Optimal Multi-Objective Best Arm Identification with Fixed Confidence

Zhirui Chen[1] **P. N. Karthik**[2] **Yeow Meng Chee**[1] **Vincent Y. F. Tan**[1]

[1]National University of Singapore [2]Indian Institute of Technology, Hyderabad

zhiruichen@u.nus.edu pnkarthik@ai.iith.ac.in {ymchee,vtan}@nus.edu.sg

Abstract

We consider a multi-armed bandit setting with finitely many arms, in which each arm yields an M-dimensional vector reward upon selection. We assume that the reward of each dimension (a.k.a. *objective*) is generated independently of the others. The best arm of any given objective is the arm with the largest component of mean corresponding to the objective. The end goal is to identify the best arm of *every* objective in the shortest (expected) time subject to an upper bound on the probability of error (i.e., fixed-confidence regime). We establish a problem-dependent lower bound on the limiting growth rate of the expected stopping time, in the limit of vanishing error probabilities. This lower bound, we show, is characterised by a max-min optimisation problem that is computationally expensive to solve at each time step. We propose an algorithm that uses the novel idea of *surrogate proportions* to sample the arms at each time step, eliminating the need to solve the max-min optimisation problem at each step. We demonstrate theoretically that our algorithm is asymptotically optimal. In addition, we provide extensive empirical studies to substantiate the efficiency of our algorithm. While existing works on pure exploration with multi-objective multi-armed bandits predominantly focus on *Pareto frontier identification*, our work fills the gap in the literature by conducting a formal investigation of the multi-objective best arm identification problem. Code: https://github.com/zchen42/simulation_mobai

1 Introduction

Multi-armed bandit (MAB) (Thompson, 1933) is a sequential decision-making paradigm where an agent sequentially pulls one out of K finitely many arms and receives a corresponding reward at each time step, with widespread applications in clinical trials, internet advertising, and recommender systems (Lattimore and Szepesvári, 2020). In the classical MAB setup, the rewards from the arms are independent and identically distributed (i.i.d.), and real-valued (one-dimensional). In contrast, the *multi-objective multi-armed bandit* (MO-MAB) setup proposed by Drugan and Nowe (2013) allows for i.i.d. multi-dimensional (vector) rewards from the arms, with the reward of any given dimension (a.k.a the *objective*) being independent of the others or, more generally, a function of the rewards of the others. Defining the best arm of an objective as the arm with the largest mean reward corresponding to the objective, it is evident that in the MO-MAB setup, distinct objectives may possess distinct best arms, leading to the possibility of an arm being optimal for one objective and sub-optimal for another, thereby amplifying the complexity of identifying one or more best arms. In this paper, we study the problem of recovering the best arm of *every* objective in the shortest (expected) time, while ensuring that the probability of error is within a prescribed threshold (*fixed-confidence* regime).

Motivation Consider the task of deploying advertisements from a candidate set of advertisements, on platforms such as YouTube and Twitch. Here, selecting an advertisement to launch on any given day is analogous to pulling an arm. The feedback obtained from deploying a specific advertisement is inherently multi-dimensional, comprising various video-specific metrics such as user engagement and view rates, as well as demography-specific metrics such as viewer age and gender. The task of identifying the optimal advertisement for different demographic segments, which is crucial for maximizing revenue, translates to finding the best arm for each objective (e.g., each age group).

As such, the problem of best arm identification (BAI)

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

poses a non-trivial challenge, primarily owing to the inherent uncertainty associated with the true reward distribution of each arm. This challenge is further exacerbated when rewards are multi-dimensional (as in the MO-MAB setting). As delineated in prior works, numerous practical applications exhibit rewards that are multi-dimensional in nature, as opposed to being solely scalar, such as hardware design (Zuluaga et al., 2016), drug development and dose identification (Lizotte and Laber, 2016) in clinical trials, and electric battery control (Busa-Fekete et al., 2017). However, the existing works on pure exploration in MO-MAB settings are mainly focused on Pareto frontier identification. The identification of the best arm for each objective, an inherent task in MO-MAB scenarios, has not received comprehensive scholarly attention. This study seeks to fill the research gap in this domain.

1.1 Overview of Existing Works

Multi-objective bandits and BAI have both been extensively investigated in the literature. In particular, the latter has been investigated under two complementary regimes: the fixed-confidence regime (Even-Dar et al., 2006; Garivier and Kaufmann, 2016) and the fixed-budget regime (Audibert and Bubeck, 2010; Chen et al., 2024). This section highlights a collection of recent studies of BAI on the fixed-confidence regime addressing these topics. Garivier and Kaufmann (2016) first proposed the well-known Track-and-Stop (TaS) algorithm with two variants (C-Tracking and D-tracking), and demonstrated the optimality of these variants in the asymptotic limit of vanishing error probabilities. The basic premise for achieving asymptotic optimality, they showed, is to pull arms according to an *oracle weight* that is derived from the problem instance-dependent lower bound. Later, Degenne et al. (2020) and Jedra and Proutiere (2020) specialised the TaS algorithm to the linear bandit setting, while still maintaining asymptotic optimality. For more general structured bandits with non-linear structural dependence between the mean rewards of the arms, Wang et al. (2021) proposed an efficient TaS-type algorithm and a novel lower bound for the structured MAB setting, and further established the asymptotic optimality of their algorithm. Mukherjee and Tajer (2023) also proposed an efficient scheme for achieving asymptotic optimality without solving for the oracle weight at each time step. It is noteworthy that the aforementioned studies deal with a *single objective*, whereas our research deals more generally with *multiple objectives*.

Degenne and Koolen (2019) explored the problem of fixed-confidence BAI with *multiple correct answers* (i.e., multiple best arms), with the objective of identifying any one of the correct answers. They proposed the Sticky Track-and-Stop algorithm along the lines of C-Tracking and demonstrated its asymptotic optimality. While their setup appears to bear similarities with ours, it is worth noting that their work aims to identify *one* among several correct answers, whereas our study focuses on uncovering *all* correct answers (i.e., the best arm of *every* objective).

Drugan and Nowe (2013) introduced the MO-MAB setting as well as two associated metrics—the *Pareto regret* and the *scalarized regret*. The authors proposed two UCB-like algorithms to optimize these two metrics. Subsequently, several factions of researchers have predominantly concentrated on the study of Pareto regret. Turgay et al. (2018) tackled the problem of Pareto regret minimization by introducing a similarity assumption regarding the means of arms, and designed an algorithm that achieves a regret upper bound of the order $\tilde{O}\left(T^{(1+d_p)/(2+d_p)}\right)$, where d_p is the number of dimensions that is a function of the arm vectors and the environmental context. Xu and Klabjan (2023) defined the notion of Pareto regret in the context of adversarial bandits (Lattimore and Szepesvári, 2020, Chapter 11), and designed an algorithm achieving near-optimality up to a factor of $\log T$ in both adversarial and stochastic bandit environments.

On the topic of pure exploration in the MO-MAB setting, a popular line of work is *Pareto optimal arm identification* or *Pareto frontier identification*. Considering specifically the fixed-confidence regime, Auer et al. (2016) proposed a successive elimination (SE)-type algorithm to identify all the Pareto optimal arms. For any given confidence level, the upper bound on the sample complexity of their algorithm matches their lower bound up to a logarithmic factor that is a function of the sub-optimality gaps of the arms. Along similar lines, Ararat and Tekin (2023) present another SE-type algorithm to identify all Pareto (ϵ, δ)-PAC arms, a generalization of Pareto optimal arms. More recently, Kim et al. (2023) developed a framework for analysing the problem of Pareto frontier identification in linear bandits. Their proposed algorithm is nearly optimal up to a logarithmic factor involving the minimum of the arm sub-optimality gaps and the algorithm's accuracy parameter. While the best arm of each objective is notably also Pareto optimal, our work goes beyond merely identifying a subset of Pareto optimal arms, hence significantly advancing the state-of-the-art in multi-objective bandits. Existing works on Pareto optimal arm identification in the MO-MAB setting lack the capability to identify the best arm of each objective, a task that our work accomplishes. We discuss in Appendix B further details of the significant differences between our setting and Pareto optimal arm identification.

In the realm of pure exploration with vector-valued payoffs, the work of Prabhu et al. (2022) studies the problem of multi-hypothesis testing with vector arm rewards under the fixed-confidence regime. The authors provide insightful contributions by establishing both lower and upper bounds on the expected stopping time. Notably, their generic problem formulation encapsulates both single-objective best arm identification (BAI) and its multi-objective counterpart. However, when specialized to multi-objective BAI, a computational bottleneck arises in their proposed algorithm due to the emergence of a multi-dimensional optimization routine akin to the core optimization procedure in TaS. This introduces computational inefficiencies, warranting computationally more efficient solutions for high-dimensional problems. In a parallel vein, the work of Shang et al. (2020) introduces the concept of the "relative vector loss," tied to the sup-norm of vector rewards. Their primary objective lies in identifying the best arm (defined in terms of vector losses and hence vastly different from ours) within the fixed-confidence regime. While their focus is on identifying the overall best arm, our framework necessitates identifying the best arm of each objective.

The MO-MAB setup (with M independent objectives) appears to bear connections to the federated multi-armed bandit setting with a single server and M independent clients, with each client having access to all arms or a subset thereof and striving to determine its best arm, as studied for instance in the recent works of Reddy et al. (2023); Shi et al. (2021); Chen et al. (2023). Despite the apparent high-level similarities between the two settings, there exists one crucial distinction. In the federated setting, each client may choose an arm of its choice independently of the other clients, thereby leading to the possibility of accruing M rewards from M distinct arms at any given time instant. However, in the MO-MAB setup, an M-dimensional reward is generated from a single arm at each time instant. In essence, the MO-MAB setup is equivalent to a federated learning setup in which every client has access to all the arms, and all clients are compelled to pull the same arm at each time instant.

1.2 Our Contributions

While existing works on pure exploration in multi-objective bandits primarily focus on Pareto frontier identification, we bridge the gap by investigating the problem of identifying the best arm of each objective under the fixed-confidence regime.

We provide an asymptotic lower bound for the problem and complement it with an algorithm that achieves the lower bound up to a multiplicative constant of $1 + \eta$, where $\eta > 0$ is a tuneable parameter that can be chosen arbitrarily close to 0. We show that the lower bound is characterised by the solution to a max-min optimisation problem that is reminiscent of fixed-confidence BAI problems. The basic premise upon which asymptotically optimal algorithms of the prior works such as D-Tracking (Garivier and Kaufmann, 2016) and Sticky TaS (Degenne and Koolen, 2019) operate is to compute the max-min optimisation at each time step to evaluate the oracle weight at each time instant, and to pull arms according to the (empirical) oracle weight in order to guarantee asymptotic optimality. However, these approaches may be inefficient, as the oracle weight, to the best of our knowledge, does not have a closed-form solution in multi-objective cases. Instead of following the (empirical) oracle weight, we propose a novel technique to sample the arms at each step, based on the idea of *surrogate proportions*. These surrogate proportions serve as proxy for the oracle weight and can be computed efficiently.

2 Preliminaries and Problem Setup

Let $\mathbb{N}$ denote the set of positive integers. For $n \in \mathbb{N}$, let $[n] := \{1, \ldots, n\}$. We consider a multi-armed bandit with K arms in which each arm is associated with M independent *objectives*. Pulling arm $A_t \in [K]$ at time step t yields an M-dimensional reward $\mathbf{r}_t = [r_{t,m} : m \in [M]]^\top \in \mathbb{R}^M$, where $r_{t,m} = \mu_{A_t,m} + \eta_{t,m}$; here, $\mu_{A_t,m} \in \mathbb{R}$ is the unknown mean corresponding to objective m of arm A_t, and $\eta_{t,m}$ is an independent standard normal random variable. Let $v = [\mu_{i,m} : (i,m) \in [K] \times [M]]^\top$ denote a *problem instance* in which $\mu_{i,m}$ is the mean corresponding to objective m of arm i. Arm i is said to be the best arm of objective m if it has the highest mean in dimension m across all arms, i.e., $\mu_{i,m} > \mu_{j,m}$ for all $j \neq i$. Without loss of generality, we assume that each objective has a unique best arm, and we write $\mathcal{P}$ to denote the set of all problem instances with a unique best arm for each objective. We write $I^*(v) = (i_1^*(v), i_2^*(v), \ldots, i_M^*(v))$ to denote the collection of best arms under instance v; here, $i_m^*(v)$ is the best arm of objective m.

Given an error probability threshold $\delta \in (0, 1)$, the goal is to identify the set of best arms $I^*(v)$ in the shortest time, while ensuring that the error probability is within δ. Formally, an *algorithm* (or *policy*) for identifying the best arms is a tuple $\pi = (A, \tau, \widehat{I})$ consisting of the following components.

- An *arms selection rule* $A = \{A_t\}_{t=1}^\infty$ for pulling the arms at each time instant. Here, $A_t = A_t(A_{1:t-1}, \mathbf{r}_{1:t-1}, \delta)$ is a (random) function that takes input as the history of all the arms pulled and rewards obtained up to time step $t-1$ as well as the confidence δ, and outputs the arm to be pulled at time step t.

- A *stopping rule* that dictates the stopping time τ at which to stop further selection of arms.
- A *recommendation* $\widehat{I} \in [K]^M$ of best arm estimates at the stopping time τ.

For simplicity, let τ_δ and $\widehat{I}_\delta$ denote the stopping time and recommendation under confidence δ, respectively. Our interest is in the class of all *δ-PAC policies*, defined as

$$\Pi(\delta) \coloneqq \{\pi : \ \mathbb{P}_v^\pi(\tau_\delta < +\infty) = 1, \\ \mathbb{P}_v^\pi(\widehat{I}_\delta \neq I^\star(v)) \leq \delta \quad \forall v \in \mathcal{P}\}. \quad (1)$$

Here, and throughout the paper, we write $\mathbb{P}_v^\pi$ and $\mathbb{E}_v^\pi$ to denote probabilities and expectations under the instance v and under the policy π. Prior works on fixed-confidence BAI show that $\inf_{\pi\in\Pi(\delta)} \mathbb{E}_v^\pi[\tau_\delta] \approx \Theta\left(\log \frac{1}{\delta}\right)$. We anticipate that a similar growth rate for the expected stopping time holds in the context of our work. Our interest is to precisely characterise the asymptotic rate

$$\liminf_{\delta\downarrow 0} \ \inf_{\pi\in\Pi(\delta)} \frac{\mathbb{E}_v^\pi[\tau_\delta]}{\log(1/\delta)}, \quad (2)$$

where the asymptotics is as the error probability $\delta \downarrow 0$.

3 Lower bound

In this section, we present an lower bound on (2) for any instance $v \in \mathcal{P}$. We first introduce the notion of sub-optimality gaps under this instance. For any arm $i \in [K]$ and objective $m \in [M]$, we define the sub-optimality gap of the tuple (i, m) under the instance v as

$$\Delta_{i,m}(v) \coloneqq \mu_{i_m^*(v),m}(v) - \mu_{i,m}(v), \quad (3)$$

where, to recall, $i_m^*(v)$ is the best arm of objective m under instance v.

Proposition 3.1. *Fix $\delta \in (0,1)$. For any δ-PAC policy π,*

$$\mathbb{E}_v^\pi[\tau_\delta] \geq c^*(v) \log\left(\frac{1}{4\delta}\right), \quad \forall v \in \mathcal{P} \quad (4)$$

where the constant $c^*(v)$ is given by

$$c^*(v)^{-1} \coloneqq \sup_{\omega\in\Gamma} \ \min_{m\in[M]} \ \min_{i\in[K]\setminus i_m^*(v)} \frac{\omega_i\, \omega_{i_m^*(v)}\, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}. \quad (5)$$

In (5), Γ denotes the set of probability distributions on $[K]$, and we use the convention $\frac{\omega_i\,\omega_{i_m^*(v)}}{\omega_i+\omega_{i_m^*(v)}} = 0$ if $\omega_i = \omega_{i_m^*(v)} = 0$. Consequently, taking limits as $\delta \downarrow 0$ in (4), we get

$$\liminf_{\delta\downarrow 0} \ \inf_{\pi\in\Pi(\delta)} \frac{\mathbb{E}_v^\pi[\tau_\delta]}{\log(1/\delta)} \geq c^*(v). \quad (6)$$

Proposition 3.1 shows that the expected stopping time of any δ-PAC policy π grows as $\Omega(\log(1/\delta))$, as is the case in the prior works, and that the smallest (best) constant multiplying $\log(1/\delta)$ is $c^*(v)$ under the instance v. The constant $c^*(v)$ quantifies the complexity of identifying the best arms; notice that the smaller the sub-optimality gaps of the arms, the larger the value of $c^*(v)$, and therefore the larger the time required to find the best arms under any δ-PAC policy. When $M = 1$, the expression in (5) specialises to that of the problem complexity of BAI for Gaussian arms with unit variance; see, for instance, Garivier and Kaufmann (2016). The proof of Proposition 3.1 uses change-of-measure techniques introduced by Garivier and Kaufmann (2016) and is presented in Appendix E.

Notice that $c^*(v)^{-1}$ is the value of a sup-min optimisation problem that is typically reminiscent of fixed-confidence BAI problems, as evident from the lower bounds in the prior works. Defining

$$g_v(\omega) \coloneqq \min_{m\in[M]} \ \min_{i\in[K]\setminus i_m^*(v)} \frac{\Delta_{i,m}^2(v)}{2} \cdot \frac{\omega_i\, \omega_{i_m^*(v)}}{\omega_i + \omega_{i_m^*(v)}}, \quad (7)$$

let $\omega^*(v) \in \arg\max_{\omega\in\Gamma} g_v(\omega)$ denote the optimal solution to (5). For the case $M = 1$, the optimal solution $\omega^*(v)$ is referred to as the *oracle weight* of arm pulls (cf. Garivier and Kaufmann (2016)), and represents the optimal proportion of times each arm must be pulled in the long run to achieve the lower bound in (4). Building on this insight, the well-known Track-and-Stop (TaS) algorithm of Garivier and Kaufmann (2016) and its variants such as Lazy TaS (Jedra and Proutiere, 2020) and Sticky TaS (Degenne and Koolen, 2019), attempt to solve the sup-min optimisation of the lower bound therein to obtain the oracle weight for the estimated problem instance (in place of the unknown instance v) at each time step, a task that is computationally demanding. The computational burden of solving the sup-min optimisation at each time step is further escalated in the multi-objective setting of our work (i.e., when $M > 1$).

In this paper, we refrain from explicitly solving the sup-min optimisation for estimating the oracle weight at each time step. Instead, we construct a *surrogate proportion* as a proxy for the oracle weight, and sample arms according to the surrogate proportion at each time step. We show that the surrogate proportion may be computed easily by solving a linear program (using a standard technique such as the simplex method).

Remark 3.2. In a recent publication, Mukherjee and Tajer (2023) introduced the concept of a *look-ahead distribution* as an approximation for the oracle weight. They devised an arm sampling strategy named Transport Cost Based Arms Elimination (or TCB in short), which samples arms at each time step according to the look-ahead distribution. While their algorithm demonstrates promising computational efficacy in the context of one-dimensional arm rewards ($M = 1$), the applicability of their approach to scenarios with $M > 1$ remains uncertain. In hindsight, we note that our scheme of sampling from surrogate proportions, although designed primarily for settings where $M > 1$, may be easily specialised to settings where $M = 1$.

