are multiple 'optimal arms' (Badanidiyuru et al., 2013). Existing works focus on maximizing the accumulated reward under the resource constraints by finding the optimal fixed distribution (Badanidiyuru et al., 2013; Agrawal & Devanur, 2014).

In this paper, we consider a common situation in which the decision-maker needs to identify all optimal arms with the knapsack constraints, and define it as the optimal arms identification (OAK) problem. In the OAK setting, there exists a fixed set of arms and the hard constrained capacity for each resource; each arm is associated with an unknown reward distribution and an unknown consumption distribution. During the exploration, the decision-maker chooses an arm in each round and only observes a scalar-valued reward independently sampled from the reward distribution and a resource consumption vector independently sampled from the consumption distribution. Once one or more resource budget constraint is violated then the exploration stops. The decision-maker aims to maximize the probability of identifying all optimal arms with the resource knapsacks.

The OAK problem encompasses a wide range of applications due to the presence of resource constraints in pure exploration decision problems. For example, during the medical testing phase, the selection of ingredients may be constrained by the supply and monetary cost of each component, and the decision-maker would like to identify the ingredients that minimize harm to the patients or maximize the medical therapeutic effect. Similarly, the dynamic pricing problem involves sellers who face limited supply and aim to determine the optimal policy for maximizing

[1] University of Science and Technology of China, Hefei, China. Correspondence to: Lan Zhang <zhanglan@ustc.edu.cn>, Xiang-Yang Li <xiangyangli@ustc.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

expected revenue. In the dynamic procurement problem, algorithms are designed to purchase items or services while adhering to budgetary and other constraints.

The knowledge of optimal arms, which is not limited to pure exploration settings, is also valuable for addressing the associated regret minimization BwK problem. Previous work demonstrates that even with sufficient exploration about the optimal solution of LP to converge on an LP-perfect distribution while avoiding obviously suboptimal strategies, an $O(\sqrt{T})$ gap with the optimum remains necessary (Badanidiyuru et al., 2013). In contrast, subsequent studies present tighter regret upper bounds with the known optimal arms set (Flajolet & Jaillet, 2015; Li et al., 2021). For instance, (Li et al., 2021) prove the $O(d^4/b^2)$ regret upper bound, omitting other problem-dependent parameters, which does not depend on T.

The OAK problem raises several challenges for designing and analyzing algorithms. One challenge is the unknown number of optimal arms, which necessitate that the algorithm explores the structure of the feasible region. Another one is to estimate the fixed optimal distribution, as with the BwK problem. The expected per-round reward is no longer a reliable estimate of the arm's value. Each pull of any arm may influence the decision-maker's estimated optimal distribution and lead to a different result. The algorithm needs to search over all possible distributions. And the large search space of possible distribution exacerbates the difficulty of the problem. This differs from the overwhelming majority of previous pure exploration bandits settings. Therefore, the design and analysis of algorithms require new technical tools for characterizing the complexity of the new exploration problem.

1.1. Our contributions

We pioneer the OAK setting under knapsacks constraints in this paper. First, we propose a BASEOAK learning algorithm based on the quarter reject/accept strategy. The analysis of our algorithm depends on our observation of the relationship between selecting optimal arms (point aspect) and the structure of the feasible domain (global aspect). We develop a new complexity measure based on this observation. Our analysis shows the 'successive elimination-style' algorithm places excessive emphasis on point aspect and veers widely from the optimal distribution.

Then, we develop a FULLOAK algorithm based on BASEOAK that strikes a balance between converging to the optimal distribution and exploring the optimality of arms. We upper bound the probability that the algorithm makes mistakes based on the new complexity measure. We establish the instance-dependent lower bound for the OAK problem. The analysis shows that FULLOAK is close to optimum - the lower bound matches the error probability in the exponential term up to a constant factor.

Last, We further investigate some special cases of OAK setting. We study the simple OAK setting for the case that influence from selections of different arms could be avoided. For the simple OAK problem, we present a near-optimal algorithm BASEOAK$^-$ based on BASEOAK. We demonstrate that BASEOAK$^-$ recovers or improves the state-of-the-art upper bounds for many classical pure exploration problems, including the BAI problem, top-K best arms identification problem, and multi-bandits best arms identification problem.

2. Problem Setup and Technical Preliminaries

In this paper, we use bold fonts to represent vectors and matrices. For a matrix $\boldsymbol{C}$, we use $\boldsymbol{C}_{j,\cdot}$ and $\boldsymbol{C}_{\cdot,i}$ to denote the j-th row vector and the i-th column vector, respectively. For a set $\mathcal{X}$, we use $|\mathcal{X}|$ to denote its cardinality.

2.1. Problem setup

We formally define the OAK problem below. Given T rounds, m arms and d types of resources being consumed, they are indexed by $[T] = 1, 2, \ldots, T$, $[m] = 1, 2, \ldots, m$, and $[d] = 1, 2, \ldots, d$, respectively. Each arm is associated with an unknown reward distribution and an unknown consumption distribution. In each round t, the algorithm plays an arm $i(t) \in [m]$, then observes a scalar-valued reward $r(t) \in [0, 1]$ and a resource consumption vector $\boldsymbol{c}(t) \in [0, 1]^d$, which are independently sampled from the reward/consumption distribution. The j-th component of $\boldsymbol{c}(t)$ represents consumption of resource j. There are some fixed unknown reward expected vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top \in [0, 1]^m$ and consumption expected matrix $\boldsymbol{C} = (\boldsymbol{C}_{\cdot,1}, \ldots, \boldsymbol{C}_{\cdot,m}) \in [0, 1]^{d\times m}$ such that $\mathbb{E}[r(t)|i(t)] = \mu_{i(t)}$ and $\mathbb{E}[\boldsymbol{c}(t)|i(t)] = \boldsymbol{C}_{\cdot,i(t)}$.

We use $\boldsymbol{B}$ to denote the hard resource constraint vector. For each resource j, there is a pre-specified knapsack B_j representing the maximum amount constraint of consumption over all time horizons. Once one or more resource budget constraint is violated then the exploration stops. We say the constraints are uniform if $B_j = B$ for all resource $j \in [d]$. And any OAK instance can be reduced to one with uniform constraints $B = \min_{j\in[d]} B_j$. For notation simplicity, we focus on the uniform OAK setting with knapsack B. Let $b = B/T$ denote the expected per-round resource constraint. Besides, we assume that the 1-st resource is the 'time' resource and each arm deterministically consumes 1 unit of it whenever it is picked. We assume the 1-st arm is the 'null' arm that can be played with no reward and only consumes the time resource. These assumptions are standard form in BwK literature (Badanidiyuru et al., 2013; Agrawal & Devanur, 2014; Li et al., 2021).

For a problem instance, we say one arm i is an optimal arm if it would be selected by the optimal dynamic policy in expectation. We give a more precise definition in the linear relaxation part below. The OAK algorithm aims to maximize the probability that correctly identifies all optimal arms with the fixed resource constraint. Formally, let $\mathcal{X}^*$ denote the index set of all optimal arms. The algorithm has to output an arm set $\mathcal{O} \subseteq [m]$ once any constraint is violated (includes the time resource). The algorithm tries to maximize $\mathbb{P}[\mathcal{O} = \mathcal{X}^*]$.

2.2. Linear relaxation

The OAK problem can be relaxed to the following linear program

$$\begin{aligned} \max \quad & \boldsymbol{\mu}^\top \boldsymbol{x}, \\ \text{s.t.} \quad & \boldsymbol{C}\boldsymbol{x} \le \boldsymbol{b}, \qquad (1) \\ & \boldsymbol{x} \ge \boldsymbol{0}. \end{aligned}$$

The $\boldsymbol{x}$ is the decision vector and x_i corresponds to the probability to select arm $i \in [m]$. The LP (1) always has feasible solutions because of the existence of null arms. Let OPT_{LP} and $\boldsymbol{x}^*$ denote the optimal value and optimal solution of (1), respectively. One arm $i \in [m]$ is an optimal arm if the corresponding variable is basic variable in $\boldsymbol{x}^*$, i.e. $x_i^* > 0$.

Formally, let $\mathcal{X}^* := \{i | x_i^* > 0, i \in [m]\}$ and $\mathcal{X}' := \{i | x_i^* = 0, i \in [m]\}$ denote the index set of optimal basic variables and non-basic variables of $\boldsymbol{x}^*$, respectively. Then each arm $i \in \mathcal{X}^*$ is optimal arm and each arm $i \in \mathcal{X}'$ is sub-optimal arm. Similarly, let $\mathcal{Y}^* := \{j | b - (\boldsymbol{x}^*)^\top \boldsymbol{C}_{j,\cdot} = 0, j \in [d]\}$ and $\mathcal{Y}' := \{j | b - (\boldsymbol{x}^*)^\top \boldsymbol{C}_{j,\cdot} > 0, j \in [d]\}$ denote the index set of active constraints and non-active constrains of (1). Then each constraint $j \in \mathcal{Y}^*$ is an active constraint and each constraint $j \in \mathcal{Y}'$ is a non-active constraint. Notice that we always have $|\mathcal{X}^*| = |\mathcal{Y}^*| \le \min\{m, d\}$.

Assumption 2.1. The LP (1) has a unique optimal solution. Moreover, the optimal solution is non-degenerate.

This assumption is a standard one in LP's literature, and any LP can satisfy this assumption with an arbitrarily small perturbation (Megiddo & Chandrasekaran, 1989; Li et al., 2021). The dual problem of (1) is

$$\begin{aligned} \min \quad & \boldsymbol{b}^\top \boldsymbol{w}, \\ \text{s.t.} \quad & \boldsymbol{C}^\top \boldsymbol{w} \ge \boldsymbol{\mu}, \qquad (2) \\ & \boldsymbol{w} \ge \boldsymbol{0}. \end{aligned}$$

Let $\boldsymbol{w}^*$ denote the optimal solution of it. Notice that for each non-active constraint $j \in \mathcal{Y}'$, there is a non-basic variable $w_j^* = 0$.

Then we introduce the sub-optimality measure and optimality measure we use in this paper. To measure the sub-optimality, we use the absolute value of reduced cost/profit $R_i := (\boldsymbol{w}^*)^\top \boldsymbol{C}_{\cdot,i} - \mu_i, i \in [m]$ in LP literature and can be regarded as the cost/profit obtained for increasing a variable by a small amount. Notice that we have $R_{i^*} = 0$ for each optimal arm $i^* \in \mathcal{X}^*$ and $R_{i'} > 0$ for each sub-optimal arm $i' \in \mathcal{X}'$. To measure the optimality, consider the following linear program

$$\begin{aligned} \max \quad & \boldsymbol{\mu}^\top \boldsymbol{x}, \\ \text{s.t.} \quad & \boldsymbol{C}\boldsymbol{x} \le \boldsymbol{b}, \qquad (3) \\ & \boldsymbol{x} \ge \boldsymbol{0}, x_i = 0. \end{aligned}$$

Let $\text{OPT}_{\text{LP}}^{-i}$ denote the optimal value of it, which adds a new constraint $x_i = 0$ for one arm $i \in [m]$ compared with (1). Then define the value $G_i := \text{OPT}_{\text{LP}} - \text{OPT}_{\text{LP}}^{-i}$, which shows the reward gap caused by one arm's deletion. Under Assumption 2.1, we have $G_{i'} = 0$ for each sub-optimal arm $i' \in \mathcal{X}'$ and $G_{i^*} > 0$ for each optimal arm $i^* \in \mathcal{X}^*$.

3. BaseOAK Algorithm and Complexity Measure

This section introduces the intuition and specification of the BASEOAK algorithm (shown in Algorithm 1). We also introduce the new complexity measure. Based on this measure, we upper bound the probability that the algorithm makes mistakes.

3.1. BaseOAK algorithm

The algorithm splits the budget B evenly into $\lceil \log_{4/3} m \rceil - 1$ phases and chooses the worst/best quarter of surviving arms to reject/accept at the end of each phase. An arm will be included in the final output if accepted during the time horizon, and an arm will be excluded if rejected at the end of one phase. The stop condition of BASEOAK is implied in the design of the number of phases and quarter elimination. They guarantee that BASEOAK does not exceed the budget B and each arm is accepted or rejected before BASEOAK ends.

We describe the procedure of the algorithm below. The algorithm maintains three arm sets: the accept arm set $\mathcal{X}_p^*$, the reject arm set $\mathcal{X}_p'$, and the active arm set $\mathcal{X}_p$. The accept/reject arm set includes all accepted/rejected arms before phase p, and the active arm set includes all remaining arms.

During phase p, the algorithm pulls all surviving arms $n(p)$ times, where the definition of $n(p)$ is given in Algorithm 1. Let $s(p)$ denote the times of one surviving arm has been selected until the end of phase p, i.e., $s(p) = \sum_{k=0}^{p} n(k)$. Let $\bar{\boldsymbol{\mu}}(p)$ and $\bar{\boldsymbol{C}}(p)$ denote the empirical mean estimator until the end of phase p for $\boldsymbol{\mu}$ and $\boldsymbol{C}$, respectively. Let $r_i(l)$ and $C_{j,i}(l)$ denote the reward and j-th resource consumption observed of the l-th pull for the arm i. Formally,

$$\bar{\mu}_i(p) = \frac{1}{s(p)} \sum_{l=1}^{s(p)} r_i(l), \quad \bar{C}_{j,i}(p) = \frac{1}{s(p)} \sum_{l=1}^{s(p)} C_{j,i}(l).$$

Algorithm 1 BaseOAK Algorithm (BaseOak)

```
Input: resource constraint $B$, number of arms $m$
$\mathcal{X}_0 \leftarrow [m]$, $\mathcal{X}'_0 \leftarrow \emptyset$, $\mathcal{X}^*_0 \leftarrow \emptyset$
for $p = 0, \dots, \lceil \log_{4/3} m \rceil - 1$ do
    Pull each arm $i \in \mathcal{X}_p \cup \mathcal{X}^*_p$ for
        $n(p) = \left\lfloor \frac{B}{|\mathcal{X}_p \cup \mathcal{X}^*_p| \lceil \log_{4/3} m \rceil} \right\rfloor$
    times
    Compute the empirical estimator of $\bar{R}_i$ and $\bar{G}_i$ for each arm $i \in \mathcal{X}_p$
    if more basic variables in $\bar{\boldsymbol{x}}^*(p)$ then
        $\mathcal{X}^*_{p+1} \leftarrow \mathcal{X}^*_p \cup \{$ the set of $\lceil |\mathcal{X}_p|/4 \rceil$ optimal arms in $\mathcal{X}_p$ with the largest $\bar{G}_i\}$
    else
        $\mathcal{X}'_{p+1} \leftarrow \mathcal{X}'_p \cup \{$ the set of $\lceil |\mathcal{X}_p|/4 \rceil$ sub-optimal arms in $\mathcal{X}_p$ with the largest $\bar{R}_i\}$
    end if
    $\mathcal{X}_{p+1} \leftarrow \mathcal{X}_0 \backslash (\mathcal{X}'_{p+1} \cup \mathcal{X}^*_{p+1})$
end for
Output $\mathcal{X}^*_{\lceil \log_{4/3} m \rceil}$
```

At the end of each phase p, with the empirical estimator $\bar{\boldsymbol{\mu}}(p)$ and $\bar{\boldsymbol{C}}(p)$, compute

$$\begin{aligned} \max \quad & \bar{\boldsymbol{\mu}}^\top \boldsymbol{x}, \\ \text{s.t.} \quad & \bar{\boldsymbol{C}} \boldsymbol{x} \le \boldsymbol{b}, \\ & \boldsymbol{x} \ge \boldsymbol{0}. \end{aligned} \tag{4}$$

Let $\bar{\boldsymbol{x}}^*(p)$ and $\bar{\boldsymbol{w}}^*$ denote the optimal solution of it and the dual problem, respectively. In the meantime, we compute the the empirical estimator of $\bar{R}_i$ and $\bar{G}_i$ for each surviving $i \in \mathcal{X}_p \cup \mathcal{X}^*_p$ with $\bar{\boldsymbol{\mu}}(p)$ and $\bar{\boldsymbol{C}}(p)$. We have $\bar{R}_i = (\bar{\boldsymbol{w}}^*)^\top \bar{\boldsymbol{C}}_{\cdot,i} - \bar{\mu}_i, i \in [m]$. Similarly, consider the following LP

$$\begin{aligned} \max \quad & \bar{\boldsymbol{\mu}}^\top \boldsymbol{x}, \\ \text{s.t.} \quad & \bar{\boldsymbol{C}} \boldsymbol{x} \le \boldsymbol{b}, \\ & \boldsymbol{x} \ge \boldsymbol{0}, x_i = 0. \end{aligned}$$

Let $\overline{\mathrm{OPT}}_{\mathrm{LP}}^{-i}$ denote the optimal value of it, we have $\bar{G}_i = \bar{\boldsymbol{\mu}}^\top \bar{\boldsymbol{x}}^*(p) - \overline{\mathrm{OPT}}_{\mathrm{LP}}^{-i}$.