Remark 3.3. Consider a simple MO-MAB setup with $M = 3$, $K = 2$, $\mu_{1,\cdot} = [100\,0\,50]^\top$, and $\mu_{2,\cdot} = [0\,100\,50+\varepsilon]^\top$, where $\varepsilon > 0$ is small. For the above instance v, we have $I^*(v) = (1, 2, 2)$. Furthermore, the set of Pareto optimal arms is $\{1, 2\}$; see Appendix B for a formula to compute the Pareto optimal arms. While the best arm of each objective is also Pareto optimal for the preceding instance, existing algorithms for Pareto optimal arm identification fail to assert that arm 2 is the best arm for objective 3, a task that is significantly complex courtesy of the small gap $\Delta_{2,3} = \varepsilon$. In contrast, Proposition 3.1 demonstrates that the complexity of identifying arm 2 as the best arm for objective 3 scales as $\Omega(1/\varepsilon^2)$.

4 Achievability: Proposed Method

In this section, we describe our computationally efficient algorithm named MO-BAI based on the idea of *surrogate proportion* for pulling arms at each time step, and we also provide the pseudocode in Algorithm 1. Before we present the algorithm formally, we introduce some notations. Let $\widehat{\mu}_{i,m}(t)$ denote the empirical mean of rewards obtained from objective m of arm i up to time t, i.e., $\widehat{\mu}_{i,m}(t) := \frac{1}{N_{i,t}} \sum_{s=1}^{t} \mathbf{1}_{\{A_s=i\}}\, r_{s,m}$, and $N_{i,t} := \sum_{s=1}^{t} \mathbf{1}_{\{A_s=i\}}$ denotes the number of times arm i is pulled until time t. Let $\widehat{\Delta}_{i,m}(t) := \widehat{\mu}_{\widehat{i}_m(t),m}(t) - \widehat{\mu}_{i,m}(t)$ denote the empirical gap of the tuple $(i,m) \in [K] \times [M]$, where $\widehat{i}_m(t) \in \arg\max_{\iota\in[K]} \widehat{\mu}_{\iota,m}(t)$ denotes the empirical best arm of objective m at time t. Let $g_v(\cdot)$ be as defined in (7), and for all $m \in [M]$ and $i \in [K]$, let

$$g_v^{(i,m)}(\omega) := \frac{\Delta_{i,m}^2(v)}{2} \cdot \frac{\omega_i\, \omega_{i_m^*(v)}}{\omega_i + \omega_{i_m^*(v)}}, \quad \omega \in \Gamma. \tag{8}$$

Given any $\eta > 0$, let $\Gamma^{(\eta)} := \left\{\omega \in \Gamma : \ \forall i \in [K],\ \omega_i \geq \frac{\eta}{K(1+\eta)}\right\}$, and for all $\omega, \mathbf{z} \in \Gamma$, let

$$h_v(\omega, \mathbf{z}) := \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \left\{ g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, \mathbf{z} - \omega \rangle \right\}. \tag{9}$$

See more details for the intuition of $h_v(\omega, \mathbf{z})$ in Appendix C. With the above notations in place, we now describe the surrogate proportion and the associated arm selection rule of our algorithm.

Algorithm 1 Multi-Objective Best Arm Identification (MO-BAI)

```
Input:
    $\delta \in (0,1)$: confidence level.
    $\eta > 0$: relaxation parameter.
Output: $\widehat{I}_\delta$: the best arms.
1: Pull each arm once.
2: Initialise the buffer $\mathbf{B}_{i,t} = 0$ for all $i \in [K]$ and $t \in [K]$.
3: for $t \in \{K+1, K+2, \ldots\}$ do
4:     Compute the empirical mean $\widehat{\mu}_{i,m}(t)$ for each $(i,m) \in [K] \times [M]$.
5:     Compute the current empirical proportion
           $\widehat{\omega}_{i,t-1} := \frac{N_{i,t-1} + \mathbf{B}_{i,t-1}}{t-1}$.
6:     Set $l_t \leftarrow \max_{k\in\mathbb{N}:2^k\leq t} 2^k$.
7:     Compute the surrogate proportion for instance $\widehat{v}_{l_t}$ via
           $\mathbf{s}_t = \arg\max_{\mathbf{s}\in\Gamma^{(\eta)}} h_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}, \mathbf{s})$.
8:     Set $t \leftarrow t+1$.
9:     Pull arm $A_t \in \arg\max_{i\in[K]} [\mathbf{B}_{\cdot,t-1} + \mathbf{s}_t]_i$.
10:    Update the Buffer $\mathbf{B}_{\cdot,t} \leftarrow \mathbf{B}_{\cdot,t-1} + \mathbf{s}_t - \mathbf{e}_{A_t}$.
11:    if $Z(t) > \beta(t,\delta)$ then
12:        $\widehat{I}_\delta \leftarrow$ the empirical best arms at time $t$.
13:        break.
14:    end if
15: end for
16: return Best arms $\widehat{I}_\delta$.
```

4.1 Surrogate Proportion and Arm Selection Rule

Let $l_t := \max_{k\in\mathbb{N}:2^k\leq t} 2^k$, and let $\widehat{v}_t$ denote the empirical instance with means $[\widehat{\mu}_{i,m}(t') : (i,m) \in [K] \times [M]]^\top$ for $t \in \mathbb{N}$ and $t' = \max\{t-1, K\}$. As indicated in Line 7 of Algorithm 1, l_t serves to regulate the frequency of updating the empirical instance used for extracting surrogate proportions. The *surrogate proportion* at time step t, denoted $\mathbf{s}_t$, is defined as

$$\mathbf{s}_t := \arg\max_{\mathbf{s}\in\Gamma^{(\eta)}} h_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}, \mathbf{s}), \tag{10}$$

where $\widehat{\omega}_{\cdot,t-1}$ is an empirical proportion that will be defined shortly. Notice that $\mathbf{s}_t$ is a probability distribution on the arms, with each component strictly positive ($\geq \frac{\eta}{K\,(1+\eta)} > 0$), thereby placing a strictly positive mass on each arm.

To describe the empirical proportion $\widehat{\omega}_{\cdot,t}$, we introduce additional notations. Given $a \in [K]$, let $\mathbf{e}_a \in \mathbb{R}^K$ denote the *one-hot* vector of length K with a '1' in the ath component and '0' in the other components.

Let $\mathbf{B}_{\cdot,t} = [\mathbf{B}_{i,t} : i \in [K]]^\top \in \mathbb{R}^K$ denote a generic vector of length K; in the sequel, we refer to $\mathbf{B}_{\cdot,t}$ as the *buffer* at time t. With the above notations in place, the quantity $\widehat{\omega}_{\cdot,t} = [\widehat{\omega}_{i,t} : i \in [K]]^\top \in \Gamma$ is defined via

$$\widehat{\omega}_{i,t} \coloneqq \frac{N_{i,t} + \mathbf{B}_{i,t}}{t}. \tag{11}$$

Arms selection rule We now describe the arms selection rule of our algorithm. To begin with, each arm is pulled once in the first K time steps $t = 1, \ldots, K$. Initialising $B_{i,t} = 0$ for all $i \in [K]$, for all $t \leq K$, the algorithm computes $\widehat{\omega}_{\cdot,t}$ according to (11). Subsequently, the algorithm computes the surrogate proportion $\mathbf{s}_t$ using (10), and pulls arm A_t according to the rule

$$A_t \in \underset{i \in [K]}{\arg\max}\, [\mathbf{B}_{\cdot,t-1} + \mathbf{s}_t]_i, \tag{12}$$

where $[\mathbf{x}]_i$ represents the i-th component of $\mathbf{x}$. Following this, the algorithm iteratively updates the buffer according to the rule

$$\mathbf{B}_{\cdot,t} = \mathbf{B}_{\cdot,t-1} - \mathbf{e}_{A_t} + \mathbf{s}_t. \tag{13}$$

By examining Lines 5, 9, and 10 of Algorithm 1 in unison, it is evident that the buffer $\mathbf{B}$ is used to record the number of times each arm is scheduled to be pulled.

The above process repeats until the stoppage. Some remarks are in order.

Remark 4.1. The algorithm is to pull arms at each time step t according to proportion $\mathbf{s}_t$, while noting that only one arm is to be pulled exactly once per time in the setting. To achieve this, we utilize a buffer to track the number of times each arm should be pulled.

For instance, consider $K = 2$. When $\mathbf{s}_1 = [0.1, 0.9]$ at $t = 1$, the agent pulls the arm 1, resulting in an updated buffer of $\mathbf{B}_{\cdot,1} = [0.1, -0.1]$. At $t = 2$, suppose $\mathbf{s}_2 = [0.5, 0.5]$. Adding this to the existing buffer, we obtain $\mathbf{B}_{\cdot,1} + \mathbf{s}_2 = [0.6, 0.4]$. Since arm 1 has the highest value, it is pulled, and the buffer is $\mathbf{B}_{\cdot,2} = [-0.4, 0.4]$.

Remark 4.2. When $M = 1$, the well-known Track-and-Stop algorithm of Garivier and Kaufmann (2016) traditionally computes the oracle weight $\widetilde{\omega}_{\cdot,t} = \arg\max_{\omega \in \Gamma} g_{\widehat{v}_t}(\omega)$ and samples arm A_t guided by $\widehat{\omega}_{\cdot,t}$ at each time step t. Solving the preceding maximisation problem at each time step becomes computationally demanding, as exemplified by the authors therein, which is further exacerbated in scenarios where $M > 1$. In fact, it is easy to show that for any instance $v \in \mathcal{P}$, the maximisation problem $\sup_{\omega \in \Gamma} g_v(\omega)$ is a convex program (noting that the function g_v defined in (7) is concave; see Appendix G.4 for the details) with $O(K)$ variables and $O(MK)$ constraints. However, solving this convex program *exactly* through an iterative algorithm (such as the ellipsoid method) would necessitate running infinitely many iterations, thus rendering this approach practically infeasible. Of course, in practical implementations, one may only need to solve the convex program *approximately*; however, in this case, providing theoretical guarantees would also be commensurately more challenging.

In contrast, our proposed strategy for computing surrogate proportions in (10) offers computational efficiency by maximising h_v defined in (9) which serves as a surrogate version of g_v (hence the name surrogate proportion for $\mathbf{s}_t$). Notice that h_v is a linear function of its second argument. Thus, the optimisation problem in (10) is a simple *linear program* that may be solved *efficiently* and *exactly* using standard techniques such as the simplex method that typically takes polynomial time (Spielman and Teng, 2004). In this manner, we sidestep the intricacies of solving for the oracle weight at each time step, hence making our approach well-suited for scenarios with $M > 1$.

Remark 4.3. The astute reader will observe a striking similarity between the expression for h_v and that of the linear approximation of g_v, specified around any point ω by $g_v(\omega) + \langle \nabla_\omega g_v(\omega), \mathbf{z} - \omega \rangle$ for $\mathbf{z}$ lying in a neighborhood of ω. However, the two expressions are vastly different. While optimising the linear approximation of g_v in place of h_v to compute surrogate proportions, it might possibly lead to the convergence arising from the surrogate proportions to a sub-optimal weight in the long run. We show that our approach leads to the convergence of the empirical proportion to an almost-optimal weight for which the evaluation of g_v matches the constant $c^*(v)$ of the lower bound up to a factor of $1 + \eta$. See Section 5 for more details.

Remark 4.4. The sequence $\{l_t\}_{t=1}^{\infty}$ is strategically designed to minimize frequent alterations to the surrogate function $h_{\widehat{v}_{l_t}}$, thereby facilitating precise control over the *estimation error*—the difference between $h_{\widehat{v}_{l_t}}$ and $g_{\widehat{v}_{l_t}}$—in our analytical framework. Detailed insights can be found in Lemma G.8. Additionally, $\eta > 0$ is judiciously chosen to prevent the estimation error from diverging to infinity.

4.2 Stopping and Recommendation Rules

We now delineate the stopping and recommendation rules employed in our algorithm. We adopt a variant of Chernoff's stopping rule (Kaufmann et al., 2016; Lattimore and Szepesvári, 2020); our design is particularly inspired by Chen et al. (2023). Specifically, let

$$Z(t) \coloneqq \min_{m \in [M]} \min_{i \in [K] \setminus \widehat{i}_m(t)} \frac{N_{i,t}\, N_{\widehat{i}_m(t),t}\, \widehat{\Delta}^2_{i,m}(t)}{2(N_{i,t} + N_{\widehat{i}_m(t),t})} \tag{14}$$

denote a test statistic at time t. We define the stopping time of our algorithm via

$$\tau_\delta = \min\{t \geq K : Z(t) > \beta(t, \delta)\}, \tag{15}$$

where the threshold $\beta(t,\delta) \coloneqq MK\log(t^2+t)+f^{-1}(\delta)$. Here, $f:(0,+\infty)\to(0,1)$ is defined as

$$f(x) \coloneqq \sum_{i=1}^{MK} \frac{x^{i-1}e^{-x}}{(i-1)!}, \quad x \in (0,+\infty). \tag{16}$$

As for the recommendation rule, for each objective, our algorithm outputs the arm with the highest empirical mean corresponding to the objective as the best arm of the objective. That is, $\widehat{I}_\delta = [\widehat{i}_1(\tau_\delta),\dots,\widehat{i}_M(\tau_\delta)]^\top \in [K]^M$, where $\widehat{i}_m(\tau_\delta) \in \arg\max_{i\in[K]} \widehat{\mu}_{i,m}(\tau_\delta)$ for each $m \in [M]$.

Integrating the arms selection rule (with a fixed parameter $\eta > 0$), stopping rule, and recommendation rule delineated above, our algorithm for multi-objective best arm identification, abbreviated as MO-BAI, is presented in Algorithm 1.

Remark 4.5. Note that MO-BAI selects each arm once in the first K time slots. That is, $N_{\cdot,K} = [1,\dots,1]^\top$. Together with (11) and (13), it implies that $\widehat{\omega}_{\cdot,t} \in \Gamma^{(\eta)}$ for all $t > K$.

4.3 Performance of MO-BAI

In this section, we characterise the performance of MO-BAI with a fixed input parameter η. The first result below asserts that MO-BAI is δ-PAC for any $\delta \in (0,1)$.

Proposition 4.6. *Fix $\eta > 0$ and $\delta \in (0,1)$. Then,* MO-BAI *with parameter η is δ-PAC, i.e., $\forall v \in \mathcal{P}$*

$$\mathbb{P}_v^{\text{MO-BAI}}(\tau_\delta < +\infty) = 1 \quad \textit{and} \tag{17}$$

$$\mathbb{P}_v^{\text{MO-BAI}}\big(\hat{I}_\delta = I^*(v)\big) \geq 1-\delta. \tag{18}$$

We present the proof in Appendix F. It is worth noting that the form of $Z(t)$ and the threshold $\beta(t,\delta)$ play important roles in the proof. The following result demonstrates an asymptotic upper bound on the stopping time of MO-BAI.

Theorem 4.7. *Under* MO-BAI *with parameter $\eta > 0$, $\forall v \in \mathcal{P}$,*

$$\limsup_{\delta\downarrow 0} \frac{\mathbb{E}_v[\tau_\delta]}{\log(\frac{1}{\delta})} \leq (1+\eta)\,c^*(v) \quad \textit{and} \tag{19}$$

$$\mathbb{P}_v^{\text{MO-BAI}}\left(\limsup_{\delta\downarrow 0} \frac{\tau_\delta}{\log(\frac{1}{\delta})} \leq (1+\eta)\,c^*(v)\right) = 1. \tag{20}$$

We present the proof in Appendix G. Observe that in the asymptotic limit of $\delta \downarrow 0$, the upper bound in (19) matches the lower bound in (6) up to a factor $1+\eta$. The parameter η can be set arbitrarily close to 0, thereby establishing the asymptotic optimality of MO-BAI.

Remark 4.8. Our novel idea of surrogate proportion and the MO-BAI algorithm are inspired, at a high level, by the gradient-based algorithms in Jaggi (2013); Wang et al. (2021); Ménard (2019). Yet, notably, while our algorithm operates with $M \geq 1$ objectives, the algorithm in Wang et al. (2021) is specifically designed to operate with a *single* objective ($M = 1$). Adapting the algorithm in Wang et al. (2021) to our setting, while maintaining computational tractability and considering the exponential $O(K^M)$ scaling of possible sets of best arms across M objectives, presents challenges in preserving asymptotic optimality (Theorem 4.7). Furthermore, even for the case when $M = 1$, our algorithm differs substantially from that in Wang et al. (2021). The latter functions on the concept of an *r-subdifferential subspace* (an idea that is in turn adapted from Ravi et al. (2019)) and involves dealing with an infinite sequence of hyperparameters $\{r_i\}$ that must be chosen carefully. In contrast, our algorithm operates *without the need to choose/tune these hyperparameters.*

5 Proof Sketch of Theorem 4.7

In this section, we outline the key ideas that go into the proof of Theorem 4.7. At a high level, the proof encompasses two important steps: (1) Deriving a bound on the limiting value of $Z(t)/t$, and (2) Deriving a upper bound on the stopping time τ_δ based on the limiting value of $Z(t)/t$. Throughout, we fix an underlying instance v.

5.1 Bounding the Limiting Value of $Z(t)/t$

The key to deriving a "good" upper bound for the limiting value of $Z(t)/t$ lies in establishing that the evaluations of g_v at the oracle weights arising from MO-BAI (with input parameter η) match, in the long run, with the constant $c^*(v)$ appearing in the lower bound up to a factor of $1+\eta$. Towards this, we introduce the following auxiliary constant $\forall \eta > 0$:

$$C(v,\eta) \coloneqq \sup_{\substack{\omega,\mathbf{y}\in\Gamma^{(\eta)},\\ \gamma\in(0,1),\\ \mathbf{z}=\omega+\gamma(\mathbf{y}-\omega)}} \frac{2}{\gamma^2}\Big(h_v\big(\omega,\ \mathbf{z}-\omega\big) - g_v\big(\mathbf{z}\big)\Big). \tag{21}$$

The above definition appears to bear close resemblance with the definition for the curvature of a function, but they are not equivalent (see details in Appendix G). We show in Lemma G.5 that $C(v,\eta) < +\infty$ for all $\eta > 0$. For all $\eta > 0$, let

$$\tilde{c}(v,\eta)^{-1} \coloneqq \sup_{\omega\in\Gamma^{(\eta)}} g_v(\omega). \tag{22}$$

In Lemma G.13, we show that $c^*(v) \leq (1+\eta)\,\tilde{c}(v,\eta)$. With the above notations in place, the the key result of this section is presented below.

Lemma G.8. Fix $\eta > 0$. For all $t_1, t_2 \in \mathbb{N}$ with $t_2 > t_1 > K$, under MO-BAI with input parameter η, we have

$$|\tilde{c}(v,\eta)^{-1} - g_v(\widehat{\omega}_{\cdot,t_2})| \le \frac{t_1}{t_2}\,\tilde{c}(v,\eta)^{-1} + 11\,\epsilon_{t_1}(v) + \frac{2\,\log(t_2)\,\overline{C}_{t_1}(\eta)}{t_2} \qquad (23)$$

almost surely, where for any time step t:

- $\epsilon_t(v) \coloneqq \sup_{t' \ge l_t} \sup_{\omega \in \Gamma^{(\eta)}} \left|g_v(\omega) - g_{\widehat{v}_{t'}}(\omega)\right|$.
- $\overline{C}_t(\eta) \coloneqq \sup_{t' \ge l_t} C(\widehat{v}_{t'}, \eta)$.

We gather from Lemma G.8 that $\lim_{t_1\to\infty} \epsilon_{t_1}(v) = 0$ and $\limsup_{t_1\to\infty} \overline{C}_{t_1}(\eta) < +\infty$ almost surely; we present the details in Appendix G. Setting $t_2 = t_1^{-\lambda}$ for some $\lambda \in (0,1)$, we demonstrate that as $t_1 \to \infty$, the right-hand side of (23) converges to 0 almost surely, which in turn implies that $g_v(\widehat{\omega}_{\cdot,t_2})$ converges to $\tilde{c}(v,\eta)^{-1}$ in the limit as $t_2 \to \infty$ almost surely. This leads to the following quantitative characterisation of the limiting value of $Z(t)/t$.

Lemma G.11. Fix $\eta > 0$ and consider the non-stopping version of MO-BAI with input parameter η (i.e., a policy in which the stopping rule corresponding to Lines 11-14 in Algorithm 1 are not executed). Under this policy,

$$\lim_{t\to\infty} \frac{Z(t)}{t} = \tilde{c}(v,\eta)^{-1} \quad \text{almost surely.}$$

5.2 Bounding the Stopping Time τ_δ

Using the limiting value of $Z(t)/t$ derived earlier, we upper bound the stopping time almost surely and in expectation. First, we introduce two auxiliary terms.

(1) $T_{\text{gap}}(v,\eta,\epsilon)$: For any $\epsilon > 0$, let $T_{\text{gap}}(v,\eta,\epsilon)$ denote the smallest positive integer-valued random variable such that

$$\left|\frac{Z(t)}{t} - \tilde{c}(v,\eta)^{-1}\right| \le \epsilon \quad \forall\, t \ge T_{\text{gap}}(v,\eta,\epsilon).$$

Thanks to Lemma G.11, we have that $T_{\text{gap}}(v,\eta,\epsilon)$ is finite almost surely. We show in Lemma G.14 that $T_{\text{gap}}(v,\eta,\epsilon)$ also has finite expectation.

(2) $T_{\text{thres}}(v,\eta,\epsilon,\delta)$: For $\delta \in (0,1)$ and $\eta > 0$, let

$$T_{\text{thres}}(v,\eta,\epsilon,\delta) \coloneqq 1 + \frac{f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1} - \epsilon} + \frac{MK}{\tilde{c}(v,\eta)^{-1} - \epsilon} \log\left(\left(\frac{2f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1} - \epsilon}\right)^2 + \frac{2f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1} - \epsilon}\right).$$

The right-hand side of the above expression is carefully designed to meet the following requirement.

Lemma G.10. Fix $\delta \in (0,1)$ and $\eta > 0$, and consider the threshold $\beta(t,\delta)$ in MO-BAI. For all $\epsilon \in \left(0, \tilde{c}(v,\eta)^{-1}\right)$, there exists $\delta_{\text{thres}}(v,\eta,\epsilon) > 0$ such that for all $\delta \in (0, \delta_{\text{thres}}(v,\eta,\epsilon))$,

$$t\tilde{c}(v,\eta)^{-1} > \beta(t,\delta) + \epsilon t \quad \forall t \ge T_{\text{thres}}(v,\eta,\epsilon,\delta).$$

Using Lemma G.10, we show that for any $\epsilon \in \left(0, \tilde{c}(v,\eta)^{-1}\right)$ and $\delta \in (0, \delta_{\text{thres}}(v,\eta,\epsilon))$, almost surely,

$$\tau_\delta \le T_{\text{gap}}(v,\eta,\epsilon) + T_{\text{thres}}(v,\eta,\delta,\epsilon) + K + 1.$$

This crucial step leads to $\limsup_{\delta\downarrow 0} \tau_\delta \cdot \frac{\tilde{c}(v,\eta)^{-1} - \epsilon}{f^{-1}(\delta)} \le 1$ almost surely, because $T_{\text{gap}}(v,\eta,\epsilon)$ is independent of δ and that $\lim_{\delta\downarrow 0} T_{\text{thres}}(v,\eta,\delta,\epsilon) \cdot \frac{\tilde{c}(v,\eta)^{-1} - \epsilon}{f^{-1}(\delta)} = 1$. Finally, using the fact that $\lim_{\delta\downarrow 0} \frac{f^{-1}(\delta)}{\log(1/\delta)} = 1$, and letting $\epsilon \downarrow 0$, we arrive at (19) and (20). It is noteworthy that our proof techniques are applicable in a wide range of pure exploration problems with single-dimensional and multi-dimensional rewards from arms.

6 Numerical Study

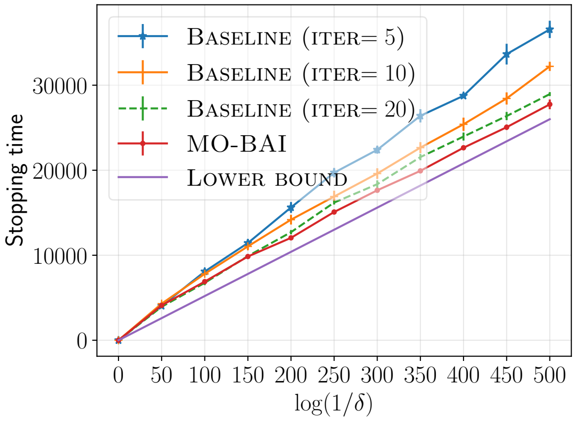

Figure 1: Plot of average stopping times of MO-BAI and Baseline (with varying iteration numbers) for the synthetic dataset.

We run experiments to validate the effectiveness of MO-BAI through empirical assessments on the SNW dataset (Zuluaga et al., 2016) and a synthetic dataset, and their detailed descriptions are presented in the below subsection. Specifically, we compare our algorithm against Baseline, a multi-objective adaptation of the D-Tracking algorithm by Garivier and Kaufmann (2016) which was originally designed for a single objective (see details in Appendix A.3), and a Successive Elimination-based (Even-Dar et al., 2006) algorithm. As alluded to in Remark 4.2, the implementation of Baseline involves solving an optimization problem, which, in scenarios with $M > 1$, may not be exactly resolved. Hence, we employ an iterative method (details

$\log(1/\delta)$	10	50	100
MO-BAI	**1,686.67 ± 21.03**	**5,097.33 ± 90.78**	**8,435.0 ± 115.23**
Baseline (iter=100)	$7,428.0 \pm 139.69$	$18,612.67 \pm 214.59$	$30,914.33 \pm 622.61$
Baseline (iter=20)	$10,584.67 \pm 313.52$	$37,316.33 \pm 406.90$	$65,395.0 \pm 602.98$

Table 1: Comparison of the empirical stopping times for the SNW dataset (Zuluaga et al., 2016).

MO-BAI	Baseline with iter Iteration Steps		
	iter = 5	iter = 10	iter = 20
10.0 ± 7.2	41.0 ± 10.0	92.2 ± 19.8	184.3 ± 41.0

Table 2: Average computation times (in ms) for a single execution of the optimisation routines in MO-BAI and Baseline on Apple M1 Chip with 16GB of RAM.

in Appendix A and pseudo code of Algorithm 3), and the accuracy of the obtained solution depends on the number of iterations.

In addition, for the fairness of comparison, we adopt the operation of force exploration of Baseline (i.e., line 3 to 4 of Algorithm 2) in the implementation of MO-BAI, which would not affect the theoretical guarantee. Also, we employ the threshold $\widehat{c}_t(\delta) = \log((1 + \log t)/\delta)$ in both MO-BAI and Baseline in the numerical study, and note that this is different from our theoretical threshold $\beta(t, \delta)$. Additional details regarding the modified threshold $\widehat{c}_t(\delta)$ and its justification in the context of our work can be found in Appendix A.2.

6.1 Simulation Environment

The experiments were executed on an Apple M1 Chip with 16 GB of memory, operating on Mac OS 14.2.1. The linear programming procedures outlined in Algorithm 1 and Algorithm 3 were executed using SciPy v1.11.3 within the Python 3.11.2 environment.

6.2 Descriptions of the Datasets

Synthetic Dataset: Our synthetic dataset is generated with parameters $K = 20$ and $M = 10$. For all pairs (i, m) where $i \neq m$, $\mu_{i,m}$ is uniformly chosen from the interval $[0, 1]$. For pairs (i, m) where $i = m$, $\mu_{i,m}$ is uniformly selected from $[1.2, 2]$. These values remain constant throughout the experiment. Let $v = [\mu_{i,m} : (i, m) \in [K] \times [M]]^\top$. It is evident that $i^*_m(v) = m$ for every $m \in [M]$. Additionally, $\Delta_{i,m}(v) > 0.2$ for all $i \neq i^*_m(v)$.

SNW Dataset: We adopt the SNW dataset introduced by Zuluaga et al. (2016), consisting of 206 distinct hardware implementations of a sorting network. Following the protocol outlined by Ararat and Tekin (2023), the objective values, represented by the negative of the area, serve as mean rewards for the designs, and Gaussian noises are added to the mean rewards in the bandit dynamic. Consequently, in this dataset, we have $K = 206$ and $M = 2$. To facilitate the simulation, we scale the rewards of each arm by a factor of 10.

6.3 Results

A comparison of MO-BAI with $\eta = 0.1$ and Baseline with varying iteration counts (iter $= 5, 10, 20$) for the synthetic dataset is depicted in Figure 1. The error bars in the figures are obtained from running three independent trials. Further, the computation times (in ms) for a single execution of the optimization routines in Baseline and MO-BAI are shown in Table 2. Clearly, the figures demonstrate the superior performance of MO-BAI compared to Baseline. As the number of iteration steps in Baseline increases, its performance approaches that of MO-BAI, an artefact of improved accuracy in solving the optimization routine in Baseline. However, this improvement comes at a significant escalated computational cost, as evidenced by Table 2. For the SNW dataset, we run three independent trials, and average the stopping times from these trials to obtain the values in Table 1. The tabulated results indicate the superior performance of our proposed MO-BAI algorithm over Baseline on the SNW dataset. These results, as well as the other experimental results of the SNW dataset in Appendix A underscore the efficacy of MO-BAI.

7 Conclusions and Future Work

This work considered a novel best arm identification setting in which a single pull of an arm yields an M-dimensional vector as its reward. The goal was to identify the M best arms, one corresponding to each dimension, under the fixed-confidence regime. We developed an efficient algorithm based on the original idea of *surrogate proportions*, that we proved is asymptotically optimal and computationally efficient. We conducted empirical studies on a synthetic dataset and the SNW datasets to substantiate the proposed algorithm's computational efficiency and asymptotic optimality. Our results are asymptotically optimal in the sense that the results are tight as the error probability $\delta \downarrow 0$. It would be fruitful to investigate whether the ideas in non-asymptotic strengthenings of single-objective pure exploration problems (e.g., Degenne et al. (2019)) carry over to our multi-objective setting.

Acknowledgement

This work is funded by Singapore Ministry of Education AcRF Tier 1 grants (A-8000189-01-00 and A-8000980-00-00).

References

Ararat, C. and Tekin, C. (2023). Vector optimization with stochastic bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 2165–2190. PMLR.

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p.

Auer, P., Chiang, C.-K., Ortner, R., and Drugan, M. (2016). Pareto front identification from stochastic bandit feedback. In *Artificial Intelligence and Statistics*, pages 939–947. PMLR.

Busa-Fekete, R., Szörényi, B., Weng, P., and Mannor, S. (2017). Multi-objective bandits: Optimizing the generalized gini index. In *International Conference on Machine Learning*, pages 625–634. PMLR.

Chen, Z., Karthik, P. N., Chee, Y. M., and Tan, V. (2024). Fixed-budget differentially private best arm identification. In *The Twelfth International Conference on Learning Representations*.

Chen, Z., Karthik, P. N., Tan, V. Y. F., and Chee, Y. M. (2023). Federated best arm identification with heterogeneous clients. *IEEE Transactions on Information Theory*, pages 1–1.

de la Pena, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564.

Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32.

Degenne, R., Koolen, W. M., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*.

Degenne, R., Ménard, P., Shang, X., and Valko, M. (2020). Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR.

Drugan, M. M. and Nowe, A. (2013). Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6).

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR.

Jaggi, M. (2013). Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR.

Jedra, Y. and Proutiere, A. (2020). Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42.

Kaufmann, E. and Koolen, W. M. (2021). Mixture martingales revisited with applications to sequential tests and confidence intervals. *The Journal of Machine Learning Research*, 22(1):11140–11183.

Kim, W., Iyengar, G., and Zeevi, A. (2023). Pareto front identification with regret minimization. *arXiv preprint arXiv:2306.00096*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lizotte, D. J. and Laber, E. B. (2016). Multi-objective Markov decision processes for data-driven decision support. *The Journal of Machine Learning Research*, 17(1):7378–7405.

Ménard, P. (2019). Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*.

Mukherjee, A. and Tajer, A. (2023). Best arm identification in stochastic bandits: Beyond β- optimality. *arXiv preprint arXiv:2301.03785*.

Prabhu, G. R., Bhashyam, S., Gopalan, A., and Sundaresan, R. (2022). Sequential multi-hypothesis testing in multi-armed bandit problems: An approach for asymptotic optimality. *IEEE Transactions on Information Theory*, 68(7):4790–4817.

Ravi, S. N., Collins, M. D., and Singh, V. (2019). A deterministic nonsmooth Frank–Wolfe algorithm with coreset guarantees. *Informs Journal on Optimization*, 1(2):120–142.

Reddy, K. S., Karthik, P. N., and Tan, V. Y. F. (2023). Almost cost-free communication in federated best arm identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8378–8385.

Shang, X., Shao, H., and Qian, J. (2020). Stochastic bandits with vector losses: Minimizing ℓ_∞-norm of relative losses. *arXiv preprint arXiv:2010.08061*.

Shi, C., Shen, C., and Yang, J. (2021). Federated multi-armed bandits with personalization. In *International Conference on Artificial Intelligence and Statistics*, pages 2917–2925. PMLR.

Spielman, D. A. and Teng, S.-H. (2004). Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Turgay, E., Oner, D., and Tekin, C. (2018). Multi-objective contextual bandit problem with similarity information. In *International Conference on Artificial Intelligence and Statistics*, pages 1673–1681. PMLR.

Wang, P.-A., Tzeng, R.-C., and Proutiere, A. (2021). Fast pure exploration via Frank–Wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821.

Xu, M. and Klabjan, D. (2023). Pareto regret analyses in multi-objective multi-armed bandit. In *International Conference on Machine Learning*, pages 38499–38517. PMLR.

Zuluaga, M., Krause, A., and Püschel, M. (2016). e-pal: An active learning approach to the multi-objective optimization problem. *Journal of Machine Learning Research*, 17(104):1–32.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, the problem setting is described in Section 2, and the algorithm is described in Section 4.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, the sample complexity is described in Section 4.3.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, assumptions are described in Section 2.]
   (b) Complete proofs of all theoretical results. [Yes, the complete proofs are described in Appendix]
   (c) Clear explanations of any assumptions. [Yes, they are described in Section 2.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes. Our proposed algorithm is rather simple, and our results mainly focus on theoretical analysis. In addition, we have provided pseudo code in the main text and the python code in https://github.com/zchen42/simulation_mobai .]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, they are described in the Section 6 and Appendix A.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, we have provided error bars in each experiment.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, they are described in Appendix 6.1]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A More Details on Numerical Study

In this section, we provide more details on our experimental implementation.

A.1 Additional Experiments for the SNW Dataset

A potential concern for the reader may be the small δ (e.g. $\log(1/\delta) = 10$) conducted in our experiments, which may be not practical in the real-world application.

To assuage the above concern, we conducted additional experiments on the practical δ (i.e., $\delta = 0.1$ and $\delta = 0.05$). The results, presented in Table 3, show the average stopping times across 100 trials. Consistent with our initial findings, we observe that our MO-BAI is again superior than Baseline algorithm in these experiments.

	$\delta = 0.1$	$\delta = 0.05$
MO-BAI	$\mathbf{968.82 \pm 58.21}$	$\mathbf{1023.77 \pm 67.42}$
Baseline	4485.98 ± 124.92	6168.29 ± 132.01
Baseline-Non-Unif	3841.05 ± 136.44	4320.55 ± 128.26
MO-SE	2322.39 ± 461.54	2411.16 ± 421.88

Table 3: Average stopping times obtained by running 100 independent trials with practical error probability $\delta = 0.1$ and $\delta = 0.05$ for the SNW dataset (Zuluaga et al., 2016). In Baseline and Baseline-Non-Unif, we set iter $= 20$.

A.2 Curated Threshold for Simulations

It is customary in the fixed-confidence BAI literature to employ thresholds in simulations that differ from theoretical thresholds. Notably, in the single-objective case, Garivier and Kaufmann (2016) utilized $\beta_{\mathrm{GK}}^{\mathrm{empirical}}(t, \delta) = \log((1 + \log t)/\delta)$ for empirical evaluation, a threshold that is (a log factor) smaller than the theoretical threshold $\beta_{\mathrm{GK}}^{\mathrm{theoretical}}(t, \delta) = \log(C_K t^{\alpha}/\delta)$ employed in their D-Tracking algorithm. Here, $\alpha > 1$ is a parameter of the D-Tracking algorithm, and C_K is a universal constant that depends only on the number of arms K. In a recent work, Kaufmann and Koolen (2021) introduced the concept of "rank" for pure exploration problems. They demonstrated that for a problem with rank R, the threshold $\widehat{c}_t(\delta) = 3R \log(1 + \log(t/R)) + O(\log(1/\delta))$, when combined with the GLR test statistic, yields a δ-PAC stopping rule; cf. Kaufmann and Koolen (2021, Proposition 21). The authors therein note that BAI with a single objective is a problem of rank 2. It is noteworthy that the same holds true of multi-objective BAI problems; for a formal proof of this, see Appendix D. In light of the above rationale, we adopt the threshold $\log((1+\log t)/\delta)$ for our simulations following Garivier and Kaufmann (2016), and note that is different from our theoretical threshold $\beta(t, \delta)$ defined in Section 4.2.

Even with the modified threshold of $\log((1 + \log t)/\delta)$, we observe near-zero empirical error rates for MO-BAI and Baseline algorithms in our experiments, thereby implying that this modified threshold is still conservative in practice. This conforms with the heuristics presented in (Garivier and Kaufmann, 2016, Section 6), providing additional practical justification for adopting the modified threshold in our experiments.