If there are more basic variables (corresponding to the optimal arms) in $\bar{\boldsymbol{x}}^*(p)$, the algorithm chooses a quarter of arms with the largest $\bar{G}_i$ from the active arm set $\mathcal{X}_p$ to accept and adds them into the accept arm set $\mathcal{X}^*_{p+1}$. Conversely, the algorithm chooses a quarter of arms with largest $\bar{R}_i$ from the active arm set $\mathcal{X}_p$ to reject and adds them into the reject arm set $\mathcal{X}'_{p+1}$. Maybe there are some arms with the same $\bar{R}_i$ /$\bar{G}_i$ such that it is difficult to decide which arm to reject/accept. We use a random strategy in this case, i.e., select a random arm to reject/accept until a quarter of the arms are eliminated. At the end of the last phase, the algorithm outputs all arms accepted during the whole time horizons.

The algorithm cannot be simplified to just eliminate and return the active set. It is important to maintain the active arm set and reject arm set simultaneously to make sure that each arm will be rejected/accepted only once during the game.

3.2. Complexity measure

We introduce the complexity measure used in our work. Let $\mathcal{D} = \{\boldsymbol{x} \in \mathbb{R}^m | \boldsymbol{C}\boldsymbol{x} \le \boldsymbol{b}, \boldsymbol{x} \ge 0\}$ denote the feasible domain of (1). Notice that $\mathcal{D}$ is a convex polyhedron. Let $\mathcal{B}$ denote the set of all vertexes (are also extreme points) of the convex polyhedron. We say $\boldsymbol{x} \in \mathcal{D}$ is a vertex if $(\forall \lambda \in (0,1), \boldsymbol{u}, \boldsymbol{v} \in \mathcal{D})[\boldsymbol{x} = \lambda \boldsymbol{u} + (1-\lambda)\boldsymbol{v} \Rightarrow \boldsymbol{u} = \boldsymbol{v}]$. We use $\boldsymbol{x}_{(k)}$ to denote the k-th optimal vertex, i.e.,

$$\begin{aligned} \boldsymbol{\mu}^\top \boldsymbol{x}^* = \boldsymbol{\mu}^\top \boldsymbol{x}_{(1)} \ge \boldsymbol{\mu}^\top \boldsymbol{x}_{(2)} \dots \\ \ge \boldsymbol{\mu}^\top \boldsymbol{x}_{(k)} \ge \dots \ge \boldsymbol{\mu}^\top \boldsymbol{x}_{(|\mathcal{B}|)}. \end{aligned}$$

Under Assumption 2.1, the number of extreme points is no less than m. We define the *vertex gap* Δ_i as

$$\Delta_i = \boldsymbol{\mu}^\top \boldsymbol{x}^* - \boldsymbol{\mu}^\top \boldsymbol{x}_{(i)}, i \in [m].$$

Our analysis relies on the following complexity measure:

$$H := \max_{i \neq 1} \frac{i}{\Delta_i^2}, i \in [m],$$

which is a generalization of the complexity measure for BAI. Note that the complexity measure captures the reduced cost of sub-optimal arm and the influence caused by that one of the optimal arm is not allowed to use simultaneously, and builds a bridge between point aspect (sub-optimality of arms) and global aspect (the feasible domain of latent structures of a problem instance, which could be induced from the BwK domain). We provide a formal description below.

Theorem 3.1. *For any optimal arms identification or bandits with knapsack problem instance, we have $R_{(i)} \ge \frac{\Delta_{i+1}}{\sqrt{2}}$ and $G_{(i)} \ge \Delta_{i+1}$.*

Proof Sketch. Let $\boldsymbol{d}_q$ denote the edge direction vector from $\boldsymbol{x}^*$ leading to the adjacent extreme points $\boldsymbol{x}^{(q)}$ corresponding to the increase of the sub-optimal variable $q \in \mathcal{X}'$. We use $\boldsymbol{x}^{(q)}$ to denote the q-th optimal adjacent vertex, i.e. $\boldsymbol{\mu}^\top \boldsymbol{x}^* > \boldsymbol{\mu}^\top \boldsymbol{x}^{(1)} \ge \dots \ge \boldsymbol{\mu}^\top \boldsymbol{x}^{(q)} \ge \dots, q \in \mathcal{X}'$. We define the *adjacent gap* $\Delta^{(q)} := \boldsymbol{\mu}^\top \boldsymbol{x}^* - \boldsymbol{\mu}^\top \boldsymbol{x}^{(q)}, q \in \mathcal{X}'$. Then we consider the relationship between $\Delta^{(q)}$ and $R_{(q)}$. We rearrange $\boldsymbol{w}^* = (w^*_B | w^*_N)^\top$, where the basis vector

$\boldsymbol{w}_B^* \in \mathbb{R}^{|\mathcal{Y}^*|}$ includes all basic variables and the non-basis vector $\boldsymbol{w}_N^* = (0, \dots, 0)^\top$. Let d_{qi} denote the i-th element of $\boldsymbol{d}_q$. Define $\alpha_q := \min_{i \in \mathcal{X}^*} \left\{ \frac{x_i^*}{-d_{qi}} \right\}$. We have

$$\begin{aligned}\Delta^{(q)} &= -\alpha_q \boldsymbol{\mu}^\top \boldsymbol{d}_q = -\alpha_q \left(\mu_q - \left(\boldsymbol{w}_B^*\right)^\top \boldsymbol{Q}_{\cdot,q}\right) \\ &= -\alpha_q \left(\mu_q - \left(\boldsymbol{w}^*\right)^\top \boldsymbol{C}_{\cdot,q}\right) = \alpha_q R_{(q)}\end{aligned}$$

From the geometry of linear programming, α_q is the distance between the vertex $\boldsymbol{x}^*$ and $\boldsymbol{x}^{(q)}$. And the distance of the polyhedron (i.e. the feasible domain of (1)) is not more than $\sqrt{2}$. From the definition of vertex gap and adjacent gap, we have $\Delta_{q+1} \le \Delta^{(q)}$. Combine the two facts together, we obtain $R_{(i)} \ge \frac{\Delta_{i+1}}{\sqrt{2}}$.

Then consider LP (3), notice that the feasible domain of (3) is a subset of $\mathcal{D}$ and the optimal extreme point of (3) is also a vertex of $\mathcal{D}$. The feasible domain of (3) is nonempty because of the existence of the null arm. Under assumption 2.1, different LP (3) with different absent optimal arms have different vertex. Based on the definition of vertex gap, we could conclude that $G_{(i)} \ge \Delta_{i+1}$.

The details of the proof are provided in Appendix A. □

3.3. Theoretical result

Theorem 3.2. *For the OAK problem, Algorithm 1 makes errors with probability at most*

$$O\left(md \log m \cdot \exp\left(-\frac{b^4 B}{90 H \max(|\mathcal{X}^*|, \log m)}\right)\right).$$

Proof Sketch. Notice that the total pulls of BASEOAK is at most B and $c_{i,j}(t) \le 1$ for all arm i and resource j during one round t, the algorithm will never exceed the consumption knapsack.

First, we argue that at the end of each phase, the optimal value of (4) is always close to $\mathrm{OPT}_{\mathrm{LP}}$. At the end of phase p, with probability at least $1 - 2md \cdot \exp\left(-2\delta_p^2 s(p)\right)$, we have

$$\begin{aligned}\bar{\boldsymbol{\mu}}^\top \bar{\boldsymbol{x}}^*(p) &\ge \left(1 - \frac{\delta_p}{b}\right) \mathrm{OPT}_{\mathrm{LP}} - \delta_p, \\ \bar{\boldsymbol{\mu}}^\top \bar{\boldsymbol{x}}^*(p) &\le \left(1 + \frac{\delta_p}{b}\right) \left(\mathrm{OPT}_{\mathrm{LP}} - \Delta_2\right) + \delta_p.\end{aligned}$$

Then based on the analysis of the gap between the optimal solution of the dual form of (4) and $\boldsymbol{w}^*$, we could bound the probability that $\bar{R}_{i^*}(p) > \bar{R}_{i'}(p)$ is at most

$$2md \cdot \exp\left(-2\left(\frac{b\Delta_2}{8} + \frac{b^2 R_{i'}}{8}\right)^2 s(p)\right).$$

And the probability that $\bar{G}_{i^*}(p) < \bar{G}_{i'}(p)$ is at most

$$2md \cdot \exp\left(-\frac{2b^2 G_{i^*}^2}{9} s(p)\right).$$

For all arms in $\mathcal{X}_p$, let $\mathcal{P}_p$ and $\bar{\mathcal{P}}_p$ denote the set of optimal arms for (1) and (4), $\mathcal{Q}_p$ and $\bar{\mathcal{Q}}_p$ denote the set of sub-optimal arms for (1) and (4). Let S_p^* denote the $\frac{1}{16}|\mathcal{X}_p|$ arms with smallest G_i and S_p' denote the $\frac{1}{16}|\mathcal{X}_p|$ arms with largest R_i, respectively. Define

$$\begin{aligned}\Phi_p^* &:= \max_{i \in \mathcal{P}_p \backslash S_p^*} \exp\left(-\frac{2b^2 G_i^2}{9} s(p)\right), \\ \Phi_p' &:= \max_{i \in \mathcal{Q}_p \backslash S_p'} \exp\left(-2\left(\frac{b\Delta_2}{8} + \frac{b^2 R_i}{8}\right)^2 s(p)\right).\end{aligned}$$

Let us start with the case that $\bar{\mathcal{Q}}_p > \bar{\mathcal{P}}_p$. Consider the number of arms in $\bar{\mathcal{Q}}_p \cap (\mathcal{P}_p \backslash S_p^*)$, then

$$\begin{aligned}\mathbb{E}[|\bar{\mathcal{Q}}_p \cap (\mathcal{P}_p \backslash S_p^*)|] &= \sum_{i \in \mathcal{P}_p \backslash S_p^*} \mathbb{P}[\bar{G}_i(p) < \bar{G}_{i'}(p)] \\ &\le \sum_{i \in \mathcal{P}_p \backslash S_p^*} 2md \cdot \exp\left(-\frac{2b^2 G_{i^*}^2}{9} s(p)\right) \\ &\le 2md \cdot |\mathcal{P}_p \backslash S_p^*| \cdot \Phi_p^*.\end{aligned}$$

Then we apply Markov's inequality and obtain

$$\mathbb{P}[|\bar{\mathcal{Q}}_p \cap (\mathcal{P}_p \backslash S_p^*)| > \frac{1}{8}|\bar{\mathcal{Q}}_p|] \le 16md \cdot \frac{|\mathcal{P}_p \backslash S_p^*|}{|\bar{\mathcal{Q}}_p|} \Phi_p^*.$$

We could bound the cardinality of the set $\bar{\mathcal{Q}}_p \cap \mathcal{Q}_p$ with high probability

$$\mathbb{P}[|\bar{\mathcal{Q}}_p \cap \mathcal{Q}_p| > \frac{3}{4}|\bar{\mathcal{Q}}_p|] \ge 1 - 16md \cdot \frac{|\mathcal{P}_p \backslash S_p^*|}{|\bar{\mathcal{Q}}_p|} \Phi_p^*.$$

Based on this event, let i_p^* denote the eliminated optimal arm in phase p. Consider the the number of arms in $(\bar{\mathcal{Q}}_p \cap \mathcal{Q}_p) \backslash S_p'$ with larger $\bar{R}_x$ than that of the eliminated optimal arm and let N_p' denote it, then

$$\begin{aligned}\mathbb{E}[N_p'] &= \sum_{i \in (\bar{\mathcal{Q}}_p \cap \mathcal{Q}_p) \backslash S_p'} \mathbb{P}[\bar{R}_{i_p^*}(p) < \bar{R}_i(p)] \\ &\le \sum_{i \in (\bar{\mathcal{Q}}_p \cap \mathcal{Q}_p) \backslash S_p'} 2md \cdot \exp\left(-\left(\frac{b\Delta_2}{4\sqrt{2}} + \frac{b^2 R_i}{4\sqrt{2}}\right)^2 s(p)\right)\end{aligned}$$

By applying Markov's inequality, we obtain

$$\mathbb{P}[N_p' > \frac{1}{6}|\bar{\mathcal{Q}}_p \cap \mathcal{Q}_p|] \le \frac{6\mathbb{E}[N_p']}{|\bar{\mathcal{Q}}_p \cap \mathcal{Q}_p|}.$$

We obtain that the probability that there is at least one eliminated optimal arms is at most $32md \cdot \Phi_p^* + 12md \cdot \Phi_p'$. Similarly, the probability that at least one sub-optimal arm is added to $\mathcal{X}_p^*$ is at most $32md \cdot \Phi_p' + 12md \cdot \Phi_p^*$. The pulls for each arm in $\mathcal{X}_p$ before phase $t+1$ satisfy

$$\begin{aligned}s(p) &\ge \frac{B}{\log_{4/3} m} \sum_{k=0}^{p} \frac{1}{|\mathcal{X}_p + \mathcal{X}_p^*|} \\ &\ge \frac{B}{\log_{4/3} m} \sum_{k=0}^{p} \min\left(\frac{2}{|\mathcal{X}_p|}, \frac{2}{|\mathcal{X}_p^*|}\right).\end{aligned}$$

Algorithm 2 FullOAK Algorithm (FULLOAK)

```
Input: resource constraint $B$, number of arms $m$
Pull each arm once
while time horizon is less than $T/3$ and the consumption of any resource is less than $B/3$ do
    for $t = 1, 2, \dots, m$ do
        Solve the linear program (5) and let $\boldsymbol{x}_t$ denote the solution for it
        Choose an arm to 'pull' as an independent sample from the distribution $\boldsymbol{x}_t$
    end for
end while
Obtain the estimate of the optimal distribution $\boldsymbol{x}_{T/3}$.
$\mathcal{X}_0 \leftarrow [m]$, $\mathcal{X}'_0 \leftarrow \emptyset$, $\mathcal{X}^*_0 \leftarrow \emptyset$
for $p = 0, \dots, \lceil \log_{4/3} m \rceil - 1$ do
    Pull each arm $i \in \mathcal{X}_p$ for
        $$n(p) = \left\lfloor \frac{B}{2|\mathcal{X}_p \cup \mathcal{X}^*_p| \lceil \log_{4/3} m \rceil} \right\rfloor$$
    times
    Compute the empirical estimator of $\bar{R}_i$ and $\bar{G}_i$ for each arm $i \in \mathcal{X}_p$
    if more basic variables in $\bar{\boldsymbol{x}}^*(p)$ then
        $\mathcal{X}^*_{p+1} \leftarrow \mathcal{X}^*_p \cup \{$ the set of $\lceil |\mathcal{X}_p|/4 \rceil$ optimal arms in $\mathcal{X}_p$ with the largest $\bar{G}_i\}$
    else
        $\mathcal{X}'_{p+1} \leftarrow \mathcal{X}'_p \cup \{$ the set of $\lceil |\mathcal{X}_p|/4 \rceil$ sub-optimal arms in $\mathcal{X}_p$ with the largest $\bar{R}_i\}$
    end if
    $\mathcal{X}_{p+1} \leftarrow \mathcal{X}_0 \backslash (\mathcal{X}'_{p+1} \cup \mathcal{X}^*_{p+1})$
end for
Output $\mathcal{X}^*_{\lceil \log_{4/3} m \rceil}$
```

Combine them together, then we complete the proof. The details of the proof are provided in Appendix B. □

4. FullOAK Algorithm and Lower Bound

This section develops an algorithm, called FULLOAK (shown in Algorithm 2), that solves the OAK problem based on some intuitions of BASEOAK. We provide the introduction, the main idea, and theoretical analysis of the algorithm. Moreover, we provide an instance-dependent lower bound for the OAK problem.

4.1. FullOAK algorithm

Notice that the theoretical analysis of BASEOAK show the dependence on $|\mathcal{X}^*|$ that the learning strategy makes mistakes for the OAK problem. The main reason is that an accurate estimator of the fixed optimal distribution suffices to guarantee algorithms with low error probability. However, BASEOAK does uniform exploration between all surviving arm during one phase, which veer widely from the optimal distribution. This also makes BASEOAK cannot delete any optimal arm during the game, so the exploration ability is limited. The dependence could be avoided if the algorithm obtains an accurate estimator before the reject/accept phase. Based on these analyses, we present the FULLOAK algorithm. There are two steps: the first step derived from the UCB family of algorithms and aims to converge the optimal solution of (1); the second step based on BASEOAK, the difference is that FULLOAK will delete all accept arms from the surviving arms set at the end of each phase. We provide the specification of the first step below.

Let $n_i(t)$ denote the number of pulls of arm i before round $t+1$. Let $\boldsymbol{\mu}^U(t)$ and $\boldsymbol{C}^L(t)$ denote the upper confidence bound reward vector and lower confidence bound consumption matrix until round t, respectively. Formally,

$$\mu_i^U(t) := proj_{[0,1]}\left(\bar{\mu}_i(t) + 2f_{rad}(\bar{\mu}_i(t), n_i(t)+1)\right),$$
$$C_{j,i}^L(t) := proj_{[0,1]}\left(\bar{C}_{i,j}(t) - 2f_{rad}(\bar{C}_{i,j}(t), n_i(t)+1)\right),$$

where $proj_{[0,1]}$ is a project function from real number to interval $[0,1]$ and $f_{rad}(v,n) = \sqrt{\frac{\gamma v}{n}} + \frac{\gamma}{n}, \gamma > 0$ is a confidence radius function. Then after the selection of round t, consider the following linear program

$$\begin{aligned} \max \quad & (\boldsymbol{\mu}^U(t))^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{C}^L(t)\boldsymbol{x} \leq (1-\epsilon)\boldsymbol{b}. \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{aligned} \tag{5}$$

The algorithm solves this linear program and selects arm according to the optimal solution of it for each round during the first step. Let $n_i(T/3)$ denote the pulls for arm i during the first phase. Then we obtain $\boldsymbol{x}_{T/3}$ with $(x_{T/3})_i = \frac{n_i(T/3)}{\sum_i n_i(T/3)}$.

4.2. Theoretical result

The following theorem expresses the error bound for FULLOAK.

Theorem 4.1. *For the OAK problem, with* $\epsilon = \sqrt{\frac{3\log(mdT)m}{B}} + \frac{3\log(mdT)m\log T}{B}$, *Algorithm 2 makes errors with probability at most*

$$O\left(mdT \cdot \exp\left(-\frac{\alpha b^2 B}{H \log m}\right)\right),$$

where α is a constant.

Proof sketch. The upper confidence bound of the expected reward $\boldsymbol{\mu}^U(T/3)$ and lower confidence bound of the expected consumption $\boldsymbol{C}^L(T/3)$ satisfy the following properties:

(1) with probability at least $1 - 2mdT \cdot \exp(-\Omega(\gamma))$,

$$|(\boldsymbol{\mu}^U(T/3))^\top \boldsymbol{x}_{T/3} - \mathrm{OPT}_{\mathrm{LP}}| \\ \leq O\left(\sqrt{\frac{\gamma m \cdot \mathrm{OPT}_{\mathrm{LP}}}{T}} + \frac{\gamma m d}{T} + \frac{\mathrm{OPT}_{\mathrm{LP}}}{B}\sqrt{\frac{\gamma m d \cdot b}{T}}\right).$$

(2) with probability at least $1 - 2mdT \cdot \exp(-\Omega(\gamma))$,

$$\sum_{t=1}^{T/3} \left|(\boldsymbol{C}^L(t))^\top \boldsymbol{x}_t - \boldsymbol{c}_t\right| \\ \leq \left(1 - O\left(\sqrt{\frac{\gamma m}{B}} + \frac{\gamma m \log T}{B}\right)\right)\frac{B\mathbf{1}}{3}.$$

Notice that the consumption during the second step is at most $\frac{B}{2}$, so the consumption of FULLOAK will less than $\frac{5B}{6}$ with high probability.

For the second step, at the end of phase p, for any optimal arm $i^* \in \mathcal{X}^*$ and any sub-optimal arm $i' \in \mathcal{X}'$, the probability that $\bar{R}_{i^*}(p) > \bar{R}_{i'}(p)$ is at most

$$2md \cdot \exp\left(-\alpha_1 \cdot b^2 R_{i'}^2 s(p)\right)$$

for some constant α_1. And the probability that $\bar{G}_{i^*}(p) < \bar{G}_{i'}(p)$ is at most

$$2md \cdot \exp\left(-\frac{2b^2 G_{i^*}^2}{9} s(p)\right).$$

Similar to the proof of Theorem 3.2, we bound the probability that the algorithm makes mistakes by ignoring the $\frac{1}{16}|\mathcal{X}_p|$ arms with the smallest G_i and the $\frac{1}{16}|\mathcal{X}_p|$ arms with largest R_i of the active arms set, then we complete the proof. The details of the proof are provided in Appendix C. □

4.3. Lower bound

We provide an instance-dependent lower bound. Our analysis ensures that any bandit strategy nevertheless makes a mistake for some OAK problem instances.

Theorem 4.2. *For some OAK problem instances, consider any bandits algorithm that output an arm set $\mathcal{O} \subseteq [m]$ at the end of the T-th round, it holds that*

$$\mathbb{P}(\mathcal{O} \neq \mathcal{X}^*) \geq \Omega\left(\exp\left(-\frac{\beta b^2 B}{H \log m}\right)\right),$$

where β is a constant.

Proof. We provide the core constructions below and give the detailed proof in Appendix D.

Let $(p_w)_{2 \leq w \leq W} \in [1/4, 1/2)$ be $(W-1)$ real numbers and let $p_1 = 1/2$. And we define the quantities $l_w := 1/2 - p_w$. Assume m is an exact multiple of W. Then we define

$$\mu_i := \frac{1}{2} - \frac{l_w}{2^{\lfloor (m-i)/W \rfloor}}, w = (i \bmod W), i \in [m].$$

Let π_i denote the Bernoulli distribution of mean μ_i and π_i' denote the Bernoulli distribution of mean $1 - \mu_i$.

Consider W problem instances with time horizon T, m arms, d types of resources being consumed, and knapsack $b = W/m$ for each type of resource. To ease the reading, assume T is a power of 2, $W \geq \Omega(\sqrt{m})$, and $d > m/W$. Let $w = (i \bmod W)$, for the u-th problem instance, the i-th arm x_i^u is associated with the reward distribution π_i^u,

$$\pi_i^u := \pi_i \mathbf{1}\{w \neq u\} + \pi_i' \mathbf{1}\{w = u\}, u \in [W], i \in [m].$$

The consumption vector $\boldsymbol{c}_i^u$ satisfies $(\boldsymbol{c}_i^u)_1 = (\boldsymbol{c}_i^u)_d = (\boldsymbol{c}_i^u)_w = 1$, and $(\boldsymbol{c}_i^u)_j = 0$ for all $j \neq 1, j \neq w, j \neq d$.

□

5. Special Cases

In this section, we investigate some special cases of the OAK problem, including simple OAK problem and some classical pure exploration problems.

5.1. Simple OAK problem

The upper and lower bounds show the dependence on $|\mathcal{X}^*|$ that BASEOAK makes mistakes for the general OAK problem. The dependence could be avoided for some simple OAK problems. We say if the deletion of any optimal arm does not change R_i and G_i of any other arm, then the OAK problem is a simple OAK problem. We provide some examples of simple OAK problem in Sec. 5.2. For the simple OAK problem, we present BASEOAK$^-$ (Algorithm 3) based on BASEOAK: the algorithm eliminates the accepted arms from the active arm set at the end of each accept phase.

Theorem 5.1. *For the simple OAK problem, Algorithm 3 makes errors with probability at most*

$$O\left(m^2 \cdot \exp\left(-\frac{\kappa T}{H \log m}\right)\right),$$

where κ is a constant.

We provide the specification of Algorithm 3 and the proof details of Theorem 5.1 in Appendix E.1.

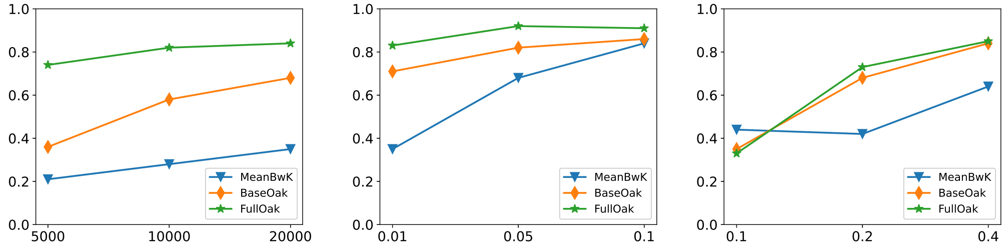
5.2. Pure exploration problems

Example 5.2 (Best arm identification)**.** *The best arm identification problem can be modeled by the OAK problem with one resource (time resource) and one optimal arm. For the BAI problem, Algorithm 3 makes errors with probability at most*

$$O\left(\log m \cdot \exp\left(-\frac{\kappa T}{H \log m}\right)\right),$$

where κ is a constant.

Notice that for the BAI problem, our complexity measure H is same as the complexity measure introduced in (Audibert



(a) Time horizon T

(b) The gap ϵ

(c) Per-round knapsack b

Figure 1. The results obtained in different environments.

et al., 2010). The result recovers the tight lower bound (Carpentier & Locatelli, 2016) up to a logarithmic factor and recovers the state-of-the-art upper bound in (Karnin et al., 2013). We provide the details of the proof in Appendix E.2.

Example 5.3 (TopK and MB problem)**.** *For any $K \in [m]$, the TopK problem can be modeled by the simple OAK problem with $d = 2$ and $|\mathcal{X}^*| = K$. The consumption vector for each arm is deterministic $(b, b/K)^\top$. Note that the number of optimal arms K is known to the learner. Let $\mathcal{P} = \{\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(K)}\}$ be a partition of $[m]$. The MB problem with $\mathcal{P}$ can be modeled by the simple OAK problem with $d = K + 1$ and $|\mathcal{X}^*| = K$. The deterministic consumption vector $\boldsymbol{C}_{\cdot,i}$ for each arm $i \in \mathcal{X}^{(k)}, k \in [K]$ is deterministic with $(C_{1,i} = b, C_{k,i} = b/|\mathcal{X}^{(k)}|, C_{j,i} = 0)[j \neq 1, j \neq k]$. Note that the number of optimal arms K and the partition $\mathcal{P}$ are known to the learner. For the TopK problem or the MB problem with K partitions, Algorithm 3 makes errors with probability at most*

$$O\left(m \cdot \exp\left(-\frac{\kappa T}{H \log m}\right)\right),$$

where κ is a constant.

Notice that the multiplicative factor in Example 5.3 is $O(m)$ while the previous upper bound is $O(m^2)$ for the TopK and MB in the fixed budget setting (Bubeck et al., 2013; Chen et al., 2014). We provide the proof in Appendix E.2.

6. Numerical Evaluations

We consider a specific instance in which there are four arms ($m = 4$), three types of resources ($d = 3$), the expected per-round resource constraint of $0 < b \leq 1$, and a parameter $0 < \epsilon \leq b$. The unknown reward vector is $\boldsymbol{r} = (0.5, 0.5 - \epsilon, 0.5, 0.5)$, and the unknown expected resource consumption is represented by the matrix:

$$\boldsymbol{C} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 + \epsilon \\ b & b & b & b \end{bmatrix}.$$

We begin by considering the case where the knapsack $b = 0.2$ and the gap $\epsilon = 0.01$. Among the available arms, the ones with indices 1 and 3 are optimal, while the rest are sub-optimal. To evaluate the performance of different algorithms, we compare the probability of outputting the index set containing all optimal arms. All results are the averages over 100 runs.