A.3 The Baseline Algorithm and its Implementation Details

The pseudo-code of Baseline is presented in Algorithm 2. It is important to note that the approach proposed by Garivier and Kaufmann (2016) for solving the optimization problem in line 6 of Baseline becomes impractical when $M > 1$ due to the various best arms across different objectives. In our implementation of Baseline, we adopt the sub-routine in Algorithm 3 to solve the optimisation problem in Line 6 of Algorithm 2 by specifying the number of iterations steps iter, and its convergence can be established by the idea of Lemma G.8. Figure 1 and Tab. 2 show respectively the stopping times and computation times (in ms) incurred under iter $\in \{5, 10, 20\}$.

Furthermore, the threshold used in the implementation of Baseline is equal to the single-objective empirical threshold of $\log((1+\log t)/\delta)$ used in Garivier and Kaufmann (2016). Notably, this threshold remains independent of the number of objectives M. The rationale behind this choice stems from the fact that the "rank" of a multi-objective BAI problem with M independent objectives is equal to the rank of the single-objective BAI problem for all values of M. See Appendix D for further details.

Algorithm 2 BASELINE (Multi-objective adaptation of D-Tracking (Garivier and Kaufmann, 2016))

```
Input:
    $K \in \mathbb{N}$: number of arms
    $\delta \in (0,1)$: confidence level.
    IT: number of iteration steps
Output: $\widehat{I}_\delta$: the best arms.
 1: Pull each arm once.
 2: for $t \in \{K+1, K+2, \ldots\}$ do
 3:     if $\min_{i\in[K]} N_{i,t-1} < \sqrt{(t-1)/K}$ then
 4:         Pull arm $A_t \in \arg\min_{i\in[K]} N_{i,t-1}$; resolve ties uniformly.
 5:     else
 6:         Compute the empirical oracle weight
                $$\widetilde{\omega}_{\cdot,t} = \textsc{SubRoutine}(K, \textsc{it}, \widehat{v}_t).$$
 7:         Pull arm $A_t \in \arg\max_{i\in[K]} N_{i,t-1} - t\,\widetilde{\omega}_{i,t}$.
 8:     end if
 9:     if $Z(t) > \log\big((1+\log t)/\delta\big)$ then
10:         $\widehat{I}_\delta \leftarrow$ the empirical best arms of time $t$.
11:         break.
12:     end if
13: end for
14: return Best arms $\widehat{I}_\delta$.
```

Algorithm 3 Sub-routine to solve the optimisation in Line 6 of Algorithm 2 – SUBROUTINE(K, IT, $\widehat{v}$)

```
Input:
    $K \in \mathbb{N}$: number of arms.
    IT: number of iteration steps.
    $\widehat{v}$: empirical instance.
Output: $\widehat{\omega}$: the oracle weight.
1: Initialise $\widehat{\omega} = [1/K, \ldots, 1/K]^\top$, $\widetilde{\mathbf{N}} = [\widetilde{N}_i : i \in [K]]^\top = \mathbf{0}$.
2: for $k \in \{1, \ldots, \textsc{it}\}$ do
3:     Compute $\mathbf{s}_k = \arg\max_{\mathbf{s}\in\Gamma}\ h_{\widehat{v}}(\widehat{\omega}, \mathbf{s})$.
4:     Set $\widetilde{\mathbf{N}} \leftarrow \widetilde{\mathbf{N}} + \mathbf{s}_k$
5:     Update $\widehat{\omega} \leftarrow \frac{\widetilde{\mathbf{N}}}{k}$.
6: end for
7: return Oracle weight $\widehat{\omega}$.
```

As such, the BASELINE algorithm, with any value of ITER $< +\infty$, is not asymptotically optimal (though practically implementable), while asymptotically optimal but practically not implementable for ITER $= +\infty$. A plausible scheme to achieve asymptotic optimality, while ensuring practical feasibility, is to let ITER grow with t. For e.g., if $\textsc{iter}(t) = O(\log t)$, solving for $\widetilde{\omega}_{\cdot,t} = \arg\max_{\omega\in\Gamma} g_{\widehat{v}_t}(\omega)$ up to a $1/\text{poly}(t)$ error at time step t requires $O(\log t)$ iterations. However, quantifying the exact growth rate (e.g., $\textsc{iter}(t) = O(\log(t)), O(\sqrt{t}), O(t)$, or $O(\exp(t))$) that is necessary to achieve asymptotic optimality is a technically challenging task; the latter involves quantifying the approximation error of each subroutine as ITER grows with t, and ensuring that these errors amortize asymptotically as $t \to \infty$. Moreover, if $ITER(t)$ growth as t (e.g., $ITER(t) = \sqrt{t}$), the number of iterations at each time step will go to infinite as $t \to \infty$, while the number of iterations at each time step is still finite as $t \to \infty$ in MO-BAI.

Nonetheless, our experiments on both synthetic and the SNW datasets indicate that the stopping times typically remain below 10^5, implying that $\textsc{iter} = \log_2 t \leq 20$ at all times t for these datasets. As shown in Table 3, the MO-BAI algorithm outperforms the BASELINE algorithm with ITER set to 20, on the SNW dataset. This suggests that, in practice, there is limited benefit to allowing ITER to grow with t.

In addition, to enhance the comprehensiveness of our comparative analyses, we investigated alternative implementations of Algorithm 3 by modifying the initialization method of $\widehat{\omega}$ in Line 1 to $\widehat{\omega} = \tilde{\omega}_{\cdot,t-1}$ (i.e., the estimate of ω from the previous time step) at time step t instead of setting it to be the uniform distribution $[1/K, \dots, 1/K]^\top$. We call this variant Baseline-non-unif, and present the experimental outcomes in Table 3. The empirical findings underscore that our proposed MO-BAI is significantly better than Baseline-non-uniform on the SNW dataset.

Algorithm 4 MO-SE (Multi-objective adaptation of Successive Elimination (Even-Dar et al., 2006))

```
Input:
   $K \in \mathbb{N}$: number of arms
   $\delta \in (0,1)$: confidence level.
   IT: number of iteration steps
Output: $\widehat{I}_\delta$: the best arms.
Initialize $t = 0$
for $m \in \{1, 2, \dots, M\}$ do
    Set $\mathcal{S} = \{1, 2, \dots, K\}$
    while $|\mathcal{S}| > 1$ do
        Pull each arm in $\mathcal{S}$ once.
        $t \leftarrow t + |\mathcal{S}|$
        $\alpha_t = \sqrt{2\ln\left(4MKt^2/\delta\right)/t}$
        Eliminate all the arms $i$ in $\mathcal{S}$ with $\max_{j\in\mathcal{S}} \widehat{\mu}_{j,m}(t) - \widehat{\mu}_{i,m}(t) > 2\alpha_t$
    end while
    $\widehat{i}_m \leftarrow$ the only arm in $\mathcal{S}$
end for
return Best arms $\widehat{I}_\delta = (\widehat{i}_1, \dots, \widehat{i}_M)$.
```

A.4 Multi-Objective Successive Elimination (MO-SE) and its Implementation Details

We also implement a multi-objective version of a classical algorithm for BAI–Successive Elimination (Even-Dar et al., 2006). This algorithm, which we call Multi-Objective Successive Elimination (MO-SE) is shown formally in Algorithm 4. Specifically, in MO-SE, there are M rounds, and we determine the empirical best arm of $m-$th objective in $m-$th round using the principle of successive elimination. In particular, we set $\alpha_t = \sqrt{2\ln\left(4MKt^2/\delta\right)/t}$ in Algorithm 3 of Even-Dar et al. (2006), which is a natural adaption to the multiobjective case as there are total MK arms and the noises are Gaussian with unit variance in our setting. From a theoretical standpoint, MO-SE is not (asymptotically) optimal even in the case of $M = 1$, which is clearly inferior to our MO-BAI. This can also be clearly seen via the experimental results shown in Table 3, which again empirically underscores the superiority of our proposed MO-BAI over all considered baselines.

A.5 Impact of η on Performance of MO-BAI

We run MO-BAI on the synthetic dataset introduced in Section 6 for $\eta \in \{2.0, 1.0, 0.5, 0.1\}$. The results are shown in Figure 2. We observe that the performance for $\eta = 0.1$ is superior to that for $\eta > 0.1$, This observation aligns with our theoretical findings. Indeed, because $N_{i,t}/t \approx \widehat{\omega}_{i,t}$ for all large t (noting that $|B_{i,t}| \leq 1$), and $\min_{i\in[K]} \widehat{\omega}_{i,t} \geq \frac{\eta}{K(1+\eta)}$, it is evident that the fraction of times each arm is pulled in the long run increases with increase in η (as $\eta \mapsto \eta/(1+\eta)$ is an increasing function), thereby leading to larger stopping times. Furthermore, when the stopping time is not excessively large, the performances of $\eta = 0.1$ and $\eta = 0.5$ empirically demonstrate a notable degree of similarity. This phenomenon may be attributed to the empirical mean necessitating a greater number of arm pulls for stabilization.

B Multi-Objective BAI vs Pareto Frontier Identification

Our research framework shares structural similarities with those used in Pareto frontier identification, yet the tasks of our investigation are distinctly different. For clarity and ease of illustration in this section, we consider that the arm means are distinct across each objective, specifically, $\mu_{i,m} \neq \mu_{j,m}$ for all $m \in [M]$ and $i \neq j$. Then,

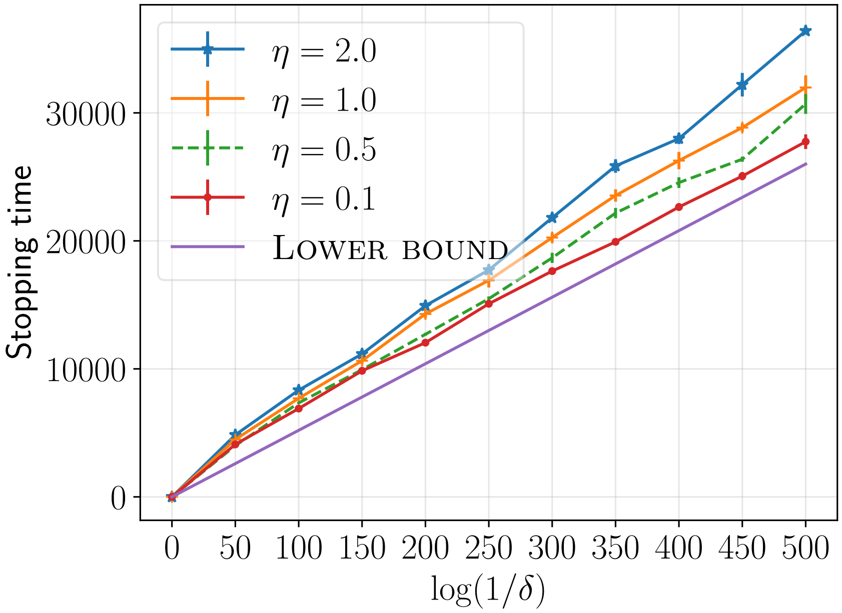

Figure 2: Comparison of the empirical stopping times with varying values of η for the synthetic dataset.

according to the extant literature (Auer et al., 2016; Ararat and Tekin, 2023; Kim et al., 2023) on Pareto frontier identification, an arm i is defined as Pareto optimal if, for every other arm j where $j \neq i$, there exists at least one objective $m \in [M]$ for which $\mu_{i,m} > \mu_{j,m}$. Consequently, under this definition, the best arm for each objective (as defined in our paper) is inherently Pareto optimal.

However, typical Pareto Frontier Identification algorithms are designed merely to determine the Pareto optimality of each arm without the capability to specifically identify the optimal arm for any given objective $m \in [M]$. Conversely, while our proposed algorithm effectively ascertains the Pareto optimality of the best arm *for each objective*, it does not ensure the identification of all Pareto optimal arms.

Thus, the nature of our research task diverges fundamentally from that of traditional Pareto Frontier Identification, reflecting distinct analytical goals and methodological requirements.

C More Intuitive Explanations for Surrogate Proportion

Our surrogate proportion and the MO-BAI algorithm are inspired, at a high level, by gradient-based methods. At time step t, the surrogate proportion under an empirical instance $\hat{v}$ is defined as

$$\mathbf{s} := \arg \max_{\mathbf{s}' \in \Gamma^{(\eta)}} h_{\hat{v}}(\widehat{\omega}_{\cdot,t-1}, \mathbf{s}'),$$

where the function $h_{\hat{v}}(\widehat{\omega}_{\cdot,t-1}, \cdot)$ may be viewed as a surrogate of $g_{\hat{v}}(\cdot)$, with $g_{\hat{v}}(\mathbf{z}) = h_{\hat{v}}(\widehat{\omega}_{\cdot,t-1}, \mathbf{z})$ for $\mathbf{z} = \widehat{\omega}_{\cdot,t-1}$. Our design of MO-BAI ensures that $h_{\hat{v}}(\widehat{\omega}_{\cdot,t-1}, \cdot)$ approximates $g_{\hat{v}}(\cdot)$ to some extent, so that

$$\arg \max_{\mathbf{s}' \in \Gamma^{(\eta)}} h_{\hat{v}}(\widehat{\omega}_{\cdot,t-1}, \mathbf{s}') \approx \arg \max_{\mathbf{s}' \in \Gamma^{(\eta)}} g_{\hat{v}}(\mathbf{s}') \text{ as } t \text{ grows} .$$

Despite selecting arms according to the surrogate proportion at each time step, the above property enables us to mimic the selection of arms as per the optimal allocation in the long run, thereby leading to matched upper and lower bounds.

D Rank of Multi-Objective Best Arm Identification

In this section, we show that the rank of multi-objective BAI problem with M independent objectives is equal to 2 for all values of M; notably, this is also the rank of the single-objective BAI problem. Before we present the formal arguments, we reproduce the definition of rank of a pure exploration problem from Kaufmann and Koolen (2021).

Definition D.1. (Kaufmann and Koolen, 2021, Definition 20) Fix constants $d, P, Q, R \in \mathbb{N}$. A sequential identification problem specified by a partition $\mathcal{O} = \bigcup_{p=1}^{P} \mathcal{O}_p$, where $O_p \subseteq \mathbb{R}^d$ for all p, is said to have rank R if for every $p \in \{1, \ldots, P\}$,

$$\mathcal{O} \setminus \mathcal{O}_p = \bigcup_{q=1}^{Q} \left\{ \boldsymbol{\lambda} \in \mathbb{R}^d : (\lambda_{k_1^{p,q}}, \ldots, \lambda_{k_R^{p,q}}) \in \mathcal{L}_{p,q} \right\} \tag{24}$$

for a family of indices $k_r^{p,q} \in [d]$ and open sets $\mathcal{L}_{p,q}$ indexed by $r \in [R]$, $q \in [Q]$, and $p \in [P]$. In other words, the problem has rank R if for each p, the set $\mathcal{O} \setminus \mathcal{O}_p$ is a finite union of sets that are each defined in terms of only R-tuples.

D.1 Rank of Single-Objective BAI Problem

Consider the classical BAI problem with a single objective ($M = 1$) and K arms, specified by a problem instance $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]^\top \in \mathbb{R}^K$. In this case, we have $d = P = K$, and for each $p \in [K]$, the set $\mathcal{O}_p = \{\boldsymbol{\lambda} \in \mathbb{R}^d : a^\star(\boldsymbol{\lambda}) = p\}$ consists of all problem instances with best arm p. Furthermore, for each $p \in [K]$,

$$\mathcal{O} \setminus \mathcal{O}_p = \bigcup_{q \neq p} \left\{ \boldsymbol{\lambda} \in \mathbb{R}^K : \lambda_q > \lambda_p \right\},$$

thereby implying that rank $R = 2$.

D.2 Rank of Multi-Objective BAI Problem

Consider now a multi-objective BAI problem with M independent objectives and K arms. In this case, a problem instance is specified by $\boldsymbol{\mu} = [\mu_{i,m} : (i,m) \in [K] \times [M]] \subseteq \mathbb{R}^{KM}$. We thus have $d = KM$. Also, for any $p = [p_1, \ldots, p_M]^\top \in [K]^M$, we have

$$\mathcal{O}_p = \{\boldsymbol{\lambda} \in \mathbb{R}^{KM} : \ \forall m \in [M], \ \ p_m = \text{best arm of objective } m \text{ under the instance } \boldsymbol{\lambda}\},$$

thereby implying that $P = K^M$. Furthermore,

$$\mathcal{O} \setminus \mathcal{O}_p = \bigcup_{m=1}^{M} \ \bigcup_{q = [q_1, \ldots, q_M]^\top \in [K]^M} \left\{ \boldsymbol{\lambda} = [\lambda_{i,m} : (i,m) \in [K] \times [M]]^\top \in \mathbb{R}^{KM} : \lambda_{q_m, m} > \lambda_{p_m, m} \right\}, \tag{25}$$

thereby implying that $R = 2$.

E Proof of Proposition 3.1

Firstly, we introduce a useful lemma adapted from Kaufmann et al. (2016).

Lemma E.1. *Fix $\delta > 0$ and a δ-PAC policy π with stopping time τ_δ. Let $\mathcal{F}_{\tau_\delta} = \sigma(\{(X_{A_t,m}(t), A_t) : t \in [\tau_\delta], m \in [M]\})$ denote the history of all the arm pulls and rewards seen up to the stopping time τ_δ under the policy π. Then, for any pair of instances $v, v' \in \mathcal{P}$ with arm means $\{\mu_{i,m} : i \in [K], m \in [M]\}$ and $\{\mu'_{i,m} : i \in [K], m \in [M]\}$ respectively, and any $\mathcal{F}_{\tau_\delta}$-measurable event E,*

$$\sum_{i=1}^{K} \sum_{m=1}^{M} \mathbb{E}_v^\pi \left[N_{i,\tau_\delta} \right] \frac{(\mu_{i,m} - \mu'_{i,m})^2}{2} \geq d_{\mathrm{KL}} \left(\mathbb{P}_v^\pi(E), \mathbb{P}_{v'}^\pi(E) \right), \tag{26}$$

where $D_{\mathrm{KL}}(p \| q)$ denotes the Kullback–Leibler (KL) divergence between distributions p and q, and $d_{\mathrm{KL}}(x, y)$ denotes the KL divergence between two Bernoulli distributions with parameters x and y.

The proof of Lemma E.1 follows along the same lines as in the proof of Kaufmann et al. (2016, Lemma 19), and is hence omitted. We then note the following lower bound derived from Lemma E.1.

Lemma E.2. *Fix $\delta > 0$ and instance $v \in \mathcal{P}$. For any δ-PAC policy π with stopping time τ_δ,*

$$\mathbb{E}_v^\pi[\tau_\delta] \geq \frac{\log(\frac{1}{4\delta})}{\sup_{\omega\in\Gamma}\ \inf_{v'\in\mathrm{Alt}(v)}\sum_{i=1}^K \omega_i \sum_{m=1}^M \frac{\left(\mu_{i,m}(v)-\mu_{i,m}(v')\right)^2}{2}},$$

where $\mathrm{Alt}(v)$ *denotes the set of problem instances with a set of best arms distinct from the set of best arm under v.*

Proof. Fix a δ-PAC policy π. Let

$$E^* \coloneqq \{\widehat{I}_\delta = I^*(v)\},$$

where $\widehat{I}_\delta$ denotes the set of best arms (one for each objective) output by policy π at stoppage. By Lemma E.1, we have

$$\sum_{m=1}^M \sum_{i=1}^K \mathbb{E}_v^\pi\left[N_{i,\tau_\delta}\right] \frac{(\mu_{i,m}(v)-\mu_{i,m}(v'))^2}{2} \geq d_{\mathrm{KL}}\left(\mathbb{P}_v^\pi(E^*), \mathbb{P}_{v'}^\pi(E^*)\right). \tag{27}$$

Because π is δ-PAC, we have $\mathbb{P}_v^\pi(E^*) \geq 1-\delta$ and $\mathbb{P}_{v'}^\pi(E^*) \leq \delta$. This in turn implies that

$$d_{\mathrm{KL}}\left(\mathbb{P}_v^\pi(E^*), \mathbb{P}_{v'}^\pi(E^*)\right) \geq \log\left(\frac{1}{4\delta}\right). \tag{28}$$

Combining (27) and (28), we obtain

$$\sum_{m=1}^M \sum_{i=1}^K \mathbb{E}_v^\pi\left[N_{i,\tau_\delta}\right] \frac{(\mu_{i,m}(v)-\mu_{i,m}(v'))^2}{2} \geq \log\left(\frac{1}{4\delta}\right), \tag{29}$$

Letting $\bar{\omega}_i \coloneqq \frac{\mathbb{E}_v^\pi[N_{i,\tau_\delta}]}{\mathbb{E}_v^\pi[\tau_\delta]}$, we have

$$\mathbb{E}_v^\pi[\tau_\delta] \geq \frac{\log(\frac{1}{4\delta})}{\sum_{m=1}^M\ \sum_{i=1}^K \bar{\omega}_i \frac{(\mu_{i,m}(v)-\mu_{i,m}(v'))^2}{2}}. \tag{30}$$

Noting that (30) holds for any $v' \in \mathrm{Alt}(v)$, we have

$$\mathbb{E}_v^\pi[\tau_\delta] \geq \frac{\log(\frac{1}{4\delta})}{\inf_{v'\in\mathrm{Alt}(v)}\sum_{m=1}^M\ \sum_{i=1}^K \bar{\omega}_i \frac{(\mu_{i,m}(v)-\mu_{i,m}(v'))^2}{2}}. \tag{31}$$

Finally, using the fact that $\tilde{\omega} = [\tilde{\omega}_i : i \in [K]]^\top \in \Gamma$, we get

$$\mathbb{E}_v^\pi[\tau_\delta] \geq \frac{\log(\frac{1}{4\delta})}{\sup_{\omega\in\Gamma}\inf_{v'\in\mathrm{Alt}(v)}\sum_{m=1}^M\ \sum_{i=1}^K \omega_i \frac{(\mu_{i,m}(v)-\mu_{i,m}(v'))^2}{2}}. \tag{32}$$

This completes the proof. □

Lemma E.3. *Fix $v \in \mathcal{P}$ with arm means $\{\mu_{i,m} : i \in [K], m \in [M]\}$. For any $\omega \in \Gamma$,*

$$g_v(\omega) = \inf_{v'\in\mathrm{Alt}(v)} \sum_{i=1}^K \omega_i \sum_{m=1}^M \frac{\left(\mu_{i,m} - \mu'_{i,m}\right)^2}{2}, \tag{33}$$

where $\mu'_{i,m}$ denotes the mean of arm i corresponding to objective m under the instance v', and $g_v(\cdot)$ is defined in (7).

Proof. Note that

$$
\begin{aligned}
&\inf_{v' \in \mathrm{Alt}(v)} \sum_{m=1}^{M} \sum_{i=1}^{K} \omega_i \frac{\left(\mu_{i,m} - \mu'_{i,m}\right)^2}{2} \\
&= \min_{m \in [M]} \min_{i \in [K] \setminus i^*_m(v)} \inf_{v' : \mu'_{i,m} > \mu'_{i^*_m(v),m}} \sum_{\jmath=1}^{M} \sum_{\imath=1}^{K} \omega_\imath \frac{(\mu_{\imath,\jmath} - \mu'_{\imath,\jmath})^2}{2} \\
&= \min_{m \in [M]} \min_{i \in [K] \setminus i^*_m(v)} \inf_{v' : \mu'_{i,m} \geq \mu'_{i^*_m(v),m}} \left[\omega_i \frac{(\mu_{i,m} - \mu'_{i,m})^2}{2} + \omega_{i^*_m(v)} \frac{(\mu_{i^*_m(v),m} - \mu'_{i^*_m(v),m})^2}{2} \right] \qquad (34) \\
&= \min_{m \in [M]} \min_{i \in [K] \setminus i^*_m(v)} \frac{\Delta^2_{i,m}(v)}{2} \left(\frac{\omega_i \, \omega_{i^*_m(v)}}{\omega_i + \omega_{i^*_m(v)}} \right) \qquad (35) \\
&= g_v(\omega),
\end{aligned}
$$

where (35) follows by using the method of Lagrange multipliers and noting that the inner infimum in (34) is attained at

$$
\begin{aligned}
\mu'_{i,m} &= \mu_{i,m} + \left(\mu_{i^*_m(v),m} - \mu_{i,m}\right) \frac{\omega_{i^*_m(v)}}{\omega_i + \omega_{i^*_m(v)}}, \\
\mu'_{i^*_m(v),m} &= \mu_{i^*_m(v),m} - \left(\mu_{i^*_m(v),m} - \mu_{i,m}\right) \frac{\omega_i}{\omega_i + \omega_{i^*_m(v)}}. \qquad (36)
\end{aligned}
$$

This completes the proof. □

Proof of Proposition 3.1. Finally, by combining Lemmas E.2 and E.3, we see that Theorem 3.1 holds. □

F Proof of Proposition 4.6

Below, we first record some useful results that will be used in the proof.

Lemma F.1. *(Lattimore and Szepesvári, 2020, Lemma 33.8) Let $Y_1, Y_2, \ldots$ be independent Gaussian random variables with mean μ and unit variance. Let $\widehat{\mu}_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} Y_i$. Then,*

$$
\mathbb{P}\left(\exists\, n \in \mathbb{N} : \frac{n}{2} (\widehat{\mu}_n - \mu)^2 \geq \log(1/\delta) + \log(n(n+1)) \right) \leq \delta.
$$

Lemma F.2. *(Chen et al., 2023, Lemma A.4) Fix $n \in \mathbb{N}$. Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with $\mathbb{P}(Y_i \leq y) \leq y$ for all $y \in [0,1]$ and $i \in [n]$. Then, for any $\epsilon > 0$,*

$$
\mathbb{P}\left(\sum_{i=1}^{n} \log(1/Y_i) \geq \epsilon \right) \leq f^{(n)}(\epsilon) \qquad (37)
$$

where $f^{(n)} : (0, +\infty) \to (0, 1)$ is defined by

$$
f^{(n)}(x) = \sum_{i=1}^{n} \frac{x^{i-1} e^{-x}}{(i-1)!}, \quad x \in (0, +\infty).
$$

Proof of Proposition 4.6. Fix any confidence level $\delta \in (0,1)$, problem instance $v \in \mathcal{P}$, and $\eta > 0$. Consider MO-BAI. We claim that $\tau_\delta < +\infty$ almost surely; a proof of this is deferred until the proof of Lemma G.12. For $m \in [M]$ and $i \in [K]$, let

$$
\xi_{i,m} \coloneqq \sup_{t \geq K} \frac{N_{i,t}}{2} \left(\widehat{\mu}_{i,m}(t) - \mu_{i,m}(v) \right)^2 - \log\left(N_{i,t}(N_{i,t} + 1) \right). \qquad (38)
$$

From Lemma F.1, we know that for any confidence level $\delta' \in (0,1)$,

$$\mathbb{P}_v^{\text{MO-BAI}}(\xi_{i,m} \geq \log(1/\delta')) \leq \delta'. \tag{39}$$

Let $\xi'_{i,m} \coloneqq \exp(-\xi_{i,m})$. From Lemma F.2, we know that for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}_v^{\text{MO-BAI}}\left(\sum_{m\in[M]}\sum_{i\in[K]} \log(1/\xi'_{i,m}) \geq \epsilon\right) \leq f^{(MK)}(\epsilon) \\
&\overset{(a)}{\Longrightarrow} \mathbb{P}_v^{\text{MO-BAI}}\left(\sum_{m\in[M]}\sum_{i\in[K]} \xi_{i,m} \geq \epsilon\right) \leq f^{(MK)}(\epsilon) \\
&\overset{(b)}{\Longrightarrow} \mathbb{P}_v^{\text{MO-BAI}}\left(\sum_{m\in[M]}\sum_{i\in[K]} \xi_{i,m} \geq \epsilon\right) \leq f(\epsilon) \\
&\overset{(c)}{\Longrightarrow} \mathbb{P}_v^{\text{MO-BAI}}\left(\sum_{m\in[M]}\sum_{i\in[K]} \xi_{i,m} \geq f^{-1}(\delta)\right) \leq \delta,
\end{aligned} \tag{40}$$

definition of $\xi'_{i,m}$, (b) follows from the definition of f in (16), and in writing (c), we (i) make use of the fact that f is continuous and strictly decreasing and therefore admits an inverse, and (ii) set $\epsilon = f^{-1}(\delta)$. Plugging the expression for $\xi_{i,m}$ from (38) in (40), and noting that $N_{i,t} \leq t$ for all $i \in [K]$ and $t \geq K$, we get

$$\begin{aligned}
&\mathbb{P}_v^{\text{MO-BAI}}\left(\forall\, t \geq K \sum_{m\in[M]}\sum_{i\in[K]} \frac{N_{i,t}}{2}\big(\widehat{\mu}_{i,m}(t) - \mu_{i,m}(v)\big)^2 \leq MK\log\big(t(t+1)\big) + f^{-1}(\delta)\right) \geq 1-\delta \\
\Longrightarrow\ &\mathbb{P}_v^{\text{MO-BAI}}\left(\forall\, t \geq K \sum_{m\in[M]}\sum_{i\in[K]} \frac{N_{i,t}}{2}\big(\widehat{\mu}_{i,m}(t) - \mu_{i,m}(v)\big)^2 \leq \beta(t,\delta)\right) \geq 1-\delta.
\end{aligned} \tag{41}$$

Note that at the stopping time $\tau_\delta = \tau_\delta$, by definition, we must have

$$Z(\tau_\delta) = \inf_{v'\in \mathrm{Alt}(\widehat{v}(\tau_\delta))} \sum_{m\in[M]}\sum_{i\in[K]} N_{i,\tau_\delta}\frac{(\mu_{i,m}(v') - \widehat{\mu}_{i,m}(\tau_\delta))^2}{2} > \beta(\tau_\delta,\delta).$$

Thus, (41) may be expressed equivalently as

$$\mathbb{P}_v^{\text{MO-BAI}}\Big(v \notin \mathrm{Alt}\big(\widehat{v}(\tau_\delta)\big)\Big) \geq 1-\delta \quad \Longleftrightarrow \quad \mathbb{P}_v^{\text{MO-BAI}}\Big(I^*(v) = I^*\big(\widehat{v}(\tau_\delta)\big)\Big) \geq 1-\delta,$$

thereby establishing the desired result. This completes the proof. □

G Proof of Theorem 4.7

Let $v \in \mathcal{P}$ be fixed. Firstly, we define the curvature (Jaggi, 2013) of a concave function.

Definition G.1 (Curvature)**.** Given a concave function $f : \mathcal{D} \to \mathbb{R}$ defined on a convex domain $\mathcal{D}$, the *curvature* of f is defined as

$$C_{\text{cur}}(f) \coloneqq \sup_{\substack{\omega,\mathbf{y}\in\mathcal{D},\\ \gamma\in(0,1),\\ \mathbf{d}\in\partial f(\omega)}} \frac{2}{\gamma^2}\Big(f(\omega) + \langle \mathbf{z}-\omega, \mathbf{d}\rangle - f\big(\mathbf{z}\big)\Big), \quad \mathbf{z} = \omega + \gamma(\mathbf{y}-\omega), \tag{42}$$

where $\partial f(\omega)$ denotes the super-differential of f at ω.

In the following, we will show that $g_v^{(i,m)}(\cdot)\big|_{\Gamma^{(\eta)}}$ (the function $g_v^{(i,m)}(\cdot)$ restricted to the set $\Gamma^{(\eta)}$) is a concave function. It is worth noting that the curvature of g_v is a function of its super-gradients, whereas the constant

$C(v,\eta)$ is a function of the gradient of $g_v^{(i,m)}(\cdot)$ for $i \neq i_m^*(v)$ (see (9)). While $C_{\mathrm{cur}}(g_v)$ may be infinitely large, we will show that $C(v,\eta)$ is finite for all $\eta > 0$. The latter, as we shall see, is an important property that will enable us to demonstrate the asymptotic optimality of our policy.

Before we proceed further, we record some useful results.

Lemma G.2. *(Adapted from Jaggi (2013, Lemma 7)) Let $f : \mathcal{D} \to \mathbb{R}$ be a concave and differentiable function defined over a convex domain $\mathcal{D}$, with a gradient function ∇f is Lipschitz-continuous w.r.t. some norm $\|.\|$ over the domain $\mathcal{D}$ with a Lipschitz constant $L > 0$. Then,*

$$C_{\mathrm{cur}}(f) \leq \mathrm{diam}_{\|\cdot\|}(\mathcal{D})^2 L,$$

where $\mathrm{diam}_{\|\cdot\|}(\mathcal{D}) \coloneqq \sup_{\mathbf{x},\mathbf{y}\in\mathcal{D}} \|\mathbf{x}-\mathbf{y}\|$.

Lemma G.3. *(Chen et al. (2023, Lemma A.7)) Given $n \in \mathbb{N}$, let $f^{(n)}(x) = \sum_{i=1}^{n} \frac{x^{i-1}e^{-x}}{(i-1)!}$ for $x \in (0,+\infty)$. Then, $(f^{(n)})^{-1}(\delta) = (1+o(1))\log(1/\delta)$ as $\delta \downarrow 0$, i.e.,*

$$\lim_{\delta\downarrow 0} \frac{(f^{(n)})^{-1}(\delta)}{\log(1/\delta)} = 1. \tag{43}$$

The below result demonstrates the concavity of the function $g_v^{(i,m)}(\cdot)\big|_{\Gamma^{(\eta)}}$ for all (i,m) with $i \neq i_m^*(v)$.

Lemma G.4. *Fix $v \in \mathbb{P}$, $\eta > 0$, $m \in [M]$, and $i \in [K]$ such that $i \neq i_m^*(v)$. Then, $g_v^{(i,m)}(\cdot)\big|_{\Gamma^{(\eta)}}$ is a concave function.*

Proof of Lemma G.4. It can be easily shown that the eigenvalues of the Hessian matrix of $g_v^{(i,m)}(\cdot)$ are

- $\dfrac{-\Delta_{i,m}^2(\omega_i^2 + \omega_{i_m^*(v)}^2)}{(\omega_i + \omega_{i_m^*(v)})^3}$ with algebraic multiplicity 1, and
- 0 with algebraic multiplicity $K-1$.

Hence, the Hessian matrix of $g_v^{(i,m)}(\cdot)$ is negative semi-definite matrix for all $\omega \in \Gamma^{(\eta)}$, thus proving the desired result. □

With the above ingredients in place, we now prove that $C(v,\eta)$ is finite for all $\eta > 0$.

Lemma G.5. *Fix $v \in \mathcal{P}$. For any $\eta > 0$,*

$$C(v,\eta) \leq \max_{i\neq i_m^*(v)} \frac{2\,\Delta_{i,m}^2(v)\,(1+\eta)\,K}{\eta}. \tag{44}$$

Proof of Lemma G.5. In this proof, we will first prove that the curvature of $g_v^{(i,m)}(\cdot)\big|_{\Gamma^{(\eta)}}$ is finite for all $i \in [K]$ and $m \in [M]$. Subsequently, we will show that $C(v,\eta)$ is no greater than the maximum curvature of $g_v^{(i,m)}(\cdot)\big|_{\Gamma^{(\eta)}}$ computed over all (i,m) pairs.