Note that the traditional regret minimization algorithms are not suitable for handling the OAK setting. This can be attributed to the fact that these algorithms rely on the "optimism under uncertainty" principle and the associated confidence radius is often too large to explore the entire latent structure adequately. Consequently, the selection probability for "not too bad" arms remains high even when the game is ending. To compare our algorithms with traditional strategies, we propose a modification to the UcbBwK algorithm (Babaioff et al., 2015) that makes it more suitable for the OAK task. Specifically, we modify the algorithm to compute the LP based on the mean estimator after the last round and output optimal arms based on the solution. We refer to this new algorithm as "MeanBwK." The algorithm maintain a distribution $\mathcal{D}^{(t)}$ over arms to select arms during the times horizons, and we assume that the algorithms identify optimal arms based on the distribution of the last round, i.e., an arm i is considered optimal if and only if $\mathcal{D}_i^{(\tau)} > 10^{-3}$, where τ is the index of the last round before the algorithms stop.

The results (accuracy) obtained in different environments are summarized in Figure 1. Figure 1(a) provides results for different time horizons. It is noteworthy that our algorithms outperform MeanBwK due to their new designs tailored specifically for the OAK setting. To further evaluate the performance of our algorithms, we vary the value of ϵ while keeping $b = 0.2$ and $T = 2 \times 10^4$ fixed, and present the results in Figure 1(b). Note that for smaller ϵ, algorithms are more susceptible to errors. We also investigate the impact of different values of b on the performance of our algorithms. We conduct experiments with $\epsilon = 0.01$ and $T = 2 \times 10^4$ fixed, while also modifying the consumption to ensure that there are at least two optimal arms. The results of these experiments are presented in Figure 1(c). Based on the results, we observe that smaller values of b make it more challenging for the algorithms to identify all optimal arms, even with a longer time horizon, which is consistent with our theoretical analysis. The code for algorithms could be available at https://github.com/ShaoangLi/OAK-problem.git.

7. Related Work

7.1. Pure exploration problems

The best arm identification with the fixed budget setting was introduced by (Audibert et al., 2010). Subsequent work (Karnin et al., 2013; Carpentier & Locatelli, 2016; Chen et al., 2017c) establish the upper bound and lower bound of BAI, respectively. There are some extensions of BAI including top-K best arms identification (Kalyanakrishnan et al., 2012; Bubeck et al., 2013; Chen et al., 2017a;b; Réda et al., 2021; Zhou & Tian, 2022), θ-threshold arms identification (Locatelli et al., 2016; Mukherjee et al., 2017; Xu et al., 2020), ϵ-best arm identification (Kano et al., 2019; Katz-Samuels & Jamieson, 2020), multi-bandits best arms identification (Gabillon et al., 2011; Bubeck et al., 2013), and other variants (Abbasi-Yadkori et al., 2018; Rizk et al., 2021; Zhang & Ong, 2021; Zhong et al., 2021; Barrier et al., 2022; Wang et al., 2022). The Feasible Arms Identification (FAI) setting (Katz-Samuels & Scott, 2018; 2019) aims to identify all feasible (distribution have means belonging to the polyhedron) arms and top-K feasible arms, respectively, while the decision-maker aims to identify all optimal arms in the OAK problem. There are several fundamental differences between the OAK setting and the Feasible Arms Identification problem: (1) the expected reward vector $\boldsymbol{\mu}$ is unknown for the OAK setting but known for the FAI setting; (2) the leaner has to consider infinite possible candidate distributions satisfy $\boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{b}$ (consumption matrix $\boldsymbol{C}$ is unknown) for the OAK setting while only needs to consider finite m distributions satisfy $\boldsymbol{x}_i \in \mathcal{D}$ (feasible domain $\mathcal{D}$ is known) for the FAI setting.

7.2. Bandits with knapsacks

Another line relevant to this paper is bandits with knapsacks. The regret minimization setting of stochastic BwK was first introduced and optimally solved in (Badanidiyuru et al., 2013) to encompass application domains the learner be limited by the resource constraints. Subsequent work provide a UCB-based algorithm for BwK problem (Agrawal & Devanur, 2014) and a 'black-box reduction' from bandits to BwK (Immorlica et al., 2019). They all achieve near-optimal worst-case regret. Some work (Flajolet & Jaillet, 2015; Sankararaman & Slivkins, 2021; Ren et al., 2021; Li et al., 2021) study the problem-dependent regret of BwK. There are some other versions of BwK including budgeted bandits (Tran-Thanh et al., 2010; 2012; Ding et al., 2013; Cayci et al., 2020; Das et al., 2022), contextual bandits with knapsacks (Badanidiyuru et al., 2014; Agrawal & Devanur, 2016; Agrawal et al., 2016; Sivakumar et al., 2022; Li & Stoltz, 2022), combinatorial semi-bandits with knapsacks (Sankararaman & Slivkins, 2018), adversarial bandits with knapsacks (Immorlica et al., 2019; Kesselheim & Singla, 2020; Castiglioni et al., 2022), other variants (Liu et al., 2022b;a; Kumar & Kleinberg, 2022), and applications (Badanidiyuru et al., 2012; Babaioff et al., 2015; Li et al., 2022). The regret minimization setting aims to trade off exploration and exploitation while the OAK setting is the pure-exploration framework. As a result, the two settings require different techniques for proving lower and upper bounds.

To achieve near-optimal instance-dependent regret for BwK, (Li et al., 2021) provide an algorithm (phase I of Algorithm 1) from the primal-dual perspective that can identify the optimal arm set. However, the exploration setting is different and this primal-dual algorithm cannot handle the OAK task due to two fundamental reasons. First, the theoretical result of the primal-dual algorithm (phase I) is based on the assumption that the algorithm will not exceed the resource constraint, while the OAK problem is motivated by pure exploration with resource consumption and hard knapsacks during the learning process. Second, the stopping condition of the primal-dual algorithm depends on the pre-set confidence radius due to the "optimism under uncertainty" strategy. With a fixed pre-set confidence radius and the unknown hardness of the problem instance, the stopping time of the primal-dual algorithm is unpredictable.

8. Conclusion

We consider the optimal arms identification with knapsacks problem, which extends the best arm identification by considering resource consumption. We present a novel, parameter-free algorithm that returns optimal arms with high probability. We propose a new complexity measure for the OAK problem, which builds a bridge between the OAK and BwK problem. We provide the error upper and lower bounds for the general OAK problem based on the new complexity measure. We further investigate some special cases and the results show that the proposed algorithm recovers or improves the state-of-the-art upper bounds for some classical pure exploration problems.

Acknowledgments

The research is partially supported by National Key R&D Program of China under Grant No.2021ZD0110400, Innovation Program for Quantum Science and Technology 2021ZD0302900, China National Natural Science Foundation with No.62132018, No. 61932016, Pioneer and Leading Goose R&D Program of Zhejiang, 2023C01029, and the Fundamental Research Funds for the Central Universities WK2150110024.

References

Abbasi-Yadkori, Y., Bartlett, P., Gabillon, V., Malek, A., and Valko, M. Best of both worlds: Stochastic & adversarial best-arm identification. In *Conference on Learning Theory*, pp. 918–949. PMLR, 2018.

Agrawal, S. and Devanur, N. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29:3450–3458, 2016.

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *Conference on Economics and Computation*, pp. 989–1006. ACM, 2014.

Agrawal, S., Devanur, N. R., and Li, L. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016.

Audibert, J.-Y., Bubeck, S., and Munos, R. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pp. 41–53. Citeseer, 2010.

Babaioff, M., Dughmi, S., Kleinberg, R. D., and Slivkins, A. Dynamic pricing with limited supply. *ACM Trans. Economics and Comput.*, 3(1):4:1–4:26, 2015.

Badanidiyuru, A., Kleinberg, R., and Singer, Y. Learning on a budget: posted price mechanisms for online procurement. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 128–145. ACM, 2012.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013.

Badanidiyuru, A., Langford, J., and Slivkins, A. Resourceful contextual bandits. In *Conference on Learning Theory*, pp. 1109–1134. PMLR, 2014.

Barrier, A., Garivier, A., and Kocák, T. A non-asymptotic approach to best-arm identification for gaussian bandits. In *International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 10078–10109. PMLR, 2022.

Bubeck, S., Wang, T., and Viswanathan, N. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pp. 258–265. PMLR, 2013.

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

Castiglioni, M., Celli, A., and Kroer, C. Online learning with knapsacks: the best of both worlds. *CoRR*, abs/2202.13710, 2022.

Cayci, S., Eryilmaz, A., and Srikant, R. Budget-constrained bandits over general cost and reward distributions. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 4388–4398. PMLR, 2020.

Chen, J., Chen, X., Zhang, Q., and Zhou, Y. Adaptive multiple-arm identification. In *International Conference on Machine Learning*, pp. 722–730. PMLR, 2017a.

Chen, L., Li, J., and Qiao, M. Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pp. 101–110. PMLR, 2017b.

Chen, L., Li, J., and Qiao, M. Towards instance optimal bounds for best arm identification. In *Conference on Learning Theory*, pp. 535–592. PMLR, 2017c.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 27, pp. 379–387, 2014.

Das, D., Jain, S., and Gujar, S. Budgeted combinatorial multi-armed bandits. In *International Conference on Autonomous Agents and Multiagent Systems*, pp. 345–353, 2022.

Ding, W., Qin, T., Zhang, X.-D., and Liu, T.-Y. Multi-armed bandit with budget constraint and variable costs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2013.

Flajolet, A. and Jaillet, P. Logarithmic regret bounds for bandits with knapsacks. *arXiv preprint arXiv:1510.01800*, 2015.

Gabillon, V., Ghavamzadeh, M., Lazaric, A., and Bubeck, S. Multi-bandit best arm identification. *Advances in Neural Information Processing Systems*, 24, 2011.

Immorlica, N., Sankararaman, K. A., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. In *Symposium on Foundations of Computer Science*, pp. 202–219. IEEE, 2019.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. Pac subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, volume 12, pp. 655–662, 2012.

Kano, H., Honda, J., Sakamaki, K., Matsuura, K., Nakamura, A., and Sugiyama, M. Good arm identification via bandit feedback. *Machine Learning*, 108(5):721–745, 2019.

Karnin, Z., Koren, T., and Somekh, O. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 1238–1246. PMLR, 2013.

Katz-Samuels, J. and Jamieson, K. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1781–1791. PMLR, 2020.

Katz-Samuels, J. and Scott, C. Feasible arm identification. In *International Conference on Machine Learning*, pp. 2535–2543. PMLR, 2018.

Katz-Samuels, J. and Scott, C. Top feasible arm identification. In *International Conference on Artificial Intelligence and Statistics*, pp. 1593–1601. PMLR, 2019.

Kesselheim, T. and Singla, S. Online learning with vector costs and bandits with knapsacks. In *Conference on Learning Theory*, pp. 2286–2305. PMLR, 2020.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In Dwork, C. (ed.), *Symposium on Theory of Computing*, pp. 681–690. ACM, 2008.

Kumar, R. and Kleinberg, R. Non-monotonic resource utilization in the bandits with knapsacks problem. In *Advances in Neural Information Processing Systems*, 2022.

Li, S., Zhang, L., and Li, X. Online pricing with limited supply and time-sensitive valuations. In *IEEE Conference on Computer Communications*, pp. 860–869. IEEE, 2022.

Li, X., Sun, C., and Ye, Y. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *International Conference on Machine Learning*, pp. 6483–6492. PMLR, 2021.

Li, Z. and Stoltz, G. Contextual bandits with knapsacks for a conversion model. In *Advances in Neural Information Processing Systems*, 2022.

Liu, Q., Xu, W., Wang, S., and Fang, Z. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. In *Advances in Neural Information Processing Systems*, 2022a.

Liu, S., Jiang, J., and Li, X. Non-stationary bandits with knapsacks. In *Advances in Neural Information Processing Systems*, 2022b.

Locatelli, A., Gutzeit, M., and Carpentier, A. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pp. 1690–1698. PMLR, 2016.

Megiddo, N. and Chandrasekaran, R. On the ε-perturbation method for avoiding degeneracy. *Operations Research Letters*, 8(6):305–308, 1989.

Mukherjee, S., Naveen, K. P., Sudarsanam, N., and Ravindran, B. Thresholding bandits with augmented ucb. *arXiv preprint arXiv:1704.02281*, 2017.

Réda, C., Kaufmann, E., and Delahaye-Duriez, A. Top-m identification for linear bandits. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 1108–1116. PMLR, 2021.

Ren, W., Liu, J., and Shroff, N. B. On logarithmic regret for bandits with knapsacks. In *Conference on Information Sciences and Systems*, pp. 1–6. IEEE, 2021.

Rizk, G., Thomas, A., Colin, I., Laraki, R., and Chevaleyre, Y. Best arm identification in graphical bilinear bandits. In *International Conference on Machine Learning*, pp. 9010–9019. PMLR, 2021.

Sankararaman, K. A. and Slivkins, A. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1760–1770. PMLR, 2018.

Sankararaman, K. A. and Slivkins, A. Bandits with knapsacks beyond the worst case. *Advances in Neural Information Processing Systems*, 34, 2021.

Sivakumar, V., Zuo, S., and Banerjee, A. Smoothed adversarial linear contextual bandits with knapsacks. In *International Conference on Machine Learning*, volume 162, pp. 20253–20277. PMLR, 2022.

Tran-Thanh, L., Chapman, A., De Cote, E. M., Rogers, A., and Jennings, N. R. Epsilon–first policies for budget–limited multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2010.

Tran-Thanh, L., Chapman, A., Rogers, A., and Jennings, N. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 1134–1140, 2012.

Wang, Z., Wagenmaker, A. J., and Jamieson, K. G. Best arm identification with safety constraints. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9114–9146. PMLR, 2022.

Xu, Y., Chen, X., Singh, A., and Dubrawski, A. Thresholding bandit problem with both duels and pulls. In *International Conference on Artificial Intelligence and Statistics*, pp. 2591–2600. PMLR, 2020.

Zhang, M. and Ong, C. S. Quantile bandits for best arms identification. In *International Conference on Machine Learning*, pp. 12513–12523. PMLR, 2021.

Zhong, Z., Cheung, W. C., and Tan, V. Probabilistic sequential shrinking: A best arm identification algorithm for stochastic bandits with corruptions. In *International Conference on Machine Learning*, pp. 12772–12781. PMLR, 2021.

Zhou, R. and Tian, C. Approximate top-m arm identification with heterogeneous reward variances. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7483–7504. PMLR, 2022.