Note that for any $i \in [K]$, $m \in [M]$, $\imath \in [K]$ and $\jmath \in [K]$, we have $\forall \omega \in \Gamma^{(\eta)}$

$$\frac{\partial}{\partial \omega_\imath} g^{(i,m)}(\omega) = \begin{cases} \dfrac{\Delta_{i,m}^2(v)\,\omega_{i_m^*(v)}^2}{2(\omega_{i_m^*(v)} + \omega_i)^2}, & \text{if } \imath = i, \\[2ex] \dfrac{\Delta_{i,m}^2(v)\,\omega_i^2}{2(\omega_{i_m^*(v)} + \omega_i)^2}, & \text{if } \imath = i_m^*(v), \\[2ex] 0, & \text{otherwise,} \end{cases} \tag{45}$$

and

$$\frac{\partial^2}{\partial\omega_\imath\partial\omega_\jmath} g^{(i,m)}(\omega) = \begin{cases} -\dfrac{\Delta_{i,m}^2(v)\,\omega_{i_m^*(v)}^2}{(\omega_{i_m^*(v)}+\omega_i)^3}, & \text{if } \imath=\jmath=i, \\[2ex] -\dfrac{\Delta_{i,m}^2(v)\,\omega_i^2}{(\omega_{i_m^*(v)}+\omega_i)^3}, & \text{if } \imath=\jmath=i_m^*(v), \\[2ex] \dfrac{\Delta_{i,m}^2(v)\,\omega_i\omega_{i_m^*(v)}}{(\omega_{i_m^*(v)}+\omega_i)^3}, & \text{if } \imath=i, \jmath=i_m^*(v) \text{ or } \imath=i_m^*(v), \jmath=i \\ 0, & \text{otherwise.} \end{cases} \tag{46}$$

Recall that $\omega\in\Gamma^{(\eta)}$ implies $\frac{\eta}{K(1+\eta)}\le\omega_i\le 1$ for all $i\in[K]$. Using this fact along with (46), we get that

$$\sup_{\omega\in\Gamma^{(\eta)}} \frac{\partial^2}{\partial\omega_\imath\partial\omega_\jmath} g^{(i,m)}(\omega) < +\infty \quad \forall i,\imath,\jmath\in[K],\ m\in[M].$$

In addition, (45) implies that the function $\nabla g^{(i,m)}:\Gamma^{(\eta)}\to\mathbb{R}^K$ is continuous. Combining the above facts, we see that $\nabla g^{(i,m)}:\Gamma^{(\eta)}\to\mathbb{R}$ is Lipschitz-continuous w.r.t. the norm $\|\cdot\|_1$ with a finite Lipschitz constant L satisfying

$$\begin{aligned} L &\le \sup_{\omega\in\Gamma^{(\eta)},\mathbf{d}\in\mathbb{R}^K} \frac{\|\mathbf{H}_{g^{(i,m)}}(\omega)\,\mathbf{d}\|_1}{\|\mathbf{d}\|_1} \\ &= \sup_{\omega\in\Gamma^{(\eta)},\ \jmath\in[K]} \|\mathbf{H}_{g^{(i,m)}}^{(\cdot,\jmath)}(\omega)\|_1 \\ &\overset{(a)}{\le} \sup_{\omega\in\Gamma^{(\eta)},\ \imath\in[K],\ \jmath\in[K]} 2|\mathbf{H}_{g^{(i,m)}}^{(\imath,\jmath)}(\omega)| \\ &\overset{(b)}{\le} \frac{2\Delta_{i,m}^2(v)(1+\eta)K}{\eta}, \end{aligned} \tag{47}$$

where $\mathbf{H}_{g^{(i,m)}}$ denotes the Hessian matrix of $g^{(i,m)}$ and $\mathbf{H}_{g^{(i,m)}}^{(\cdot,\jmath)}$ (resp. $\mathbf{H}_{g^{(i,m)}}^{(\imath,\jmath)}$) denotes its $\jmath$-th column (resp. its $(\imath,\jmath)$)-th element), (a) above follows the fact that the Hessian matrix of $g^{(i,m)}(\cdot)\big|_{\Gamma^{(\eta)}}$ has at most two non-zero elements in each column, a fact that in turn follows from (46), and (b) above follows the fact that

$$\frac{\partial^2}{\partial\omega_\imath\partial\omega_\jmath} g^{(i,m)}(\omega) \le \frac{\Delta_{i,m}^2(v)(1+\eta)K}{\eta} \quad \forall\omega\in\Gamma^{(\eta)}.$$

Using Lemma G.2 along with the fact that $\operatorname{diam}_{\|\cdot\|_1}(\Gamma^{(\eta)})\le 1$, we have for any $i\in[K]$ and $m\in[M]$

$$C_{\mathrm{cur}}\left(g_v^{(i,m)}\big|_{\Gamma^{(\eta)}}\right) \le \frac{2\,\Delta_{i,m}^2(v)\,(1+\eta)\,K}{\eta}. \tag{48}$$

In addition, from (21) and (9), we note by setting $\mathbf{z}=\omega+\gamma(\mathbf{y}-\omega)$ that

$$C(v,\eta) = \sup_{\substack{\omega,\mathbf{y}\in\Gamma^{(\eta)},\\ \gamma\in(0,1)}} \frac{2}{\gamma^2}\left(\min_{(i,m):i\ne i_m^*(v)} g_v^{(i,m)}(\omega) + \langle\nabla_\omega g_v^{(i,m)}(\omega), \mathbf{z}-\omega\rangle - g_v(\mathbf{z})\right), \tag{49}$$

Note that for any $\mathbf{z}\in\Gamma^{(\eta)}$, there exist $i'\in[K]$ and $m'\in[M]$ with $i'\ne i_{m'}^*(v)$ such that $g_v(\mathbf{z})=g_v^{(i',m')}(\mathbf{z})$. Therefore,

$$\begin{aligned} &\min_{(i,m):i\ne i_m^*(v)} g_v^{(i,m)}(\omega) + \langle\nabla_\omega g_v^{(i,m)}(\omega), \mathbf{z}-\omega\rangle - g_v(\mathbf{z}) \\ &= \min_{(i,m):i\ne i_m^*(v)} g_v^{(i,m)}(\omega) + \langle\nabla_\omega g_v^{(i,m)}(\omega), \mathbf{z}-\omega\rangle - g_v^{(i',m')}(\mathbf{z}) \\ &\le g_v^{(i',m')}(\omega) + \langle\nabla_\omega g_v^{(i',m')}(\omega), \mathbf{z}-\omega\rangle - g_v^{(i',m')}(\mathbf{z}) \end{aligned}$$

$$\leq \max_{(i,m):i\neq i^*_m(v)} \left[g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega), \mathbf{z}-\omega\rangle - g_v^{(i,m)}(\mathbf{z})\right]. \tag{50}$$

Combining (49) and (50), and setting $\mathbf{z} = \omega + \gamma(\mathbf{y}-\omega)$ throughout, we get that

$$\begin{aligned} C(v,\eta) &\leq \sup_{\substack{\omega,\mathbf{y}\in\Gamma^{(\eta)},\\ \gamma\in(0,1)}} \max_{(i,m):i\neq i^*_m(v)} \left[g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega), \mathbf{z}-\omega\rangle - g_v^{(i,m)}(\mathbf{z})\right], \\ &= \max_{(i,m):i\neq i^*_m(v)} \sup_{\substack{\omega,\mathbf{y}\in\Gamma^{(\eta)},\\ \gamma\in(0,1)}} \left[g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega), \mathbf{z}-\omega\rangle - g_v^{(i,m)}(\mathbf{z})\right], \\ &\overset{(a)}{=} \max_{(i,m):i\neq i^*_m(v)} C_{\text{cur}}\left(g_v^{(i,m)}\big|_{\Gamma^{(\eta)}}\right), \end{aligned} \tag{51}$$

where (a) follows from the definition in (42). Finally, combining (48) and (51), we arrive at the desired result. □

Lemma G.6. *Fix $\eta > 0$. Consider the sequence $\{v_{l_t} : t \geq K\}$ generated by* MO-BAI*. Then,*

$$\limsup_{t\to\infty} C(\widehat{v}_{l_t}, \eta) < +\infty \quad \textit{almost surely.}$$

Proof of Lemma G.6. Fix $v \in \mathcal{P}$. By the strong law of large numbers, we have for all $i \in [K]$ and $m \in [M]$

$$\lim_{t\to+\infty} \Delta_{i,m}(\widehat{v}_{l_t}) = \Delta_{i,m}(v) \quad \text{almost surely}$$

under the instance v, which by Lemma G.5 implies that

$$\limsup_{t\to\infty} C(\widehat{v}_{l_t}, \eta) < +\infty \quad \text{almost surely.}$$

This completes the proof. □

Lemma G.7. *Fix $v \in \mathcal{P}$ and $\eta > 0$. Under* MO-BAI*, we have*

$$\lim_{t\to+\infty} \max_{\omega\in\Gamma^{(\eta)}} |g_v(\omega) - g_{\widehat{v}_{l_t}}(\omega)| = 0 \quad \textit{almost surely.}$$

Proof of Lemma G.7. By the strong law of large numbers and the fact that under v the best arm of each objective is unique, we have for all $m \in [M]$

$$\lim_{t\to+\infty} i^*_m(\widehat{v}_{l_t}) = i^*_m(v) \quad \text{almost surely,} \tag{52}$$

and for all $i \in [K]$ and $m \in [M]$

$$\lim_{t\to+\infty} \Delta_{i,m}(\widehat{v}_{l_t}) = \Delta_{i,m}(v) \quad \text{almost surely.} \tag{53}$$

For $m \in [M]$, let T_m be the smallest integer such that

$$i^*_m(\widehat{v}_{l_t}) = i^*(v) \quad \forall t \geq T_m. \tag{54}$$

Then, (52) implies that $T_m < +\infty$ almost surely. In addition, by the definition of $g_v^{(i,m)}(\cdot)$ in (8), we have $\forall t \geq T_m$ and $\omega \in \Gamma^{(\eta)}$ that

$$\begin{aligned} |g_v^{(i,m)}(\omega) - g_{\widehat{v}_{l_t}}^{(i,m)}(\omega)| &= \frac{\omega_i\, \omega_{i^*_m(v)}}{\omega_i + \omega_{i^*_m(v)}} |\Delta_{i,m}(\widehat{v}_{l_t}) - \Delta_{i,m}(v)| \\ &\leq |\Delta_{i,m}(\widehat{v}_{l_t}) - \Delta_{i,m}(v)|, \end{aligned} \tag{55}$$

which implies that $\forall t \geq \max_{m\in[M]} T_m$,

$$|g_v(\omega) - g_{\widehat{v}_{l_t}}(\omega)| \leq \max_{(i,m)\in[K]\times[M]} |\Delta_{i,m}(\widehat{v}_{l_t}) - \Delta_{i,m}(v)|. \tag{56}$$

Notice that the right-hand side of (56) does not depend on ω. Hence, combing (53) and (56) along with the almost sure finiteness of T_m for each $m \in [M]$, we arrive at the desired result. □

Next, we show that the empirical proportion of arms pulls under MO-BAI converges to the oracle weight in the long run.

Lemma G.8. *Fix $v \in \mathcal{P}$ and $\eta > 0$. For all $t_1, t_2 \in \mathbb{N}$ with $t_2 > t_1 > K$, under* MO-BAI*, we have*

$$|\tilde{c}(v,\eta)^{-1} - g_v(\widehat{\omega}_{\cdot,t_2})| \leq \frac{t_1}{t_2}\,\tilde{c}(v,\eta)^{-1} + 11\,\epsilon_{t_1}(v) + \frac{2\,\log(t_2)\,\overline{C}_{t_1}(\eta)}{t_2}, \tag{57}$$

where for any time step t:

- *The quantity $\epsilon_t(v)$ is defined as*

$$\epsilon_t(v) \coloneqq \sup_{t' \geq l_t}\left(\sup_{\omega \in \Gamma^{(\eta)}} \left|g_v(\omega) - g_{\widehat{v}_{t'}}(\omega)\right|\right). \tag{58}$$

- *The quantity $\overline{C}_t(\eta)$ is defined as*

$$\overline{C}_t(\eta) \coloneqq \sup_{t' \geq l_t} C(\widehat{v}_{t'}, \eta). \tag{59}$$

Consequently, letting $t_1 = \lceil\sqrt{t_2}\rceil$, we have

$$\tilde{c}(v,\eta)^{-1} - g_v(\widehat{\omega}_{\cdot,t_2}) \leq \frac{2}{\sqrt{t_2}}\tilde{c}(v,\eta)^{-1} + 11\epsilon_{\lceil\sqrt{t_2}\rceil}(v) + \frac{2\,\log(t_2)\,\overline{C}_{\lceil\sqrt{t_2}\rceil}(\eta)}{t_2}. \tag{60}$$

Proof of Lemma G.8. Fix t_1, t_2 with $K < t_1 < t_2$.

Step 1: Bound on the empirical instance $\widehat{v}_{l_t}$.

Fix $t \in \{t_1+1, \ldots, t_2\}$. Let $\widehat{v}_{l_t} \in \mathcal{P}$ denote the instance in which the mean of arm i corresponding to objective m is given by $\widehat{\mu}_{i,m}(l_t)$. Making the substitutions $v \leftarrow \widehat{v}_{l_t}$, $\omega \leftarrow \widehat{\omega}_{\cdot,t-1}$, and $\mathbf{s} \leftarrow \widehat{\omega}_{\cdot,t}$ in (9), we may deduce the following: there exists $(i',m') \in [K]\times[M]$, $i' \neq i^*_{m'}(\widehat{v}_{l_t})$, such that

$$g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) + \langle \nabla_\omega\, g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}),\ \widehat{\omega}_{\cdot,t} - \widehat{\omega}_{\cdot,t-1}\rangle = h_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}, \widehat{\omega}_{\cdot,t}). \tag{61}$$

Recall that $\mathbf{s}_t = \arg\max_{\mathbf{s}\in\Gamma^{(\eta)}} h_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}, \mathbf{s})$. We then have from (9) that

$$\begin{aligned}
g^{(i',m')}_{\widehat{v}_{l_t}}&(\widehat{\omega}_{\cdot,t-1}) + \langle \nabla_\omega\, g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}),\ \mathbf{s}_t - \widehat{\omega}_{\cdot,t-1}\rangle \\
&\geq h_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}, \mathbf{s}_t) \\
&\overset{(a)}{\geq} \tilde{c}(\widehat{v}_{l_t}, \eta)^{-1},
\end{aligned} \tag{62}$$

where (a) follows from Lemma G.4. Let

$$U_t \coloneqq g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) + \langle \nabla_\omega\, g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}),\ \mathbf{s}_t - \widehat{\omega}_{\cdot,t-1}\rangle - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}), \tag{63}$$

From (62), we have

$$U_t \geq \tilde{c}(\widehat{v}_{l_t}, \eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}). \tag{64}$$

We then note that

$$\begin{aligned}
&h_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}, \widehat{\omega}_{\cdot,t}) - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) \\
&\overset{(a)}{=} g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) + \langle \nabla_\omega\, g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}),\ \widehat{\omega}_{\cdot,t} - \widehat{\omega}_{\cdot,t-1}\rangle - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) \\
&\overset{(b)}{=} \langle \nabla_\omega g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}),\ \frac{1}{t}(\mathbf{s}_t - \widehat{\omega}_{\cdot,t-1})\rangle + g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) \\
&\geq \frac{1}{t}\Big(\langle \nabla_\omega g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}),\ \mathbf{s}_t - \widehat{\omega}_{\cdot,t-1}\rangle + g^{(i',m')}_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1})\Big)
\end{aligned}$$

$$\begin{aligned}
&= \frac{U_t}{t} \\
&\geq \frac{1}{t}\left(\tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1})\right),
\end{aligned} \tag{65}$$

where (a) above follows from (61), and (b) above follows from the construction of $\widehat{\omega}_{\cdot,t}$ (see Algorithm 1).

Step 2: Bound on the instance v.

For $t \in \{t_1+1,\ldots,t_2\}$, by the definition of $C(\cdot)$ in (21), we have

$$g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t}) \geq h_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1},\widehat{\omega}_{\cdot,t}) - \frac{2\,\overline{C}_t(\eta)}{t^2}. \tag{66}$$

Then, combining (65) and (66), we get that for all $t \in \{t_1+1,\ldots,t_2\}$,

$$g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t}) \geq g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) + \left(\frac{1}{t}\left(\tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1})\right) - \frac{2\,\overline{C}_t(\eta)}{t^2}\right). \tag{67}$$

It then follows from (67) that

$$\begin{aligned}
&\tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t}) \\
&\leq \tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) - \left(\frac{1}{t}\left(\tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1})\right) - \frac{2\,\overline{C}_t(\eta)}{t^2}\right) \\
&= \frac{t-1}{t}\left(\tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1})\right) + \frac{2\,\overline{C}_t(\eta)}{t^2}.
\end{aligned} \tag{68}$$

Let

$$G_t \coloneqq \tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t}). \tag{69}$$

Observe that $l_{t-1} = l_t$ if and only if $t \notin \{2^i \mid i \in \mathbb{N}\}$. Using the definition of ϵ_t from the statement of the lemma, and combining (68) and (69), we get that for all $t \in \{t_1+1,\ldots,t_2\}$,

$$G_t \leq \begin{cases} \dfrac{t-1}{t}\,G_{t-1} + \dfrac{2\,\overline{C}_t(\eta)}{t^2} + 4\,\epsilon_{t_1}(v), & \text{if } t \in \{2^i \mid i \in \mathbb{N}\}, \\[2ex] \dfrac{t-1}{t}\,G_{t-1} + \dfrac{2\,\overline{C}_t(\eta)}{t^2}, & \text{if } t \notin \{2^i \mid i \in \mathbb{N}\}. \end{cases} \tag{70}$$

The first line above follows by noting that for $t \in \{2^i \mid i \in \mathbb{N}\}$,

$$\begin{aligned}
&\tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) - G_{t-1} \\
&= \tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) - \tilde{c}(\widehat{v}_{l_{t-1}},\eta)^{-1} + g_{\widehat{v}_{l_{t-1}}}(\widehat{\omega}_{\cdot,t-1}) \\
&= \tilde{c}(\widehat{v}_{l_t},\eta)^{-1} - \tilde{c}(\widehat{v}_{l_{t-1}},\eta)^{-1} + g_{\widehat{v}_{l_{t-1}}}(\widehat{\omega}_{\cdot,t-1}) - g_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}) \\
&\leq \sup_{\omega\in\Gamma^{(\eta)}} g_{\widehat{v}_{l_t}}(\omega) - \sup_{\omega\in\Gamma^{(\eta)}} g_{\widehat{v}_{l_{t-1}}}(\omega) + g_{\widehat{v}_{l_{t-1}}}(\omega) - g_{\widehat{v}_{l_t}}(\omega) \\
&\leq \sup_{\omega\in\Gamma^{(\eta)}} \left[g_{\widehat{v}_{l_t}}(\omega) - g_{\widehat{v}_{l_{t-1}}}(\omega)\right] + |g_{\widehat{v}_{l_{t-1}}}(\omega) - g_{\widehat{v}_{l_t}}(\omega)| \\
&\leq 2\sup_{\omega\in\Gamma^{(\eta)}} |g_{\widehat{v}_{l_{t-1}}}(\omega) - g_{\widehat{v}_{l_t}}(\omega)| \\
&\leq 2\left[\sup_{\omega\in\Gamma^{(\eta)}} |g_v(\omega) - g_{\widehat{v}_{l_t}}(\omega)| + \sup_{\omega\in\Gamma^{(\eta)}} |g_v(\omega) - g_{\widehat{v}_{l_{t-1}}}(\omega)|\right] \\
&\leq 2\left[\sup_{t'\geq l_t}\ \sup_{\omega\in\Gamma^{(\eta)}} |g_v(\omega) - g_{\widehat{v}_{t'}}(\omega)| + \sup_{t'\geq l_{t-1}}\ \sup_{\omega\in\Gamma^{(\eta)}} |g_v(\omega) - g_{\widehat{v}_{t'}}(\omega)|\right] \\
&\leq 4\,\epsilon_{t-1}(v) \\
&\leq 4\,\epsilon_{t_1}(v).
\end{aligned} \tag{71}$$

The penultimate step above follows by noting that $l_{t-1} < l_t$ for $t \in \{2^i \mid i \in \mathbb{N}\}$, and the last step above follows by using the fact that $\epsilon_{t-1} \leq \epsilon_{t_1}$ for all $t \in \{t_1+1, \ldots, t_2\}$. Using the mathematical induction formula given in the following Lemma G.9, it follows from (70) that for all $K < t_1 < t_2$,

$$G_{t_2} \leq \frac{t_1}{t_2}\, G_{t_1} + 4\,\epsilon_{t_1}(v) \left(\sum_{t=t_1}^{t_2} \frac{t}{t_2}\, \mathbf{1}_{\{t \in \{2^i | i \in \mathbb{N}\}\}}\right) + \frac{2\,\overline{C}_{t_1}(\eta)}{t_2} \left(\sum_{j=t_1}^{t_2} \frac{1}{j}\right). \tag{72}$$

Note that

$$\sum_{t=t_1}^{t_2} \frac{t}{t_2}\, \mathbf{1}_{\{t \in \{2^i | i \in \mathbb{N}\}\}} \leq \sum_{i=0}^{\infty} 2^{-i} = 2. \tag{73}$$