A. Analysis of Complexity Measure (Theorem 3.1)

After removing all non-active constraints of (1), we can get a standard form LP. Formally, base on the matrix $\boldsymbol{C}$, by arranging the $|\mathcal{X}^*|$ basic columns and $|\mathcal{Y}^*|$ active rows next to each other, we obtain a $|\mathcal{Y}^*| \times |\mathcal{X}^*|$ optimal basis matrix corresponding to $\boldsymbol{x}^*$ and let $\boldsymbol{P}$ denote the basis matrix. Similarly, by arranging the $|\mathcal{X}'|$ non-basic columns and $|\mathcal{Y}^*|$ active rows next to each other, we obtain a $|\mathcal{Y}^*| \times |\mathcal{X}'|$ non-basis matrix and let $\boldsymbol{Q}$ denote it. Then we could construct a matrix $\boldsymbol{A} := [\boldsymbol{P}|\boldsymbol{Q}]$ with $|\mathcal{X}^*|$ linearly independent columns (rows). We use $\boldsymbol{b}^*$ to denote the vector $(b, \ldots, b)^\top \in (0,1]^{|\mathcal{X}^*|}$. Then the standard form LP is

$$\begin{aligned} \max \quad & \boldsymbol{\mu}^\top \boldsymbol{x}, \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}^*, \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned} \tag{6}$$

It is oblivious that (1) and (6) have same feasible domain $\mathcal{D}$, optimal solution $\boldsymbol{x}^*$, and optimal value OPT_{LP}. The dual problem of (6) is

$$\begin{aligned} \min \quad & (\boldsymbol{b}^*)^\top \boldsymbol{w}, \\ \text{s.t.} \quad & \boldsymbol{A}^\top \boldsymbol{w} \geq \boldsymbol{\mu}, \boldsymbol{w} \geq \boldsymbol{0}. \end{aligned}$$

And $\boldsymbol{w}_B^*$ is the optimal solution of it. We define a new $m \times m$ square matrix:

$$\boldsymbol{M} := \begin{bmatrix} \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}.$$

We use $\boldsymbol{I}$ to denote the identity matrix. We also get the inverse of $\boldsymbol{M}$:

$$\boldsymbol{M}^{-1} := \begin{bmatrix} \boldsymbol{P}^{-1} & -\boldsymbol{P}^{-1}\boldsymbol{Q} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}.$$

Define a m dimension vector

$$\boldsymbol{d}_q := (\boldsymbol{M}^{-1})_{\cdot,q}, q \in \mathcal{X}'.$$

Notice that $\boldsymbol{x}^*$ is the optimal extreme point in $\mathcal{D}$. Under the Assumption 2.1, there are $|\mathcal{X}'|$ neighbors of $\boldsymbol{x}^*$. And $\boldsymbol{d}_q$ is the edge direction from $\boldsymbol{x}^*$ leading to the adjacent extreme points $\boldsymbol{x}^{(q)}$ corresponding to the increase of the sub-optimal variable $q \in \mathcal{X}'$. We use $\boldsymbol{x}^{(q)}$ to denote the q-th optimal adjacent vertex, i.e.,

$$\boldsymbol{\mu}^\top \boldsymbol{x}^* > \boldsymbol{\mu}^\top \boldsymbol{x}^{(1)} \geq \ldots \geq \boldsymbol{\mu}^\top \boldsymbol{x}^{(q)} \geq \ldots, q \in \mathcal{X}'.$$

We define the *adjacent gap* $\Delta^{(q)}$ as

$$\Delta^{(q)} = \boldsymbol{\mu}^\top \boldsymbol{x}^* - \boldsymbol{\mu}^\top \boldsymbol{x}^{(q)}, q \in \mathcal{X}'.$$

Then consider the relationship between $\Delta^{(q)}$ and $R_{(q)}$. Let d_{qi} denote the i-th element of $\boldsymbol{d}_q$. Define

$$\alpha_q = \min_{i \in \mathcal{X}^*} \left\{ \frac{x_i^*}{-d_{qi}} \right\},$$

we have

$$\begin{aligned} \Delta^{(q)} &= -\alpha_q \boldsymbol{\mu}^\top \boldsymbol{d}_q = -\alpha_q \big(\mu_q - \big(\boldsymbol{w}_B^*\big)^\top \boldsymbol{Q}_{\cdot,q}\big) \\ &= -\alpha_q \big(\mu_q - \big(\boldsymbol{w}^*\big)^\top \boldsymbol{C}_{\cdot,q}\big) = \alpha_q R_{(q)}. \end{aligned}$$

From the geometry of linear programming, α_q is the distance between the vertex $\boldsymbol{x}^*$ and $\boldsymbol{x}^{(q)}$. The distance of the polyhedron (i.e., the feasible domain of (1)) will be not more than $\sqrt{2}$. From the definition of vertex gap and adjacent gap, we have $\Delta_{q+1} \leq \Delta^{(q)}$. Combine the two facts together, we obtain $R_{(i)} \geq \frac{\Delta_{i+1}}{\sqrt{2}}$.

Then consider LP (3), notice that the feasible domain of (3) is a subset of $\mathcal{D}$ and the optimal extreme point of (3) is also a vertex of $\mathcal{D}$. The feasible domain of (3) is nonempty because of the existence of null arms. Under Assumption 2.1, different LP (3) with different absent optimal arms have different vertex. Based on the definition of vertex gap, we could conclude that $G_{(i)} \geq \Delta_{i+1}$.

B. Analysis of Algorithm 1 (Theorem 3.2)

In this section, we analyze the probability bound of Algorithm 1 and prove Theorem 3.2.

B.1. Main lemmas

Lemma B.1. *At the end of phase p, for the optimal value of (4), if the algorithm makes no errors before the beginning of phase p, with probability at least $1-2md\cdot\exp\left(-2\delta_p^2 s(p)\right)$), we have*

$$\bar{\boldsymbol{\mu}}^\top \bar{\boldsymbol{x}}^*(p) \geq \left(1-\frac{\delta_p}{b}\right)\mathrm{OPT}_{\mathrm{LP}} - \delta_p.$$

Proof. According to the Hoeffding's inequality, with probability at least $1-2md\cdot\exp\left(-2\delta_p^2 s(p)\right)$, for all optimal arms i^* and all active constraints j^* in (1) and (4), we have

$$\begin{aligned}
&\max(\mu_{i^*}-\delta_p,0)\leq\bar{\mu}_{i^*}(p)\leq\min(1,\mu_{i^*}+\delta_p),\\
&\max(C_{j^*,i^*}-\delta_p,0)\leq\bar{C}_{j^*,i^*}(p)\leq\min(1,C_{j^*,i^*}+\delta_p).
\end{aligned}$$

Under this event, consider the following linear program

$$\begin{aligned}
\max \quad & \boldsymbol{\mu}^\top\boldsymbol{x},\\
\text{s.t.} \quad & \boldsymbol{x}^\top\boldsymbol{C}_{j,\cdot}^\top\leq b-\delta_p,\forall j\in[d],\\
& \boldsymbol{x}\geq\mathbf{0}.
\end{aligned}$$

Let $\boldsymbol{x}^*(\boldsymbol{b}^L)$ and $\mathrm{OPT}_{\mathrm{LP}}(\boldsymbol{b}^L)$ denote the optimal solution and optimal value of it, respectively. According to (Agrawal & Devanur, 2014), we have

$$\mathrm{OPT}_{\mathrm{LP}}(\boldsymbol{b}^L)\geq\left(1-\frac{\delta_p}{b}\right)\mathrm{OPT}_{\mathrm{LP}}.$$

Then we show that $\boldsymbol{x}^*(\boldsymbol{b}^L)$ is also a feasible solution for $LP(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{C}})$ of phase p. Consider the j-th resource

$$\begin{aligned}
&\sum_{i=1}^m\bar{C}_{j,i}x_i^*(\boldsymbol{b}^L)\\
&=\sum_{i=1}^m(\bar{C}_{j,i}-C_{j,i})x_i^*(\boldsymbol{b}^L)+\sum_{i=1}^m C_{j,i}x_i^*(\boldsymbol{b}^L)\\
&\leq\max_{i\in[m]}\left(\bar{C}_{j,i}-C_{j,i}\right)+(b-\delta_p)\\
&\leq b.
\end{aligned}$$

With this feasibility, we have

$$\begin{aligned}
&\bar{\boldsymbol{\mu}}^\top\boldsymbol{x}^*(\boldsymbol{b}^L)\\
&=\boldsymbol{\mu}^\top\boldsymbol{x}^*(\boldsymbol{b}^L)-\left(\boldsymbol{\mu}^\top\boldsymbol{x}^*(\boldsymbol{b}^L)-\bar{\boldsymbol{\mu}}^\top\boldsymbol{x}^*(\boldsymbol{b}^L)\right)\\
&\geq\mathrm{OPT}_{\mathrm{LP}}(\boldsymbol{b}^L)-\left\|\boldsymbol{\mu}^\top-\bar{\boldsymbol{\mu}}^\top\right\|_\infty\cdot\left\|\boldsymbol{x}^*(\boldsymbol{b}^L)\right\|_1\\
&\geq\left(1-\frac{\delta_p}{b}\right)\mathrm{OPT}_{\mathrm{LP}}-\delta_p.
\end{aligned}$$

Then we get

$$\bar{\boldsymbol{\mu}}^\top\bar{\boldsymbol{x}}^*(p)\geq\bar{\boldsymbol{\mu}}^\top\boldsymbol{x}^*(\boldsymbol{b}^L)\geq\left(1-\frac{\delta_p}{b}\right)\mathrm{OPT}_{\mathrm{LP}}-\delta_p.$$

□

Lemma B.2. *At the end of phase p, for the optimal value of (4), if one of optimal arms is not active in $\bar{\boldsymbol{x}}^*(p)$, with probability at least $1-2md\cdot\exp\left(-2\delta_p^2 s(p)\right)$), we have*

$$\bar{\boldsymbol{\mu}}^\top\bar{\boldsymbol{x}}^*(p)\leq\left(1+\frac{\delta_p}{b}\right)(\mathrm{OPT}_{\mathrm{LP}}-\Delta_2)+\delta_p.$$

Proof. According to Theorem 3.1, for any arm $i \in [m]$, we have

$$\text{OPT}_{\text{LP}}^{-i} \leq \text{OPT}_{\text{LP}} - \Delta_2. \tag{7}$$

According to the Hoeffding's inequality, with probability at least $1 - 2md \cdot \exp\left(-2\delta_p^2 s(p)\right)$, for all optimal arms i^* and all active constraints j^* in (1) and (4), we have

$$\begin{aligned}
&\max(\mu_{i^*} - \delta_p, 0) \leq \bar{\mu}_{i^*}(p) \leq \min(1, \mu_{i^*} + \delta_p),\\
&\max(C_{j^*,i^*} - \delta_p, 0) \leq \bar{C}_{j^*,i^*}(p) \leq \min(1, C_{j^*,i^*} + \delta_p).
\end{aligned}$$

Under this event, assume one of the optimal arm $i^* \in \mathcal{X}^*$ is eliminated in round t. Consider the following linear program

$$\begin{aligned}
\max \quad & \boldsymbol{\mu}^\top \boldsymbol{x},\\
\text{s.t.} \quad & \boldsymbol{x}^\top \boldsymbol{C}_{j,\cdot}^\top \leq b + \delta_p, \forall j \in [d],\\
& \boldsymbol{x} \geq \mathbf{0}, x_{i^*} = 0.
\end{aligned} \tag{8}$$

Let $\boldsymbol{x}^{-i*}(\boldsymbol{b}^U)$ and $\text{OPT}_{\text{LP}}^{i^*}(\boldsymbol{b}^U)$ denote the optimal solution and optimal value of it, respectively. According to (Agrawal & Devanur, 2014), we have

$$\text{OPT}_{\text{LP}}^{i^*} \geq \frac{b}{b + \delta_p} \text{OPT}_{\text{LP}}^{i^*}(\boldsymbol{b}^U).$$

Combine it with Theorem 3.1, we get

$$\text{OPT}_{\text{LP}}^{i^*}(\boldsymbol{b}^U) \leq \left(1 + \frac{\delta_p}{b}\right)\left(\text{OPT}_{\text{LP}} - \Delta_2\right).$$

Then we show that $\bar{\boldsymbol{x}}^*(p)$ is also a feasible solution for (8) of phase p. Consider the j-th resource

$$\begin{aligned}
&\sum_{i=1}^m C_{j,i} \bar{x}_i^*(p)\\
&= \sum_{i=1}^m (C_{j,i} - \bar{C}_{j,i}) \bar{x}_i^*(p) + \sum_{i=1}^m \bar{C}_{j,i} \bar{x}_i^*(p)\\
&\leq \max_{i \in [m]} \left(C_{j,i} - \bar{C}_{j,i}\right) + b\\
&\leq b + \delta_p.
\end{aligned}$$

With this feasibility, we have

$$\begin{aligned}
&\bar{\boldsymbol{\mu}}^\top \bar{\boldsymbol{x}}^*(p)\\
&= \boldsymbol{\mu}^\top \bar{\boldsymbol{x}}^*(p) + \left(\bar{\boldsymbol{\mu}}^\top \bar{\boldsymbol{x}}^*(p) - \boldsymbol{\mu}^\top \bar{\boldsymbol{x}}^*(p)\right)\\
&\leq \text{OPT}_{\text{LP}}^{i^*}(\boldsymbol{b}^U) + \left\|\bar{\boldsymbol{\mu}}^\top - \boldsymbol{\mu}^\top\right\|_\infty \cdot \|\bar{\boldsymbol{x}}^*(p)\|_1\\
&\leq \left(1 + \frac{\delta_p}{b}\right)\left(\text{OPT}_{\text{LP}} - \Delta_2\right) + \delta_p.
\end{aligned}$$

Then we complete the proof. □

Lemma B.3. *At the end of phase p, for any optimal arm $i^* \in \mathcal{X}^*$ and any active sub-optimal arm $i' \in \mathcal{X}' \cap \mathcal{X}_p$, the probability that $\bar{R}_{i^*}(p) > \bar{R}_{i'}(p)$ is at most*

$$2md \cdot \exp\left(-2\left(\frac{b\Delta_2}{8} + \frac{b^2 R_{i'}}{8}\right)^2 s(p)\right).$$

Proof. Consider the dual form

$$\begin{aligned}
\min \quad & \boldsymbol{b}^\top \boldsymbol{w},\\
\text{s.t.} \quad & \bar{\boldsymbol{C}}^\top \boldsymbol{w} \geq \bar{\boldsymbol{\mu}}, \boldsymbol{w} \geq \mathbf{0}.
\end{aligned}$$

Let $\bar{\boldsymbol{w}}^*(p)$ denote the optimal solution of it, according to Lemma B.1 and B.2, we have

$$\left(1-\frac{\delta_p}{b}\right)\boldsymbol{b}^\top\boldsymbol{w}^*-\delta_p \le \boldsymbol{b}^\top\bar{\boldsymbol{w}}^*(p) \le \left(1+\frac{\delta_p}{b}\right)\left(\boldsymbol{b}^\top\boldsymbol{w}^*-\Delta_2\right)+\delta_p.$$

Then we get

$$\begin{aligned}
&\sum_{j=1}^{d}(\bar{w}^*(p))_j-\sum_{j=1}^{d}w_j^* \ge -\frac{\delta_p}{b}\left(1+\sum_{j=1}^{d}w_j^*\right),\\
&\sum_{j=1}^{d}(\bar{w}^*(p))_j-\sum_{j=1}^{d}w_j^* \le \frac{\delta_p}{b}\left(\sum_{j=1}^{d}w_j^*-\frac{\Delta_2}{b}\right)+\frac{\delta_p-\Delta_2}{b}.
\end{aligned} \tag{9}$$

Consider the R_i of i^* and i'

$$\begin{aligned}
R_{i^*} &= (\boldsymbol{w}^*)^\top\boldsymbol{C}_{i^*}-\mu_{i^*} &&\le 0,\\
\bar{R}_{i^*} &= (\bar{\boldsymbol{w}}^*(p))^\top\bar{\boldsymbol{C}}_{i^*}-\bar{\mu}_{i^*} &&> 0,\\
\bar{R}_{i'} &= (\bar{\boldsymbol{w}}^*(p))^\top\bar{\boldsymbol{C}}_{i'}-\bar{\mu}_{i'} &&< \bar{R}_{i^*}.
\end{aligned} \tag{10}$$

From (9) and (10), we have

$$\begin{aligned}
&\sum_{j=1}^{d}(\bar{w}^*(p))_j\bar{C}_{j,i^*}-\bar{\mu}_{i^*}\\
&\le \sum_{j=1}^{d}(\bar{w}^*(p))_j\left(C_{j,i^*}+\delta_p\right)-(\mu_{i^*}-\delta_p)\\
&= \sum_{j=1}^{d}w_j^*\left(C_{j,i^*}+\delta_p\right)-\left[\sum_{j=1}^{d}w_j^*\left(C_{j,i^*}+\delta_p\right)-\sum_{j=1}^{d}(\bar{w}^*(p))_j\left(C_{j,i^*}+\delta_p\right)\right]-(\mu_{i^*}-\delta_p)\\
&\le \sum_{j=1}^{d}w_j^*\left(C_{j,i^*}+\delta_p\right)+\frac{\delta_p}{b}\left(\sum_{j=1}^{d}w_j^*-\frac{\Delta_2}{b}\right)+\frac{\delta_p-\Delta_2}{b}-(\mu_{i^*}-\delta_p)\\
&\le \left(\delta_p+\frac{\delta_p}{b}\right)\left(1+\sum_{j=1}^{d}w_j^*\right)-\frac{\Delta_2}{b}-\frac{\delta_p\Delta_2}{b^2}.
\end{aligned}$$

And

$$\begin{aligned}
&\sum_{j=1}^{d}(\bar{w}^*(p))_j\bar{C}_{j,i'}-\bar{\mu}_{i'}\\
&\ge \sum_{j=1}^{d}(\bar{w}^*(p))_j(C_{j,i'}-\delta_p)-(\mu_{i'}+\delta_p)\\
&= \sum_{j=1}^{d}w_j^*(C_{j,i'}-\delta_p)-\left[\sum_{j=1}^{d}w_j^*\left(C_{j,i^*}-\delta_p\right)-\sum_{j=1}^{d}(\bar{w}^*(p))_j\left(C_{j,i^*}-\delta_p\right)\right]-(\mu_{i'}+\delta_p)\\
&\ge \sum_{j=1}^{d}w_j^*(C_{j,i'}-\delta_p)-\frac{\delta_p}{b}\left(1+\sum_{j=1}^{d}w_j^*\right)-(\mu_{i'}+\delta_p)\\
&\ge R_{i'}-\left(\delta_p+\frac{\delta_p}{b}\right)\left(1+\sum_{j=1}^{d}w_j^*\right).
\end{aligned}$$

According to the Strong duality theorem, we have

$$\sum_{j=1}^{d} w_j^* = \frac{1}{b}\sum_{i=1}^{m} \mu_i x_i^* = \frac{1}{b}\text{OPT}_{\text{LP}} \leq \frac{1}{b}.$$

From the Hoeffding's inequality, Lemma B.1 and B.2, we have

$$\begin{aligned}
&\mathbb{P}[\bar{R}_{i^*}(p) > \bar{R}_{i'}(p)]\\
&\leq \mathbb{P}\left[\frac{\Delta_2}{b} + \frac{\delta_p \Delta_2}{b^2} + R_{i'} < 2\left(\delta_p + \frac{\delta_p}{b}\right)\left(1 + \sum_{j=1}^{d} w_j^*\right)\right]\\
&\leq \mathbb{P}\left[\frac{\Delta_2}{b} + R_{i'} < 2\delta_p\left(1 + \frac{1}{b}\right)^2\right]\\
&\leq 2md \cdot \exp\left(-2\left(\frac{b\Delta_2}{8} + \frac{b^2 R_{i'}}{8}\right)^2 s(p)\right).
\end{aligned}$$

Then we complete the proof. □

Lemma B.4. *During phase p, for any optimal arm $i^* \in \mathcal{X}^*$ and any active sub-optimal arm $i' \in \mathcal{X}' \cap \mathcal{X}_p$, the probability that $\bar{G}_{i^*}(p) < \bar{G}_{i'}(p)$ is at most*

$$2md \cdot \exp\left(-\frac{2b^2 G_{i^*}^2}{9} s(p)\right).$$

Proof. During phase p, for each arm $i \in [m]$, consider the linear programming

$$\begin{aligned}
\max \quad & \bar{\boldsymbol{\mu}}^\top \boldsymbol{x},\\
\text{s.t.} \quad & \bar{\boldsymbol{C}}\boldsymbol{x} \leq \boldsymbol{b},\\
& \boldsymbol{x} \geq \boldsymbol{0}, x_i = 0.
\end{aligned}$$

Let $\overline{\text{OPT}}_{\text{LP}}^{-i}$ denote the optimal value of it. According to Lemma B.1 and the proof of Lemma B.2, with probability at least $1 - 2|\bar{\mathcal{P}}_p \cup \mathcal{P}_p \cup \mathcal{X}_p^*|^2 \cdot \exp\left(-2\delta_p^2 s(p)\right)$, we have

$$\begin{aligned}
\overline{\text{OPT}}_{\text{LP}}^{-i'} &\geq \left(1 - \frac{\delta_p}{b}\right)\text{OPT}_{\text{LP}} - \delta_p,\\
\overline{\text{OPT}}_{\text{LP}}^{-i^*} &\leq \left(1 + \frac{\delta_p}{b}\right)(\text{OPT}_{\text{LP}} - G_{i*}) + \delta_p.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{P}[\bar{G}_{i^*}(p) < \bar{G}_{i'}(p)]\\
&\leq \mathbb{P}\left[\left(1 + \frac{\delta_p}{b}\right)(\text{OPT}_{\text{LP}} - G_{i^*}) + \delta_p \geq \left(1 - \frac{\delta_p}{b}\right)\text{OPT}_{\text{LP}} - \delta_p\right]\\
&\leq \mathbb{P}\left[\left(1 + \frac{2(\text{OPT}_{\text{LP}} - G_{i^*})}{b}\right)\delta_p \geq G_{i^*}\right]\\
&\leq 2md \cdot \exp\left(-\frac{2b^2 G_{i^*}^2}{9} s(p)\right).
\end{aligned}$$

□

B.2. Proof of Theorem 3.2

For all arms in $\mathcal{X}_p$, let $\mathcal{P}_p$ and $\bar{\mathcal{P}}_p$ denote the set of optimal arms for (1) and (4), $\mathcal{Q}_p$ and $\bar{\mathcal{Q}}_p$ denote the set of sub-optimal arms for (1) and (4), respectively. Consider the first phase p such that there is at least one eliminated optimal arm or one sub-optimal arm is added to $\mathcal{X}_p^*$. Let S_p^* denote the $\frac{1}{16}|\mathcal{X}_p|$ arms with smallest G_i and S_p' denote the $\frac{1}{16}|\mathcal{X}_p|$ arms with largest R_i, respectively. Define

$$\begin{aligned}\Phi_p^* &:= \max_{i\in\mathcal{P}_p\backslash S_p^*} \exp\left(-\frac{2b^2G_i^2}{9}s(p)\right),\\ \Phi_p' &:= \max_{i\in\mathcal{Q}_p\backslash S_p'} \exp\left(-2\left(\frac{b\Delta_2}{8}+\frac{b^2R_i}{8}\right)^2 s(p)\right).\end{aligned}$$

Consider the case that $\bar{\mathcal{Q}}_p > \bar{\mathcal{P}}_p$. Consider the number of arms in $\bar{\mathcal{Q}}_p \cap (\mathcal{P}_p\backslash S_p^*)$, then

$$\begin{aligned}\mathbb{E}[|\bar{\mathcal{Q}}_p\cap(\mathcal{P}_p\backslash S_p^*)|] &= \sum_{i\in\mathcal{P}_p\backslash S_p^*}\mathbb{P}[\bar{G}_i(p)<\bar{G}_{i'}(p)]\\ &\le \sum_{i\in\mathcal{P}_p\backslash S_p^*} 2md\cdot\exp\left(-\frac{2b^2G_{i^*}^2}{9}s(p)\right)\\ &\le 2md\cdot|\mathcal{P}_p\backslash S_p^*|\cdot\Phi_p^*.\end{aligned}$$

Then we apply Markov's inequality

$$\begin{aligned}&\mathbb{P}[|\bar{\mathcal{Q}}_p\cap(\mathcal{P}_p\backslash S_p^*)|>\frac{1}{8}|\bar{\mathcal{Q}}_p|]\\ &\le \frac{8\mathbb{E}[|\bar{\mathcal{Q}}_p\cap(\mathcal{P}_p\backslash S_p^*)]}{|\bar{\mathcal{Q}}_p|}\\ &\le 16md\cdot\frac{|\mathcal{P}_p\backslash S_p^*|}{|\bar{\mathcal{Q}}_p|}\Phi_p^*.\end{aligned} \tag{11}$$

Then we have

$$\mathbb{P}[|\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p|>\frac{3}{4}|\bar{\mathcal{Q}}_p|]\ge 1-16md\cdot\frac{|\mathcal{P}_p\backslash S_p^*|}{|\bar{\mathcal{Q}}_p|}\Phi_p^*. \tag{12}$$

Based on this event, let i_p^* denote the eliminated optimal arm in phase p. Consider the the number of arms in $(\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p)\backslash S_p'$ with larger $\bar{R}_x$ than that of the eliminated optimal arm and let N_p' denote it, then

$$\begin{aligned}\mathbb{E}[N_p'] &= \sum_{i\in(\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p)\backslash S_p'}\mathbb{P}[\bar{R}_{i_p^*}(p)<\bar{R}_i(p)]\\ &\le \sum_{i\in(\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p)\backslash S_p'} 2md\cdot\exp\left(-2\left(\frac{b\Delta_2}{8}+\frac{b^2R_i}{8}\right)^2 s(p)\right)\\ &\le 2md\cdot|(\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p)\backslash S_p'|\cdot\max_{i\in(\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p)\backslash S_p'}\exp\left(-2\left(\frac{b\Delta_2}{8}+\frac{b^2R_i}{8}\right)^2 s(p)\right).\end{aligned}$$

Then we apply Markov's inequality

$$\begin{aligned}\mathbb{P}[N_p'>\frac{1}{6}|\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p|] &\le \frac{6\mathbb{E}[N_p']}{|\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p|}\\ &\le 12md\cdot\max_{i\in(\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p)\backslash S_p'}\exp\left(-2\left(\frac{b\Delta_2}{8}+\frac{b^2R_i}{8}\right)^2 s(p)\right).\end{aligned}$$

Then we obtain that the probability that there is at least one eliminated optimal arms is at most

$$32md\cdot\Phi_p^*+12md\cdot\Phi_p'.$$

Consider the case that $\bar{\mathcal{P}}_p > \bar{\mathcal{Q}}_p$. Similarly, we have

$$\begin{aligned}
&\mathbb{P}[|\bar{\mathcal{P}}_p \cap (\mathcal{Q}_p \backslash S'_p)| > \frac{1}{8}|\bar{\mathcal{P}}_p|] \\
&\leq \frac{8\mathbb{E}[|\bar{\mathcal{P}}_p \cap (\mathcal{Q}_p \backslash S'_p)]}{|\bar{\mathcal{P}}_p|} \\
&\leq 16md \cdot \frac{|\mathcal{Q}_p \backslash S'_p|}{|\bar{\mathcal{P}}_p|}\Phi'_p,
\end{aligned} \tag{13}$$

and

$$\mathbb{P}[|\bar{\mathcal{P}}_p \cap \mathcal{P}_p| > \frac{3}{4}|\bar{\mathcal{P}}_p|] \geq 1 - 16md \cdot \frac{|\mathcal{Q}_p \backslash S'_p|}{|\bar{\mathcal{P}}_p|}\Phi'_p. \tag{14}$$

Let i'_t denote the added sub-optimal arm in phase p. Consider the the number of arms in $(\bar{\mathcal{P}}_p \cap \mathcal{P}_p) \backslash S^*_p$ with smaller $\bar{G}_x$ than that of the sub-optimal arm i' and let N^*_t denote it, then

$$\begin{aligned}
\mathbb{E}[N^*_t] &= \sum_{i \in (\bar{\mathcal{P}}_p \cap \mathcal{P}_p) \backslash S^*_p} \mathbb{P}[\bar{G}_i(p) < \bar{G}_{i'}(p)] \\
&\leq \sum_{i \in (\bar{\mathcal{P}}_p \cap \mathcal{P}_p) \backslash S^*_p} 2md \cdot \exp\left(-\frac{2b^2 G_i^2}{9} s(p)\right) \\
&\leq 2md \cdot |\bar{\mathcal{P}}_p \cap \mathcal{P}_p| \cdot \max_{i \in (\bar{\mathcal{P}}_p \cap \mathcal{P}_p) \backslash S^*_p} \exp\left(-\frac{2b^2 G_i^2}{9} s(p)\right).
\end{aligned}$$

Then we apply Markov's inequality

$$\begin{aligned}
&\mathbb{P}[N^*_t > \frac{1}{6}|\bar{\mathcal{P}}_p \cap \mathcal{P}_p|] \\
&\leq \frac{6\mathbb{E}[N'_p]}{|\bar{\mathcal{P}}_p \cap \mathcal{P}_p|} \\
&\leq 12|\bar{\mathcal{P}}_p \cup \mathcal{P}_p \cup \mathcal{X}^*_p|^2 \cdot \max_{i \in (\bar{\mathcal{P}}_p \cap \mathcal{P}_p) \backslash S^*_p} \exp\left(-\frac{2b^2 G_i^2}{9} s(p)\right).
\end{aligned}$$

Then the probability that at least one sub-optimal arm is added to $\mathcal{X}^*_p$ is at most

$$32md \cdot \Phi'_p + 12md \cdot \Phi^*_p.$$

In conclude, the probability that the algorithm makes mistakes in round t is at most

$$32md \cdot (\Phi'_p + \Phi^*_p). \tag{15}$$

Next, we analyse the algorithm to bound the probability of making mistakes over all rounds. Clearly, the algorithm does not exceed the budget B.