In addition, we have

$$\begin{aligned}\sum_{j=t_1}^{t_2} \frac{1}{j} &\overset{(a)}{\leq} \sum_{j=2}^{t_2} \frac{1}{j} \\ &\leq \int_1^{t_2} \frac{1}{x}\, \mathrm{d}x \qquad (74)\\ &= \log(t_2), \qquad (75)\end{aligned}$$

where (a) above follows from the fact that $t_1 \geq K \geq 2$. Combining (72), (73), and (75), we have

$$\begin{aligned}G_{t_2} &\leq \frac{t_1}{t_2}\, G_{t_1} + 8\,\epsilon_{t_1}(v) + \frac{2\,\log(t_2)\,\overline{C}_{t_1}(\eta)}{t_2} \\ &\overset{(a)}{\leq} \frac{t_1}{t_2}\, \tilde{c}(\widehat{v}_{l_{t_1}}, \eta)^{-1} + 8\,\epsilon_{t_1}(v) + \frac{2\,\log(t_2)\,\overline{C}_{t_1}(\eta)}{t_2} \\ &\overset{(b)}{\leq} \frac{t_1}{t_2}\, \left(\tilde{c}(v, \eta)^{-1} + \epsilon_{t_1}(v)\right) + 8\,\epsilon_{t_1}(v) + \frac{2\,\log(t_2)\,\overline{C}_{t_1}(\eta)}{t_2}, \qquad (76)\end{aligned}$$

where (a) above follows by noting that $G_{t_1} = \tilde{c}(\widehat{v}_{l_{t_1}}, \eta)^{-1} - g_{\widehat{v}_{l_{t_1}}}(\widehat{\omega}_{\cdot,t_1}) \leq \tilde{c}(\widehat{v}_{l_{t_1}}, \eta)^{-1}$, and (b) above follows from (58). Additionally, we note that (58) implies

$$\begin{aligned}\tilde{c}(v, \eta)^{-1} - g_v(\widehat{\omega}_{\cdot,t_2}) &= \tilde{c}(v, \eta)^{-1} - g_v(\widehat{\omega}_{\cdot,t_2}) - G_{t_2} + G_{t_2} \\ &= \tilde{c}(v, \eta)^{-1} - g_v(\widehat{\omega}_{\cdot,t_2}) - \left(\tilde{c}(\widehat{v}_{l_{t_2}}, \eta)^{-1} - g_{\widehat{v}_{l_{t_2}}}(\widehat{\omega}_{\cdot,t_2})\right) + G_{t_2} \\ &= \tilde{c}(v, \eta)^{-1} - \tilde{c}(\widehat{v}_{l_{t_2}}, \eta)^{-1} + g_{\widehat{v}_{l_{t_2}}}(\widehat{\omega}_{\cdot,t_2}) - g_v(\widehat{\omega}_{\cdot,t_2}) + G_{t_2} \\ &= \sup_{\omega \in \Gamma^{(\eta)}} g_v(\omega) - \sup_{\omega \in \Gamma^{(\eta)}} g_{\widehat{v}_{l_{t_2}}}(\omega) + \sup_{\omega \in \Gamma^{(\eta)}} |g_v(\omega) - g_{\widehat{v}_{l_{t_2}}}(\omega)| + G_{t_2} \\ &\leq 2 \sup_{\omega \in \Gamma^{(\eta)}} |g_v(\omega) - g_{\widehat{v}_{l_{t_2}}}(\omega)| + G_{t_2} \\ &\leq G_{t_2} + 2\,\epsilon_{t_2}(v) \\ &\leq G_{t_2} + 2\,\epsilon_{t_1}(v). \qquad (77)\end{aligned}$$

Finally, combining (76) and (77), we arrive at the desired result. □

Lemma G.9. *Fix $v \in \mathcal{P}$, $\eta > 0$, and $t_1, t_2 \in \mathbb{N}$ such that $t_1 < t_2$. Then, under* MO-BAI,

$$G_{t_2} \leq \frac{t_1}{t_2}\, G_{t_1} + 4\,\epsilon_{t_1}(v) \sum_{t=t_1}^{t_2} \frac{t}{t_2}\, \mathbf{1}_{\{t \in \{2^i | i \in \mathbb{N}\}\}} + \frac{2\,\overline{C}_{t_1}(\eta)}{t_2} \sum_{j=t_1}^{t_2} \frac{1}{j}, \tag{78}$$

where G_t, $\epsilon_t(v)$, and $\overline{C}_t(\eta)$ are as defined in (69), (59) *and* (58) *respectively for $t \in \mathbb{N}$.*

Proof of Lemma G.9. Using the principle of mathematical induction, we shall demonstrate that for any $t' \in \{t_1, \ldots, t_2\}$,

$$G_{t'} \leq \frac{t_1}{t'}\, G_{t_1} + 4\,\epsilon_{t_1}(v) \sum_{t=t_1}^{t'} \frac{t}{t'}\, \mathbf{1}_{\{t \in \{2^i | i \in \mathbb{N}\}\}} + \frac{2\,\overline{C}_{t_1}(\eta)}{t'} \sum_{j=t_1}^{t'} \frac{1}{j}. \tag{79}$$

Base case: For $t' = t_1$, we can immediately verify that (79) holds.

Induction Step: Suppose that (79) holds for $t' = s-1$ for some $s \in \{t_1+1, \ldots, t_2\}$. Then, (70) implies that

$$\begin{aligned}
G_s &\leq \frac{s-1}{s}\, G_{s-1} + \frac{2\,\overline{C}_s(\eta)}{s^2} + \mathbf{1}_{\{s\in\{2^i|i\in\mathbb{N}\}\}}\, 4\epsilon_s(v) \\
&\overset{(a)}{\leq} \frac{s-1}{s}\, G_{s-1} + \frac{2\,\overline{C}_s(\eta)}{s^2} + \mathbf{1}_{\{s\in\{2^i|i\in\mathbb{N}\}\}}\, 4\epsilon_{t_1}(v) \\
&\overset{(b)}{\leq} \frac{s-1}{s}\, \left(\frac{t_1}{s-1}\, G_{t_1} + 4\,\epsilon_{t_1}(v) \sum_{t=t_1}^{s-1} \frac{t}{s-1} \mathbf{1}_{\{t\in\{2^i|i\in\mathbb{N}\}\}} + \frac{2\,\overline{C}_{t_1}(\eta)}{s-1} \sum_{j=t_1}^{s-1} \frac{1}{j} \right) \\
&\qquad + \frac{2\,\overline{C}_s(\eta)}{s^2} + \mathbf{1}_{\{s\in\{2^i|i\in\mathbb{N}\}\}}\, 4\,\epsilon_{t_1}(v) \\
&= \frac{t_1}{s}\, G_{t_1} + 4\,\epsilon_{t_1}(v) \sum_{t=t_1}^{s} \frac{t}{s}\, \mathbf{1}_{\{t\in\{2^i|i\in\mathbb{N}\}\}} + \frac{2\,\overline{C}_{t_1}(\eta)}{s} \sum_{j=t_1}^{s} \frac{1}{j},
\end{aligned} \tag{80}$$

where (a) above follows from the fact that $\epsilon_s(v) < \epsilon_{t_1}(v)$ for $s > t_1$, and (b) above follows from the induction hypothesis. We have thus demonstrated that (79) holds for $t' = s$, thereby the desired proof. □

Lemma G.10. *Fix $v \in \mathcal{P}$ and $\eta > 0$. Under* MO-BAI*, for any $\epsilon \in \big(0, \tilde{c}(v,\eta)^{-1}\big)$, there exists $\delta_{\mathrm{thres}}(v,\eta,\epsilon) > 0$ such that for all $\delta \in (0, \delta_{\mathrm{thres}}(v,\eta,\epsilon))$,*

$$t\tilde{c}(v,\eta)^{-1} > \beta(t,\delta) + \epsilon t \quad \forall t \geq T_{\mathrm{thres}}(v,\eta,\epsilon,\delta), \tag{81}$$

where $T_{\mathrm{thres}}(v,\eta,\epsilon,\delta)$ is defined as

$$T_{\mathrm{thres}}(v,\eta,\epsilon,\delta) \coloneqq \frac{f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon} + \frac{MK}{\tilde{c}(v,\eta)^{-1}-\epsilon} \log\left(\left(\frac{2f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon} \right)^2 + \frac{2f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon} \right) + 1. \tag{82}$$

In (82), $f^{-1}(\delta)$ is as defined in Section 4.2.

Proof of Lemma G.10. Fix $\eta > 0$, $v \in \mathcal{P}$ and $\epsilon \in \big(0, \tilde{c}(v,\eta)^{-1}\big)$ arbitrarily. Recall the relation $\beta(t,\delta) = MK\log(t^2+t) + f^{-1}(\delta)$ for the threshold $\beta(t,\delta)$ employed by MO-BAI. For any $\bar{t} > 0$ and $\delta > 0$, let

$$H(\bar{t},\delta) \coloneqq \mathbf{1}\big\{\forall t \geq \bar{t},\ t\,\tilde{c}(v,\eta)^{-1} > MK\,\log(\bar{t}^2+\bar{t}) + f^{-1}(\delta) + \epsilon\, t\big\}. \tag{83}$$

To complete the proof, it suffices to show that $H(T_{\mathrm{thres}}(v,\eta,\epsilon,\delta),\delta) = 1$ for all sufficiently small values of δ. Note that the following is a sufficient condition for $H(\bar{t},\delta) = 1$:

$$\begin{cases} \bar{t}\,\tilde{c}(v,\eta)^{-1} > MK\,\log(\bar{t}^2+\bar{t}) + f^{-1}(\delta) + \epsilon\,\bar{t}, \\[2ex] \forall t' \geq \bar{t}, \quad \dfrac{\mathrm{d}}{\mathrm{d}t}\, t\tilde{c}(v,\eta)^{-1}\bigg|_{t=t'} \geq \dfrac{\mathrm{d}}{\mathrm{d}t}\left[MK\log(t^2+t) + f^{-1}(\delta) + \epsilon t\right]\bigg|_{t=t'}. \end{cases} \tag{84}$$

By rearranging the inequalities of (84), we get that the following is a sufficient condition for $H(\bar{t},\delta) = 1$:

$$\begin{cases} \bar{t} > \dfrac{MK\,\log(\bar{t}^2+\bar{t}) + f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}, \\[3ex] \bar{t} \geq \max\left\{ \dfrac{3}{\tilde{c}(v,\eta)^{-1}-\epsilon},\ 1 \right\}. \end{cases} \tag{85}$$

Noting that appending an extra condition on $\bar{t}$ to the conditions in (85) does not affect the sufficiency of the conditions in (85) for $H(\bar{t},\delta)=1$, we get that the following set of conditions is sufficient for $H(\bar{t},\delta)=1$:

$$\begin{cases} \bar{t} > \dfrac{MK\,\log(\bar{t}^2+\bar{t})+f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}, \\[2ex] \bar{t} \geq \max\left\{\dfrac{3}{\tilde{c}(v,\eta)^{-1}-\epsilon},\ 1\right\}, \\[2ex] \bar{t} \leq \dfrac{2\,f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}. \end{cases} \tag{86}$$

Note that for $\bar{t} \leq \frac{2f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}$, we have

$$\log\left(\left(\frac{2f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}\right)^2+\frac{2f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}\right) > \log(\bar{t}^2+\bar{t}).$$

Continuing with (86), we get that the following set of conditions is sufficient to guarantee that $H(\bar{t},\delta)=1$:

$$\begin{cases} \bar{t} > \dfrac{f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}+\dfrac{MK}{\tilde{c}(v,\eta)^{-1}-\epsilon}\,\log\left(\left(\dfrac{2\,f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}\right)^2+\dfrac{2\,f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}\right), \\[2ex] \bar{t} \geq \max\left\{\dfrac{3}{\tilde{c}(v,\eta)^{-1}-\epsilon},\ 1\right\}, \\[2ex] \bar{t} \leq \dfrac{2\,f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}. \end{cases} \tag{87}$$

For brevity in notation, let the right-hand sides of the first, second, and third lines in (87) be denoted respectively as $a(\delta)$, $b(\delta)$, and $c(\delta)$. Using the fact that $\lim_{\delta\downarrow 0} f^{-1}(\delta)=+\infty$, we get that

$$\lim_{\delta\downarrow 0}(c(\delta)-a(\delta))=+\infty, \quad \lim_{\delta\downarrow 0}(a(\delta)-b(\delta))=+\infty.$$

Let $T_{\text{thres}}(v,\eta,\epsilon,\delta)$ be as defined in the statement of the lemma. The above limiting relations imply that there exists $\delta_{\text{thres}}(v,\eta,\epsilon)>0$ sufficiently small such that $H(T_{\text{thres}}(v,\eta,\epsilon,\delta),\delta)=1$ for all $\delta\in(0,\delta_{\text{thres}}(v,\eta,\epsilon))$. This completes the desired proof. □

Lemma G.11. *Fix $v\in\mathcal{P}$, $\eta>0$, and consider the non-stopping version of* MO-BAI *(one in which the stopping rule corresponding to Lines 11-14 in Algorithm 1 are not executed). Under the aforementioned policy and under the instance v,*

$$\lim_{t\to\infty}\frac{Z(t)}{t}=\tilde{c}(v,\eta)^{-1} \quad \textit{almost surely.} \tag{88}$$

Proof of Lemma G.11. Let $\widehat{N}_{\cdot,t}=[\widehat{N}_{i,t}: i\in[K]]^\top\in\Gamma$ be defined as

$$\widehat{N}_{i,t} \coloneqq \frac{N_{i,t}}{t}, \quad i\in[K]. \tag{89}$$

In (89), $N_{i,t}$ is the total number of times arm i is pulled up to time t. Then, from (14), we have

$$\begin{aligned} \frac{Z(t)}{t} &= \frac{1}{t}\min_{m\in[M]}\ \min_{i\in[K]\setminus\widehat{i}_m(t)} \frac{N_{i,t}\,N_{\widehat{i}_m(t),t}\,\widehat{\Delta}^2_{i,m}(t)}{2(N_{i,t}+N_{\widehat{i}_m,t})} \\ &= \min_{m\in[M]}\ \min_{i\in[K]\setminus\widehat{i}_m(t)} \frac{\widehat{N}_{i,t}\,\widehat{N}_{\widehat{i}_m(t),t}\,\widehat{\Delta}^2_{i,m}(t)}{2(\widehat{N}_{i,t}+\widehat{N}_{\widehat{i}_m,t})}. \end{aligned} \tag{90}$$

In addition, by Lemma G.8, we have

$$\lim_{t\to+\infty} g_v(\widehat{\omega}_{\cdot,t}) = \tilde{c}(v,\eta)^{-1} = \sup_{\omega\in\Gamma^{(\eta)}} g_v(\omega) \quad \text{almost surely.} \tag{91}$$

Also, noting that $g_v(\cdot)$ is a continuous function on Γ with respect to the Euclidean norm $\|\cdot\|_2$, and that that Γ is compact, the Heine–Cantor theorem implies that we get that $g_v(\cdot)$ is uniformly continuous on Γ. Using this fact and taking limits as $t\to+\infty$ in (90), we get

$$\begin{aligned}
\lim_{t\to+\infty} \frac{Z(t)}{t} &= \lim_{t\to+\infty} \min_{m\in[M]} \min_{i\in[K]\setminus\widehat{i}_m(t)} \frac{\widehat{N}_{i,t}\,\widehat{N}_{\widehat{i}_m(t),t}\,\widehat{\Delta}^2_{i,m}(t)}{2(\widehat{N}_{i,t}+\widehat{N}_{\widehat{i}_m,t})} \\
&\overset{(a)}{=} \lim_{t\to+\infty} \min_{m\in[M]} \min_{i\in[K]\setminus i^*_m(v)} \frac{\widehat{N}_{i,t}\,\widehat{N}_{\widehat{i}_m(t),t}\,\Delta^2_{i,m}(v)}{2(\widehat{N}_{i,t}+\widehat{N}_{\widehat{i}_m,t})} \\
&\overset{(b)}{=} \lim_{t\to+\infty} g_v(\widehat{N}_{\cdot,t}) \\
&\overset{(c)}{=} \lim_{t\to+\infty} g_v(\widehat{\omega}_{\cdot,t}) \\
&\overset{(d)}{=} \tilde{c}(v,\eta)^{-1},
\end{aligned} \tag{92}$$

where (a) above follows from the strong law of large numbers, (b) follows from the definition of $g_v(\cdot)$, (c) follows from the fact that $\lim_{t\to\infty}\|\widehat{N}_{\cdot,t}-\widehat{\omega}_{\cdot,t}\|_2 = 0$ and that $g_v(\cdot)$ is uniformly continuous on Γ, and (d) follows from (91). □

Lemma G.12. *Fix $v\in\mathcal{P}$, $\delta\in(0,1)$, and $\eta>0$. Under* MO-BAI*,*

$$\tau_\delta < +\infty \quad \textit{almost surely.} \tag{93}$$

Proof of Lemma G.12. Note that we have almost surely

$$\begin{aligned}
\lim_{t\to\infty} \frac{\beta(t,\delta)}{Z(t)} &\overset{(a)}{=} \lim_{t\to\infty} \frac{MK\log(t^2+t)+f^{-1}(\delta)}{t\,\dfrac{Z(t)}{t}} \\
&\overset{(b)}{=} \lim_{t\to\infty} \frac{MK\log(t^2+t)+f^{-1}(\delta)}{t} \cdot \frac{1}{\tilde{c}(v,\eta)^{-1}} \\
&= 0,
\end{aligned} \tag{94}$$

where (a) follows from the definitions of $Z(t)$ and $\beta(t,\delta)$, and (b) follows from Lemma G.11. Therefore, there exists a random variable T' such that $0<T'<+\infty$ almost surely, and $Z(t)>\beta(t,\delta)$ for all $t\geq T'$, thereby proving that τ_δ is finite almost surely. □

Lemma G.13. *Fix $v\in\mathcal{P}$. For any $\eta>0$,*

$$c^*(v) \leq (1+\eta)\,\tilde{c}(v,\eta). \tag{95}$$

Proof of Lemma G.13. Fix an arbitrary $\eta>0$. Recall that

$$c^*(v)^{-1} = \sup_{\omega\in\Gamma} \min_{m\in[M]} \min_{i\in[K]\setminus i^*_m(v)} \frac{\omega_i\,\omega_{i^*_m(v)}\,\Delta^2_{i,m}(v)}{2(\omega_i+\omega_{i^*_m(v)})}, \tag{96}$$

$$\tilde{c}(v,\eta)^{-1} = \sup_{\omega\in\Gamma^{(\eta)}} \min_{m\in[M]} \min_{i\in[K]\setminus i^*_m(v)} \frac{\omega_i\,\omega_{i^*_m(v)}\,\Delta^2_{i,m}(v)}{2(\omega_i+\omega_{i^*_m(v)})}. \tag{97}$$

Let

$$\omega^*(v) \in \arg\sup_{\omega\in\Gamma} \min_{m\in[M]} \min_{i\in[K]\setminus i^*_m(v)} \frac{\omega_i\,\omega_{i^*_m(v)}\,\Delta^2_{i,m}(v)}{2(\omega_i+\omega_{i^*_m(v)})} \tag{98}$$

be chosen arbitrarily, and let $\omega' \in \mathbb{R}^K$ be defined as

$$\omega'_i \coloneqq \frac{\omega_i^*(v)}{1+\eta} + \frac{\eta}{(1+\eta)\,K}, \quad i \in [K].$$

Note that

$$\begin{aligned}\sum_{i=1}^K \omega'_i &= \sum_{i=1}^K \frac{\omega_i^*(v)}{1+\eta} + \frac{\eta}{(1+\eta)K} \\ &= \frac{1}{1+\eta} + \frac{\eta}{1+\eta} \\ &= 1, \end{aligned} \tag{99}$$

and for all $i \in [K]$,

$$\omega'_i \geq \frac{\eta}{(1+\eta)K}. \tag{100}$$

Then, (99) and (100) together imply that $\omega' \in \Gamma^{(\eta)}$, as a result of which we have

$$\begin{aligned}\tilde{c}(v,\eta)^{-1} &= \sup_{\omega \in \Gamma^{(\eta)}} \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \frac{\omega_i\, \omega_{i_m^*(v)}\, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})} \\ &\overset{(a)}{\geq} \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \frac{\omega'_i\, \omega'_{i_m^*(v)}\, \Delta_{i,m}^2(v)}{2(\omega'_i + \omega'_{i_m^*(v)})} \\ &\overset{(b)}{\geq} \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \frac{(\omega_i^*(v)/(1+\eta))\, \left(\omega_{i_m^*(v)}/(1+\eta)\right)\, \Delta_{i,m}^2(v)}{2\left(\omega'_i/(1+\eta) + \omega_{i_m^*(v)}/(1+\eta)\right)} \\ &\overset{(c)}{=} \frac{c^*(v)^{-1}}{1+\eta}, \end{aligned} \tag{101}$$

where (a) follows from the fact that $\omega' \in \Gamma^{(\eta)}$, (b) follows from the fact that $\omega'_i \geq \frac{\omega_i^*(v)}{1+\eta}$ for all $i \in [K]$, and (c) follows from the definition of $c^*(v)^{-1}$ in (5). □

Given any $\epsilon > 0$, let $T_{\text{gap}}(v,\eta,\epsilon)$ denote the smallest positive integer-valued random variable such that

$$\left| \frac{Z(t)}{t} - \tilde{c}(v,\eta)^{-1} \right| \leq \epsilon \quad \forall\ t \geq T_{\text{gap}}(v,\eta,\epsilon). \tag{102}$$

Lemma G.14. *Fix instance $v \in \mathcal{P}$ with arm means $\{\mu_{i,m} : i \in [K], m \in [M]\}$ and $\eta > 0$. For every $\epsilon > 0$,*

$$\mathbb{E}_v^{\text{MO-BAI}}[T_{\text{gap}}(v,\eta,\epsilon)] < +\infty. \tag{103}$$

Proof of Lemma G.14. Fix $\epsilon > 0$. Recall the quantity $\widehat{N}_{i,t} = \frac{N_{i,t}}{t}$. For any $\xi > 0$, let $T_{\widehat{N}}(\xi)$ denote the smallest positive integer such that

$$\max_{i \in [K]} \left| \widehat{N}_{i,t} - \widehat{\omega}_{i,t} \right| \leq \xi \quad \forall t > T_{\widehat{N}}(\xi), \tag{104}$$

let $T_\mu(\xi)$ denote the smallest positive integer such that

$$\max_{(i,m) \in [K] \times [M]} |\mu_{i,m}(v) - \widehat{\mu}_{i,m}(l_t)| \leq \xi \quad \forall t > T_\mu(\xi), \tag{105}$$

and let $T_{\widehat{\omega}}(\xi)$ denote the smallest positive integer such that

$$\max \left\{ \left| \frac{Z(t)}{t} - g_v(\widehat{N}_{\cdot,t}) \right|, \quad \left| \tilde{c}(v,\eta)^{-1} - g_v(\widehat{\omega}_{\cdot,t}) \right| \right\} < \xi \quad \forall t > T_{\widehat{\omega}}(\xi). \tag{106}$$

Recall from the proof of Lemma G.11 that $g_v(\cdot)$ is uniformly continuous on Γ. Then, from (104) and (106), we get that there exists $\epsilon_1 > 0$ such that

$$T_{\text{gap}}(v,\eta,\epsilon) < \max\{T_{\widehat{N}}(\epsilon_1), T_{\widehat{\omega}}(\epsilon_1)\} \quad \text{almost surely}. \tag{107}$$

Replacing t_2 with t in (60), we note that almost surely,

$$\begin{aligned}
&\left|\tilde{c}(v,\eta)^{-1} - g_v(\widehat{\omega}_{\cdot,t})\right| \\
&\leq \frac{2}{\sqrt{t}}\,\tilde{c}(v,\eta)^{-1} + 11\epsilon_{\lceil\sqrt{t}\rceil}(v) + \frac{2\,\log(t)\,\overline{C}_{\lceil\sqrt{t}\rceil}(\eta)}{t} \\
&= \frac{2}{\sqrt{t}}\tilde{c}(v,\eta)^{-1} + 11\sup_{t'\geq l_{\lceil\sqrt{t}\rceil}}\left(\sup_{\omega\in\Gamma^{(\eta)}}\left|g_v(\omega) - g_{\widehat{v}_{t'}}(\omega)\right|\right) + \frac{2\log(t)}{t}\sup_{t'\geq l_{\lceil\sqrt{t}\rceil}} C(\widehat{v}_{t'},\eta) \\
&\leq \frac{2}{\sqrt{t}}\,\tilde{c}(v,\eta)^{-1} + 11\sup_{t'\geq\lceil\sqrt{t}\rceil/2}\left(\sup_{\omega\in\Gamma^{(\eta)}}\left|g_v(\omega) - g_{\widehat{v}_{t'}}(\omega)\right|\right) + \frac{2\log(t)}{t}\sup_{t'\geq\lceil\sqrt{t}\rceil/2} C(\widehat{v}_{t'},\eta), \qquad (108)
\end{aligned}$$

where the last line above follows by noting that $l_t \geq t/2$ for any t (see the definition of l_t in Line 8 of Algorithm 1). Also, (56) implies that almost surely,

$$\sup_{\omega\in\Gamma^{(\eta)}}|g_v(\omega) - g_{\widehat{v}_{l_t}}(\omega)| \leq \max_{(i,m)\in[K]\times[M]}|\Delta_{i,m}(\widehat{v}_{l_t}) - \Delta_{i,m}(v)| \quad \forall t \geq \max_{m\in[M]} T_m, \qquad (109)$$

where T_m (for any $m \in [M]$) is as defined in (54). Notice that for each $m \in [M]$, we have $i^*_m(\widehat{v}_{l_t}) = i^*_m(v)$ for all $t \geq T_m$ almost surely by the definition of T_m. Therefore, it follows that almost surely,

$$\begin{aligned}
&\max_{(i,m)\in[K]\times[M]}|\Delta_{i,m}(\widehat{v}_{l_t}) - \Delta_{i,m}(v)| \\
&= \max_{(i,m)\in[K]\times[M]}\left|\mu_{i^*_m(v),m} - \mu_{i,m}(v) - \left(\widehat{\mu}_{i^*_m(\widehat{v}_{l_t}),m}(l_t) - \widehat{\mu}_{i,m}(l_t)\right)\right| \\
&= \max_{(i,m)\in[K]\times[M]}\left|\mu_{i^*_m(v),m} - \widehat{\mu}_{i^*_m(v),m}(l_t) - \left(\mu_{i,m}(v) - \widehat{\mu}_{i,m}(l_t)\right)\right| \\
&\leq 2\max_{(i,m)\in[K]\times[M]}|\mu_{i,m}(v) - \widehat{\mu}_{i,m}(l_t)| \qquad (110)
\end{aligned}$$

for all $t \geq \max_{m\in[M]} T_m$. Hence, we get that there exists $\epsilon_2 > 0$ such that

$$\max_{m\in[M]} T_m < T_\mu(\epsilon_2) \quad \text{almost surely.} \qquad (111)$$

In addition, by Lemma G.5, we have $\forall t' > 0$,

$$C(\widehat{v}_{t'},\eta) \leq \max_{i\neq i^*_m(\widehat{v}_{t'})}\frac{2\Delta^2_{i,m}(\widehat{v}_{t'})K(1+\eta)}{\eta} \quad \text{almost surely.} \qquad (112)$$

Then, combining (108), (109), (111) and (112), we get that there exists $\epsilon_3 \in (0,\epsilon_2)$ and $T' \in \mathbb{N}$ such that

$$T_{\widehat{\omega}}(\epsilon_1) < \max\{T', T_\mu(\epsilon_3)\} \quad \text{almost surely.} \qquad (113)$$

Combining (107) and (113), and using the fact that $\max\{a,b\} \leq a+b$ for all $a,b \geq 0$, we have

$$T_{\text{gap}}(v,\eta,\epsilon) \leq T_{\widehat{N}}(\epsilon_1) + T_\mu(\epsilon_3) + T' \quad \text{almost surely.} \qquad (114)$$

Notice that

$$\left|\widehat{N}_{i,t} - \widehat{\omega}_{i,t}\right| = \left|\frac{B_{i,t}}{t}\right| \leq \frac{1}{t}, \qquad (115)$$

where the equality follows from the definition of $\widehat{\omega}_{i,t}$ in Line 7 of Algorithm 1, and the inequality follows by noting that the buffer size $|B_{i,t}| \leq 1$ for all t. Therefore, we have

$$T_{\widehat{N}}(\epsilon_1) < \frac{1}{\epsilon_1} \quad \text{almost surely.} \qquad (116)$$

For any $i \in [K]$, $m \in [M]$, $t \in \mathbb{N}$, and $\xi > 0$, let

$$\mathcal{E}(t,i,m,\xi) \coloneqq \left\{|\widehat{\mu}_{i,m}(t) - \mu_{i,m}(v)| > \xi\right\}.$$

Then, we have

$$\begin{aligned}
\mathbb{E}_v^{\text{MO-BAI}}\left[T_\mu(\epsilon_3)\right] &= \sum_{t=1}^{\infty} \mathbb{P}_v^{\text{MO-BAI}}\left(T_\mu(\epsilon_3) > t\right) \\
&\leq K + \sum_{t=K+1}^{\infty} \mathbb{P}_v^{\text{MO-BAI}}\left(T_\mu(\epsilon_3) > t\right) \\
&\leq K + \sum_{i=1}^{K} \sum_{m=1}^{M} \sum_{t=K+1}^{\infty} \mathbb{P}_v^{\text{MO-BAI}}\left(\bigcup_{t' > l_t} \mathcal{E}(t', i, m, \epsilon_3)\right) \\
&\overset{(a)}{\leq} K + \sum_{i=1}^{K} \sum_{m=1}^{M} \sum_{t=K+1}^{\infty} \sum_{t'=t}^{\infty} \mathbb{P}_v^{\text{MO-BAI}}\left(\mathcal{E}(t', i, m, \epsilon_3)\right), \qquad (117)
\end{aligned}$$

where (a) above follows from the union bound. We now observe that for any $t' > K$,

$$\begin{aligned}
\mathcal{E}(t', i, m, \xi) &= \left\{\left|\widehat{\mu}_{i,m}(t') - \mu_{i,m}(v)\right| > \xi\right\} \\
&= \left\{\frac{\left|\sum_{s=1}^{t'} \mathbf{1}_{\{A_s = i\}} \left(X_{A_s,m}(s) - \mu_{i,m}(v)\right)\right|}{N_{i,t'}} > \xi\right\} \\
&\subseteq \left\{\left|\sum_{s=1}^{t'} \mathbf{1}_{\{A_s = i\}} \left(X_{A_s,m}(s) - \mu_{i,m}(v)\right)\right| > \xi\left(\frac{\eta}{K(1+\eta)} t' - 1\right)\right\}, \qquad (118)
\end{aligned}$$

where the last line above follows by noting that for all t', under MO-BAI,

$$\begin{aligned}
N_{i,t'} &= t'\,\widehat{\omega}_{i,t'} - B_{i,t'} \\
&\geq \frac{\eta}{K(1+\eta)} t' - 1. \qquad (119)
\end{aligned}$$

The inequality above follows by observing that $\widehat{\omega}_{\cdot,t'} \in \Gamma^{(\eta)}$ for all t'. We note that

$$\left\{\sum_{s=1}^{t'} \mathbf{1}_{\{A_s = i\}} \left(X_{A_s,m}(s) - \mu_{i,m}(v)\right)\right\}_{t'=K+1}^{\infty}$$

is a bounded martingale difference sequence with finite variance. Using (118) in (117) together with martingale concentration bounds (de la Pena, 1999, Theorem 1.2A), we get

$$\begin{aligned}
\mathbb{E}_v^{\text{MO-BAI}}\left[T_\mu(\epsilon_3)\right] &\leq K + \sum_{i=1}^{K} \sum_{m=1}^{M} \sum_{t=K+1}^{\infty} \sum_{t'=t}^{\infty} \exp\left(-\Big(\frac{\eta\, t'}{K(1+\eta)} - 1\Big) c_{\epsilon_3}\right) \\
&= K + \sum_{i=1}^{K} \sum_{m=1}^{M} \sum_{t'=K+1}^{\infty} t' \exp\left(-\Big(\frac{\eta\, t'}{K(1+\eta)} - 1\Big) c_{\epsilon_3}\right) \\
&< +\infty. \qquad (120)
\end{aligned}$$

In the above set of relations, c_{ϵ_3} is a positive constant that depends only on ϵ_3. Finally, combining (114), (116), and (120), we arrive at the desired result. This completes the proof. □

With the ingredient of above lemmas, we are ready to prove Theorem 4.7.

Proof of Theorem 4.7. Fix $\eta > 0$ and instance $v \in \mathcal{P}$. Consider MO-BAI. By Lemma G.11, we have $T_{\text{gap}}(v, \eta, \epsilon) < +\infty$ almost surely. For any $\epsilon \in \left(0, \tilde{c}(v, \eta)^{-1}\right)$ and $\delta \in (0, \delta_{\text{thres}}(v, \eta, \epsilon))$, it follows from Lemma G.10 that

$$Z(t) > \beta(t, \delta) \quad \forall\, t \geq \max\left\{T_{\text{gap}}(v, \eta, \epsilon), T_{\text{thres}}(v, \eta, \delta, \epsilon), K\right\} \quad \text{almost surely,}$$

which implies that almost surely,

$$\begin{aligned}\tau_\delta &\le \max\left\{T_{\mathrm{gap}}(v,\eta,\epsilon), T_{\mathrm{thres}}(v,\eta,\delta,\epsilon), K\right\}+1\\ &\le T_{\mathrm{gap}}(v,\eta,\epsilon)+T_{\mathrm{thres}}(v,\eta,\delta,\epsilon)+K+1. \end{aligned} \tag{121}$$

Hence, for any $\epsilon \in \left(0, \tilde{c}(v,\eta)^{-1}\right)$, we have almost surely

$$\begin{aligned}&\limsup_{\delta\downarrow 0}\frac{\tau_\delta}{\log\left(\frac{1}{\delta}\right)}\\ &\overset{(a)}{\le} \limsup_{\delta\downarrow 0}\ \frac{T_{\mathrm{gap}}(v,\eta,\epsilon)+T_{\mathrm{thres}}(v,\eta,\delta,\epsilon)+K+1}{\log\left(\frac{1}{\delta}\right)}\\ &\overset{(b)}{=} \limsup_{\delta\downarrow 0}\ \frac{T_{\mathrm{thres}}(v,\eta,\delta,\epsilon)}{\log\left(\frac{1}{\delta}\right)}\\ &\overset{(c)}{=} \limsup_{\delta\downarrow 0}\ \frac{\dfrac{f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}+\dfrac{MK}{\tilde{c}(v,\eta)^{-1}-\epsilon}\ \log\left(\left(\dfrac{2\,f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}\right)^2+\dfrac{2\,f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}\right)+1}{\log\left(\frac{1}{\delta}\right)}\\ &\overset{(d)}{=} \frac{1}{\tilde{c}(v,\eta)^{-1}-\epsilon}\\ &\overset{(e)}{\le} \frac{1}{((1+\eta)\,c^*(v))^{-1}-\epsilon}, \end{aligned} \tag{122}$$

where (a) follows from (121), (b) follows from the fact that $T_{\mathrm{gap}}(v,\eta,\epsilon) < +\infty$ almost surely and that $T_{\mathrm{gap}}(v,\eta,\epsilon)$ does not depend on δ, (c) follows from the definition of $T_{\mathrm{thres}}(\cdot)$, (d) follows from Lemma G.3, and (e) follows from Lemma G.13. Letting $\epsilon \to 0$, we have

$$\limsup_{\delta\downarrow 0}\frac{\tau_\delta}{\log\left(\frac{1}{\delta}\right)} \le (1+\eta)\,c^*(v) \quad \text{almost surely.}$$

In addition, we have

$$\begin{aligned}&\limsup_{\delta\downarrow 0}\ \frac{\mathbb{E}_v^{\text{MO-BAI}}[\tau_\delta]}{\log\left(\frac{1}{\delta}\right)}\\ &\overset{(a)}{\le} \limsup_{\delta\downarrow 0}\ \frac{\mathbb{E}_v^{\text{MO-BAI}}[T_{\mathrm{gap}}(v,\eta,\epsilon)]+T_{\mathrm{thres}}(v,\eta,\delta,\epsilon)+K+1}{\log\left(\frac{1}{\delta}\right)}\\ &\overset{(b)}{=} \limsup_{\delta\downarrow 0}\ \frac{T_{\mathrm{thres}}(v,\eta,\delta,\epsilon)}{\log\left(\frac{1}{\delta}\right)}\\ &\overset{(c)}{=} \limsup_{\delta\downarrow 0}\ \frac{\dfrac{f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}+\dfrac{MK}{\tilde{c}(v,\eta)^{-1}-\epsilon}\ \log\left(\left(\dfrac{2\,f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}\right)^2+\dfrac{2\,f^{-1}(\delta)}{\tilde{c}(v,\eta)^{-1}-\epsilon}\right)+1}{\log\left(\frac{1}{\delta}\right)}\\ &\overset{(d)}{=} \frac{1}{\tilde{c}(v,\eta)^{-1}-\epsilon}\\ &\overset{(e)}{\le} \frac{1}{((1+\eta)\,c^*(v))^{-1}-\epsilon}, \end{aligned} \tag{123}$$

where (a) follows from (121), (b) follows from the fact that $\mathbb{E}_v^{\text{MO-BAI}}[T_{\mathrm{gap}}(v,\eta,\epsilon)] < +\infty$ (see Lemma G.14) and that $\mathbb{E}_v^{\text{MO-BAI}}(T_{\mathrm{gap}}(v,\eta,\epsilon))$ does not depend on δ, (c) follows from the definition of $T_{\mathrm{thres}}(\cdot)$, (d) follows from Lemma G.3, and (e) follows from Lemma G.13. Letting $\epsilon \to 0$, we get

$$\limsup_{\delta\downarrow 0}\frac{\mathbb{E}_v^{\text{MO-BAI}}[\tau_\delta]}{\log\left(\frac{1}{\delta}\right)} \le (1+\eta)\,c^*(v),$$

thereby arriving at the desired result. This completes the proof. □