Consider the pulls for each arm in $\mathcal{X}_p$ before phase $p+1$

$$\begin{aligned}
s(p) &\geq \frac{B}{\log_{4/3} m} \sum_{k=0}^{p} \frac{1}{|\mathcal{X}_p + \mathcal{X}^*_p|} \\
&\geq \frac{B}{\log_{4/3} m} \sum_{k=0}^{p} \min\left(\frac{2}{|\mathcal{X}_p|}, \frac{2}{|\mathcal{X}^*_p|}\right).
\end{aligned}$$

Let $i_p = \frac{m}{16}\left(\frac{3}{4}\right)^p$, then we have

$$\begin{aligned}
\Phi_p^* &= \max_{i\in\mathcal{P}_p\setminus S_p^*} \exp\left(-\frac{2b^2G_i^2}{9}s(p)\right) \\
&\le \exp\left(-\frac{2b^2\Delta_{i_p}^2}{9}s(p)\right) \\
&\le \exp\left(-\frac{4b^2\Delta_{i_p}^2}{9}\frac{(p+1)B}{|\mathcal{X}_p^*|\log_{4/3} m}\right) + \exp\left(-\frac{4b^2\Delta_{i_p}^2}{9}\frac{B}{m\log_{4/3} m}\left(\frac{4}{3}\right)^t\right) \\
&= \exp\left(-\frac{4b^2\Delta_{i_p}^2}{9i_p}\frac{i_p(p+1)B}{|\mathcal{X}_p^*|\log_{4/3} m}\right) + \exp\left(-\frac{4b^2\Delta_{i_p}^2}{9i_p}\frac{B}{16\log_{4/3} m}\right) \\
&\le \exp\left(-\frac{4b^2}{9H}\frac{B}{|\mathcal{X}^*|}\right) + \exp\left(-\frac{4b^2}{9H}\frac{B}{16\log_{4/3} m}\right),
\end{aligned}\tag{16}$$

and

$$\begin{aligned}
\Phi_p' &= \max_{i\in\mathcal{Q}_p\setminus S_p'} \exp\left(-2\left(\frac{b\Delta_2}{8}+\frac{b^2R_i}{8}\right)^2 s(p)\right) \\
&\le \exp\left(-2\left(\frac{b\Delta_2}{8}+\frac{b^2\Delta_{i_p}}{8\sqrt{2}}\right)^2 s(p)\right) \\
&\le \exp\left(-2\left(\frac{b\Delta_2}{8}+\frac{b^2\Delta_{i_p}}{8\sqrt{2}}\right)^2 \frac{(p+1)B}{|\mathcal{X}_p^*|\log_{4/3} m}\right) + \exp\left(-4\left(\frac{b\Delta_2}{8}+\frac{b^2\Delta_{i_p}}{8\sqrt{2}}\right)^2 \frac{1}{i_p}\frac{B}{16\log_{4/3} m}\right) \\
&\le \exp\left(-\frac{4b^4}{9H}\frac{B}{|\mathcal{X}^*|}\right) + \exp\left(-\frac{4b^4}{9H}\frac{B}{16\log_{4/3} m}\right).
\end{aligned}\tag{17}$$

Combine (15), (16), (17) together, then we complete the proof of Theorem 3.2.

Last, we analyse the algorithm's behavior. Consider the case that $\mathcal{Q}_p > 2\mathcal{P}_p$, according to (12) (which does not rely on $\bar{\mathcal{Q}}_p > \bar{\mathcal{P}}_p$), with probability at least $1 - 16md\cdot\frac{|\mathcal{P}_p\setminus S_p^*|}{|\mathcal{Q}_p|}\Phi_p^*$, we have

$$\begin{aligned}
|\bar{\mathcal{P}}_p| &= |\bar{\mathcal{P}}_p\cap\mathcal{Q}_p| + |\bar{\mathcal{P}}_p\cap\mathcal{P}_p| \\
&= |\mathcal{Q}_p| - |\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p| + |\mathcal{P}_p| - |\bar{\mathcal{Q}}_p\cap\mathcal{P}_p| \\
&< \frac{3}{2}|\mathcal{Q}_p| - |\bar{\mathcal{Q}}_p| \\
&\le |\bar{\mathcal{Q}}_p|.
\end{aligned}$$

Similarly, consider the case that $\mathcal{P}_p > 2\mathcal{Q}_p$, according to (14), with probability at least $1 - 16md\cdot\frac{|\mathcal{Q}_p\setminus S_p'|}{|\bar{\mathcal{P}}_p|}\Phi_p'$, we have

$$|\bar{\mathcal{Q}}_p| = |\bar{\mathcal{Q}}_p\cap\mathcal{Q}_p| + |\bar{\mathcal{Q}}_p\cap\mathcal{P}_p| < \frac{3}{2}|\mathcal{P}_p| - |\bar{\mathcal{P}}_p| \le |\bar{\mathcal{P}}_p|.$$

C. Analysis of Algorithm 2

Our proof based on the following concentration inequality.

Lemma C.1 ((Kleinberg et al., 2008; Babaioff et al., 2015)). *Consider some distribution with values in $[0,1]$, let v and $\bar{v}$ be the expectation and average of n independent samples $x_1, x_2, \ldots, x_n$ from this distribution, respectively. Then for each $\gamma > 0$,*

$$\mathbb{P}[|v-\bar{v}| \le f_{rad}(\bar{v},n) \le 3f_{rad}(v,n)] \ge 1-\exp(-\Omega(\gamma)),\tag{18}$$

where $f_{rad}(v,n) = \sqrt{\frac{\gamma v}{n}} + \frac{\gamma}{n}$. More generally, equation (18) holds if $v = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[x_t|x_1,\ldots,x_{t-1}]$.

We first prove the clean event that the upper confidence bound of the expected reward $\boldsymbol{\mu}^U(T/3)$ and lower confidence bound of the expected consumption $\boldsymbol{C}^L(T/3)$satisfy the following properties:

(1) with probability at least $1-2mdT\cdot\exp(-\Omega(\gamma))$,

$$|(\boldsymbol{\mu}^U(T/3))^\top \boldsymbol{x}_{T/3}-\text{OPT}_{\text{LP}}|\le O\left(\sqrt{\frac{\gamma m\cdot\text{OPT}_{\text{LP}}}{T}}+\frac{\gamma md}{T}+\frac{\text{OPT}_{\text{LP}}}{B}\sqrt{\frac{\gamma md\cdot b}{T}}\right).$$

(2) with probability at least $1-2mdT\cdot\exp(-\Omega(\gamma))$,

$$\sum_{t=1}^{T/3}\left|(\boldsymbol{C}^L(t))^\top \boldsymbol{x}_t-\boldsymbol{c}_t\right|\le\left(1-O\left(\sqrt{\frac{\gamma m}{B}}+\frac{\gamma m\log T}{B}\right)\right)\frac{B\mathbf{1}}{3}.$$

The proof is standard and similar to Lemma 7.4 in (Badanidiyuru et al., 2013) and the theoretical analysis in Appendix B.3 for the UCB algorithm for BwK (Agrawal & Devanur, 2014). We provide the details for completeness. Let $\hat{v}$ denote the empirical average of n samples, with probability at least $1-\exp(-\Omega(\gamma))$, we have

$$\begin{aligned}|\hat{v}-\bar{v}|&\le\frac{n}{n+1}\cdot f_{rad}(\hat{v},n)+\frac{\bar{v}}{n+1}\\&\le f_{rad}(\hat{v},n+1)+\frac{\bar{v}}{n+1}\\&\le 2f_{rad}(\hat{v},n+1).\end{aligned}$$

By take a union bound, with probability $1-mT\cdot\exp(-\Omega(\gamma))$, we have

$$\begin{aligned}\left|\frac{3}{T}\sum_{t=1}^{T/3}(r(t)-\mu_{i(t)})\right|&\le O\left(f_{rad}\left(\frac{3}{T}\sum_{t=1}^{T/3}(\boldsymbol{\mu}^U(t))_{i(t)},\frac{T}{3}\right)\right),\\ \left|(\boldsymbol{\mu}^U(T/3))^\top\boldsymbol{x}_{T/3}-\frac{3}{T}\sum_{t=1}^{T/3}(\boldsymbol{\mu}^U(t))_{i(t)}\right|&\le O\left(f_{rad}\left(\frac{3}{T}\sum_{t=1}^{T/3}(\boldsymbol{\mu}^U(t))_{i(t)},\frac{T}{3}\right)\right).\end{aligned}\tag{19}$$

According to (Badanidiyuru et al., 2013), for any two vectors $\boldsymbol{a},\boldsymbol{n}\in\mathbb{R}_+^m$, the following inequality always hold:

$$\sum_{i=1}^m f_{rad}(a_i,n_i)n_i\le\sqrt{\gamma m(\boldsymbol{a}\cdot\boldsymbol{n})}+\gamma m$$

Therefore,

$$\begin{aligned}\left|\sum_{t=1}^{T/3}(\mu_{i(t)}-(\boldsymbol{\mu}^U(t))_{i(t)})\right|&\le O\left(\sum_t f_{rad}(\mu_{i(t)},n_{i(t)}(t)+1)\right)\\&\le O\left(\sum_i(n_{i(t)}(T/3)+1)f_{rad}(\mu_{i(t)},n_{i(t)}(T/3)+1)\right)\\&\le O\left(\sqrt{\gamma m\left(\sum_i\mu_i(n_{i(t)}(T/3)+1)\right)}+\gamma m\right)\\&\le O\left(\sqrt{\gamma m\left(\sum_t\mu_{i_t}\right)}+\gamma m\right)\\&\le O\left(\sqrt{\gamma m\left(\sum_t(\boldsymbol{\mu}^U(t))_{i(t)}\right)}+\gamma m\right)\end{aligned}\tag{20}$$

Combine (19) and (20) together, we obtain

$$\sqrt{\sum_{t=1}^{T/3}(\boldsymbol{\mu}^U(t))_{i(t)}} \leq \sqrt{\sum_{t=1}^{T/3} r(t)} + O(\sqrt{\gamma m}),$$

and

$$\left|(\boldsymbol{\mu}^U(T/3))^\top \boldsymbol{x}_{T/3} - \frac{3}{T}\sum_{t=1}^{T/3} r(t)\right| \leq O\left(\sqrt{\gamma m(\sum_t r(t))} + \gamma m\right). \tag{21}$$

Similarly, we could also prove that with probability $1 - mdT \cdot \exp(-\Omega(\gamma))$, we have

$$\sum_{t=1}^{T/3}\left|(\boldsymbol{C}^L(t))^\top \boldsymbol{x}_t - \boldsymbol{c}_t\right| \leq O\left(\sqrt{\gamma m B} + \gamma m\right)\mathbf{1}. \tag{22}$$

Combine (21) and (22) with the following inequality

$$\frac{3}{T}\sum_{t=1}^{T/3} r(t) \geq (1-\epsilon)\text{OPT}_{\text{LP}}$$

and substituting the specification of ϵ and $\gamma = O(\log(mdT))$, we obtain the desired inequalities. Notice that the consumption during the second step is at most $\frac{B}{2}$, so the consumption of FULLOAK will less than $\frac{5B}{6}$ with high probability. Define

$$\Psi := \left(\sqrt{\frac{\gamma m \cdot \text{OPT}_{\text{LP}}}{T}} + \frac{\gamma m d}{T} + \frac{\text{OPT}_{\text{LP}}}{B}\sqrt{\frac{\gamma m d \cdot b}{T}}\right).$$

Then we prove the following lemma.

Lemma C.2. *At the end of phase p, for any optimal arm $i^* \in \mathcal{X}^*$ and any sub-optimal arm $i' \in \mathcal{X}'$, the probability that $\bar{R}_{i^*}(p) > \bar{R}_{i'}(p)$ is at most*

$$2md \cdot \exp\left(-\alpha_1 \cdot b^2 R_{i'}^2 s(p)\right).$$

for some constant α_1.

Proof. Consider the dual form

$$\begin{aligned} \min \quad & \boldsymbol{b}^\top \boldsymbol{w}, \\ \text{s.t.} \quad & \bar{\boldsymbol{C}}^\top \boldsymbol{w} \geq \bar{\boldsymbol{\mu}}, \boldsymbol{w} \geq \mathbf{0}. \end{aligned}$$

Let $\bar{\boldsymbol{w}}^*(p)$ denote the optimal solution of it, we have

$$|\boldsymbol{b}^\top \bar{\boldsymbol{w}}^*(p) - \boldsymbol{b}^\top \boldsymbol{w}| \leq O(\Psi).$$

Then we get

$$\begin{aligned} \sum_{j=1}^d (\bar{w}^*(p))_j - \sum_{j=1}^d w_j^* &\geq -\frac{1}{b}O(\Psi), \\ \sum_{j=1}^d (\bar{w}^*(p))_j - \sum_{j=1}^d w_j^* &\leq \frac{1}{b}O(\Psi). \end{aligned} \tag{23}$$

Consider R_i of optimal arm i^* and suboptimal i'

$$\begin{aligned} R_{i^*} &= (\boldsymbol{w}^*)^\top \boldsymbol{C}_{i^*} - \mu_{i^*} && \leq 0, \\ \bar{R}_{i^*} &= (\bar{\boldsymbol{w}}^*(p))^\top \bar{\boldsymbol{C}}_{i^*} - \bar{\mu}_{i^*} && > 0, \\ \bar{R}_{i'} &= (\bar{\boldsymbol{w}}^*(p))^\top \bar{\boldsymbol{C}}_{i'} - \bar{\mu}_{i'} && < \bar{R}_{i^*}. \end{aligned} \tag{24}$$

From (23) and (24), we have

$$
\begin{aligned}
&\sum_{j=1}^{d}(\bar{w}^*(p))_j\bar{C}_{j,i^*} - \bar{\mu}_{i^*} \\
\le&\sum_{j=1}^{d}(\bar{w}^*(p))_j\left(C_{j,i^*} + 3f_{rad}(C_{j,i^*}, s^*(p))\right) - \left(\mu_{i^*} - 3f_{rad}(\mu_{i^*}, s^*(p))\right) \\
=&\sum_{j=1}^{d} w_j^*\left(C_{j,i^*} + 3f_{rad}(C_{j,i^*}, s^*(p))\right) - \Bigg[\sum_{j=1}^{d} w_j^*\left(C_{j,i^*} + 3f_{rad}(C_{j,i^*}, s^*(p))\right) \\
&\quad - \sum_{j=1}^{d}(\bar{w}^*(p))_j\left(C_{j,i^*} + 3f_{rad}(C_{j,i^*}, s^*(p))\right)\Bigg] - \left(\mu_{i^*} - 3f_{rad}(\mu_{i^*}, s^*(p))\right) \\
\le&\sum_{j=1}^{d} w_j^*\left(C_{j,i^*} + 3f_{rad}(C_{j,i^*}, s^*(p))\right) + \frac{1}{b}O(\Psi) - \left(\mu_{i^*} - 3f_{rad}(\mu_{i^*}, s^*(p))\right) \\
\le&3f_{rad}(C_{j,i^*}, s^*(p))\left(1 + \sum_{j=1}^{d} w_j^*\right) + \frac{1}{b}O(\Psi).
\end{aligned}
$$

And

$$
\begin{aligned}
&\sum_{j=1}^{d}(\bar{w}^*(p))_j\bar{C}_{j,i'} - \bar{\mu}_{i'} \\
\ge&\sum_{j=1}^{d}(\bar{w}^*(p))_j(C_{j,i'} - 3f_{rad}(C_{j,i'}, s'(p))) - (\mu_{i'} + 3f_{rad}(\mu_{i'}, s'(p))) \\
=&\sum_{j=1}^{d} w_j^*(C_{j,i'} - 3f_{rad}(C_{j,i'}, s'(p))) - \Bigg[\sum_{j=1}^{d} w_j^*\left(C_{j,i^*} - 3f_{rad}(C_{j,i'}, s'(p))\right) \\
&\quad - \sum_{j=1}^{d}(\bar{w}^*(p))_j\left(C_{j,i^*} - 3f_{rad}(C_{j,i'}, s'(p))\right)\Bigg] - (\mu_{i'} + 3f_{rad}(\mu_{i'}, s'(p))) \\
\ge&\sum_{j=1}^{d} w_j^*(C_{j,i'} - 3f_{rad}(C_{j,i'}, s'(p))) - \frac{1}{b}O(\Psi) - (\mu_{i'} + 3f_{rad}(\mu_{i'}, s'(p))) \\
\ge&R_{i'} - 3f_{rad}(C_{j,i'}, s'(p))\left(1 + \sum_{j=1}^{d} w_j^*\right) - \frac{1}{b}O(\Psi).
\end{aligned}
$$

According to the Strong duality theorem, we have

$$
\sum_{j=1}^{d} w_j^* = \frac{1}{b}\sum_{i=1}^{m}\mu_i x_i^* = \frac{1}{b}\text{OPT}_{\text{LP}} \le \frac{1}{b}.
$$

Then we have

$$
\begin{aligned}
&\mathbb{P}[\bar{R}_{i^*}(p) > \bar{R}_{i'}(p)] \\
&\le \mathbb{P}\left[R_{i'} < 3\left(f_{rad}(C_{j,i^*}, s^*(p)) + 3f_{rad}(C_{j,i'}, s'(p))\right)\left(1 + \frac{1}{b}\right) + \frac{1}{b}O(\Psi)\right] \\
&\le \mathbb{P}\left[R_{i'} < \frac{1}{b}O\left(\Psi + f_{rad}(1, s(p))\right)\right].
\end{aligned}
$$

Notice that we have $f_{rad}(1, s(p)) = \Omega(\Psi)$. Combine them together, then we complete the proof. □

According to Lemma B.4, for any optimal arm $i^* \in \mathcal{X}^*$ and any active sub-optimal arm $i' \in \mathcal{X}' \cap \mathcal{X}_p$, the probability that $\bar{G}_{i^*}(p) < \bar{G}_{i'}(p)$ is at most

$$2md \cdot \exp\left(-\frac{2b^2 G_{i^*}^2}{9} s(p)\right)$$

during phase p. Similar to the proof of Theorem 3.2, we bound the probability that the algorithm makes mistakes by ignoring the $\frac{1}{16}|\mathcal{X}_p|$ arms with the smallest G_i and the $\frac{1}{16}|\mathcal{X}_p|$ arms with largest R_i of the active arms set. For all arms in $\mathcal{X}_p$, let $\mathcal{P}_p$ and $\bar{\mathcal{P}}_p$ denote the set of optimal arms for (1) and (4), $\mathcal{Q}_p$ and $\bar{\mathcal{Q}}_p$ denote the set of sub-optimal arms for (1) and (4). Let S_p^* denote the $\frac{1}{16}|\mathcal{X}_p|$ arms with smallest G_i and S_p' denote the $\frac{1}{16}|\mathcal{X}_p|$ arms with largest R_i, respectively. Define

$$\Phi_p^* := \max_{i \in \mathcal{P}_p \setminus S_p^*} \exp\left(-\frac{2b^2 G_i^2}{9} s(p)\right),$$
$$\Phi_p' := \max_{i \in \mathcal{Q}_p \setminus S_p'} \exp\left(-\alpha_1 \cdot b^2 R_{i'}^2 s(p)\right).$$

We obtain that the probability that there is at least one eliminated optimal arms is at most

$$32md \cdot \Phi_p^* + 12md \cdot \Phi_p'.$$

Similarly, the probability that at least one sub-optimal arm is added to $\mathcal{X}_p^*$ is at most

$$32md \cdot \Phi_p' + 12md \cdot \Phi_p^*.$$

Notice that FULLOAK will delete all accept arms from the surviving arms set at the end of each phase, the pulls for each arm in $\mathcal{X}_p$ before phase $t+1$ satisfy

$$s(p) \geq \frac{B}{\log_{4/3} m} \sum_{k=0}^{p} \frac{1}{|\mathcal{X}_p|}$$

Let $i_p = \frac{m}{16}\left(\frac{3}{4}\right)^p$, then we have

$$\begin{aligned}
\Phi_p^* &= \max_{i \in \mathcal{P}_p \setminus S_p^*} \exp\left(-\frac{2b^2 G_i^2}{9} s(p)\right) \\
&\leq \exp\left(-\frac{2b^2 \Delta_{i_p}^2}{9} s(p)\right) \\
&\leq \exp\left(-\frac{4b^2 \Delta_{i_p}^2}{9} \frac{B}{m \log_{4/3} m} \left(\frac{4}{3}\right)^p\right) \\
&= \exp\left(-\frac{4b^2 \Delta_{i_p}^2}{9 i_p} \frac{B}{16 \log_{4/3} m}\right) \\
&\leq \exp\left(-\frac{4b^2}{9H} \frac{B}{16 \log_{4/3} m}\right)
\end{aligned}$$

Similarly, we have

$$\Phi_p' = \max_{i \in \mathcal{Q}_p \setminus S_p'} \exp\left(-\alpha_1 \cdot b^2 R_{i'}^2 s(p)\right) \leq \exp\left(-\frac{\alpha_2 b^2 B}{H \log m}\right),$$

where α_2 is a constant. Combine them together, then we complete the proof.

The proof is similar to Lemma 7.4 in (Badanidiyuru et al., 2013) and the theoretical analysis in Appendix B.3 for the UCB algorithm for BwK (Agrawal & Devanur, 2014).

D. Analysis of Lower Bound (Theorem 4.2)

Let $(p_w)_{2 \leq w \leq W} \in [1/4, 1/2)$ be $(W-1)$ real numbers and let $p_1 = 1/2$. And we define the quantities $l_w := 1/2 - p_w$. Assume m is an exact multiple of W. Then we define

$$\mu_i := \frac{1}{2} - \frac{l_w}{2^{\lfloor (m-i)/W \rfloor}}, w = (i \bmod W), i \in [m].$$

Let π_i denote the Bernoulli distribution of mean μ_i and π_i' denote the Bernoulli distribution of mean $1-\mu_i$.

Consider W problem instances with time horizon T, m arms, d types of resources being consumed, and knapsack $b = W/m$ for each type of resource. To ease the reading, assume T is a power of 2, $W \geq \Omega(\sqrt{m})$, and $d > m/W$. Let $w = (i \bmod W)$, for the u-th problem instance, the i-th arm x_i^u is associated with the reward distribution π_i^u,

$$\pi_i^u := \pi_i \mathbf{1}\{w \neq u\} + \pi_i' \mathbf{1}\{w = u\}, u \in [W], i \in [m].$$

The consumption vector $\boldsymbol{c}_i^u$ satisfies $(\boldsymbol{c}_i^u)_1 = (\boldsymbol{c}_i^u)_d = (\boldsymbol{c}_i^u)_w = 1$, and $(\boldsymbol{c}_i^u)_j = 0$ for all $j \neq 1, j \neq w, j \neq d$. Then there are $|\mathcal{X}^*| = m/W = 1/b$ optimal arms and their indexes satisfy $(i \bmod W) = u$. For the hardness measure of the u-th problem instance $H(u)$, we have

$$H(u) = \max_{i \in [m]} \frac{i}{\Delta_{i,u}^2} \leq b \cdot 2^{\frac{1}{b}+1} \sum_{w \neq u} (l_w + l_u)^{-2},$$

where $\Delta_{i,u}$ is the vertex gap of the i-th arm for u-th instance.

Consider any algorithm $\mathcal{A}$ and let $(T_k)_{1 \leq k \leq |\mathcal{X}^*|}$ denote the number of samples by $\mathcal{A}$ on arms from index $(k-1) \cdot W + 1$ to $k \cdot W$. These quantities are random but satisfy $\sum_{1 \leq k \leq |\mathcal{X}^*|} T_k = B$. We have

$$\begin{aligned}
\mathbb{P}(\mathcal{O} \neq \mathcal{X}^*) &\geq \sum_{i \in \mathcal{X}^*} \mathbb{P}(i \notin \mathcal{O}) \\
&\geq \sum_{1 \leq k \leq |\mathcal{X}^*|} \exp\left(-\frac{\beta_1 T_k}{2^{m-k} \cdot \log(W) \sum_{w \neq u} (l_w + l_u)^{-2}}\right) \\
&\geq \exp\left(-\frac{\beta_2 b B}{2^{\frac{1}{b}-1} \log(W) \sum_{w \neq u} (l_w + l_u)^{-2}}\right) \\
&\geq \exp\left(-\frac{\beta_2 b^2 \cdot 2^{\frac{1}{b}+1} B}{2^{\frac{1}{b}-1} H(u) \log(W)}\right) \geq \exp\left(-\frac{2\beta_2 b^2 B}{H(u) \log m}\right),
\end{aligned}$$

where β_1, β_2 are some constants. The second inequality comes from Theorem 2 of (Carpentier & Locatelli, 2016). Then we complete the proof.

E. Analysis of Special Cases

E.1. Simple OAK Problem

In this section, we provide the specification of BASEOAK$^-$ and prove Theorem 5.1.

The algorithm (shown in Algorithm 3) also splits the budget evenly into phases and chooses the worst/best quarter of surviving arms to reject/accept at the end of each phase. The difference is that the Algorithm 3 eliminates the accepted arms from the active arm set at the end of each accept phase.

We provide the proof of Theorem 5.1 below. Again, the probability that the algorithm makes mistakes in phase p is at most

$$32|\bar{\mathcal{P}}_p \cup \mathcal{P}_p|^2 \cdot (\Phi_p' + \Phi_p^*). \tag{25}$$

And the algorithm does not exceed the budget T.

Consider the pulls for each arm in $\mathcal{X}_p$ before phase $p+1$

$$s(p) \geq \frac{T}{\log_{4/3} m} \sum_{k=0}^{p} \frac{1}{|\mathcal{X}_p|}.$$

Algorithm 3 BASEOAK$^-$

```
Input: rounds $T$, number of arms $m$
$\mathcal{X}_0 \leftarrow [m]$, $\mathcal{X}'_0 \leftarrow \emptyset$, $\mathcal{X}^*_0 \leftarrow \emptyset$
for $p = 0, \dots, \lceil \log_{4/3} m \rceil - 1$ do
    Pull each arm $i \in \mathcal{X}_p$ for
        $n(p) = \left\lfloor \frac{T}{|\mathcal{X}_p| \lceil \log_{4/3} m \rceil} \right\rfloor$
    times
    Compute the empirical estimator of the reduced gap $\bar{R}_i$ and deletion gap $\bar{G}_i$ for each arm $i \in \mathcal{X}_p$
    if more basic variables in $\bar{\boldsymbol{x}}^*(p)$ then
        $\mathcal{X}^*_{p+1} \leftarrow \mathcal{X}^*_p \cup \{$ the set of $\lceil |\mathcal{X}_p|/4 \rceil$ optimal arms in $\mathcal{X}_p$ with the largest $\bar{G}_i\}$
    else
        $\mathcal{X}'_{p+1} \leftarrow \mathcal{X}'_p \cup \{$ the set of $\lceil |\mathcal{X}_p|/4 \rceil$ sub-optimal arms in $\mathcal{X}_p$ with the largest $\bar{R}_i\}$
    end if
    $\mathcal{X}_{p+1} \leftarrow \mathcal{X}_0 \backslash (\mathcal{X}'_{p+1} \cup \mathcal{X}^*_{p+1})$
end for
Output $\mathcal{X}^*_{\lceil \log_{4/3} m \rceil}$
```

Let $i_p = \frac{m}{16}\left(\frac{3}{4}\right)^p$, then we have

$$\begin{aligned}
\Phi^*_p &= \max_{i \in \mathcal{P}_p \backslash S^*_p} \exp\left(-\frac{2b^2 G_i^2}{9} s(p)\right) \\
&\leq \exp\left(-\frac{2b^2 \Delta_{i_p}^2}{9} s(p)\right) \\
&\leq \exp\left(-\frac{4b^2 \Delta_{i_p}^2}{9} \frac{T}{m \log_{4/3} m} \left(\frac{4}{3}\right)^p\right) \\
&= \exp\left(-\frac{4b^2 \Delta_{i_p}^2}{9 i_p} \frac{T}{16 \log_{4/3} m}\right) \\
&\leq \exp\left(-\frac{4b^2}{9H} \frac{T}{16 \log_{4/3} m}\right),
\end{aligned} \tag{26}$$

and

$$\begin{aligned}
\Phi'_p &= \max_{i \in \mathcal{Q}_p \backslash S'_p} \exp\left(-2\left(\frac{b\Delta_2}{8} + \frac{b^2 R_i}{8}\right)^2 s(p)\right) \\
&\leq \exp\left(-2\left(\frac{b\Delta_2}{8} + \frac{b^2 \Delta_{i_p}}{8\sqrt{2}}\right)^2 s(p)\right) \\
&\leq \exp\left(-4\left(\frac{b\Delta_2}{8} + \frac{b^2 \Delta_{i_p}}{8\sqrt{2}}\right)^2 \frac{1}{i_p} \frac{T}{16 \log_{4/3} m}\right) \\
&\leq \exp\left(-\frac{4b^4}{9H} \frac{T}{16 \log_{4/3} m}\right).
\end{aligned} \tag{27}$$

For all rounds t, we have $|\bar{\mathcal{P}}_p \cup \mathcal{P}_p| \leq \left(\frac{3}{4}\right)^p m$. Combine (25), (26), (27) together, then we complete the proof.

E.2. Pure Exploration Problems

In this section, we prove the results in Example 5.2 and 5.3.

For the BAI problem, there is one optimal arm i^*. According to the proof of Lemma B.3, assume the optimal arm i^* is not eliminated at the end of phase p. For the optimal arm i^* and any active sub-optimal arm $i' \in \mathcal{X}' \cap \mathcal{X}_p$, the probability that $\bar{R}_{i^*}(p) > \bar{R}_{i'}(p)$ is at most

$$O\left(\exp\left(-2\left(\frac{b\Delta_2}{8} + \frac{b^2 R_{i'}}{8}\right)^2 s(p)\right)\right).$$

According to the proof of Theorem 3.2, the probability that the algorithm makes mistakes in round p is at most

$$32(\Phi'_p + \Phi^*_p).$$

By equation (26), (27), and a union bound, we complete the proof of the result in Example 5.2.

For the TopK and MB problem, due to the deterministic resource consumption, the probability that the algorithm makes mistakes in round p is at most

$$32|\bar{\mathcal{P}}_p \cup \mathcal{P}_p| \cdot (\Phi'_p + \Phi^*_p).$$

For all rounds p, we have $|\bar{\mathcal{P}}_p \cup \mathcal{P}_p| \leq \left(\frac{3}{4}\right)^p m$. By equation (26), (27), and a union bound, we obtain the result in Example 5.3.