FEW-FOR-MANY BEST-ARM IDENTIFICATION IN MULTI-OBJECTIVE BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world systems often serve many objectives but can deploy only a few options, whose performance must be learned from noisy feedback. Modeling each option as an arm in a multi-objective bandit, we introduce *few-for-many best-arm identification*, which seeks a small set of arms minimizing the worst weighted degradation from the objective-wise optima, rather than the best arm for every objective. We consider both known objective groups and latent groups specified only by a representative budget, and propose GFM-ELIM and AFM-ELIM with instance-dependent sample-complexity guarantees. For Gaussian rewards, we further derive a lower bound for the latent-group setting and show that it is asymptotically attainable with high probability by an oracle-based Track-and-Stop procedure. Our results show that sharing observations across objectives and ignoring irrelevant objective-wise distinctions can substantially reduce the sampling cost, as confirmed by experiments on synthetic and real-world tasks.

1 INTRODUCTION

Modern decision systems frequently need to serve many evaluation criteria or user groups with only a handful of deployable choices. This motivates the *few-for-many* paradigm, which selects a small set of complementary solutions that jointly cover many objectives (Liu et al., 2024; Lin et al., 2025; Liu et al., 2025; Shang et al., 2026). For example, clinical trials may seek a few treatment options that collectively perform well across multiple efficacy and safety outcomes (Kavelaars et al., 2020; Villar et al., 2015); online advertising platforms may use a few campaigns to reach heterogeneous target audiences (Geng et al., 2020; Schwartz et al., 2017); and LLM services may route diverse queries among a limited pool of models with different capabilities and inference costs (Chen et al., 2024; Ong et al., 2025).

A key challenge in these applications is that the multi-objective performance of each candidate is unknown and must be learned from noisy observations, such as patient outcomes, user responses to advertisements, or repeated model evaluations. Few-for-many selection thus becomes a statistical identification problem: the learner must find a small representative set whose members jointly serve many objectives, while simultaneously learning their unknown performance. Multi-objective multi-armed bandits offer a natural framework for this setting, associating each arm with an unknown vector-valued mean reward (Drugan & Nowe, 2013; Lattimore & Szepesvári, 2020). In the pure-exploration regime, the goal is to identify the desired representative set with probability at least $1-\delta$ using as few arm pulls as possible (Audibert & Bubeck, 2010; Garivier & Kaufmann, 2016).

Existing multi-objective formulations typically lie at one of two extremes. At one end, Pareto-front identification (Auer et al., 2016; Ararat & Tekin, 2023; Crepon et al., 2024) seeks all nondominated arms; as the number of objectives grows, however, the Pareto set can become large, increasing both the statistical cost of recovering it and the burden of choosing among its members. At the other end, scalarization-based methods (Paria et al., 2020; Lin et al., 2024) collapse multiple objectives into a scalar utility and return a single solution for a given preference, which can be overly restrictive when different objectives favor different arms. These limitations call for an intermediate goal: identifying a few complementary arms that jointly serve many objectives.

The closest bandit formulation, *multi-objective best-arm identification* (MO-BAI), seeks the best arm for every objective (Chen et al., 2025). This target is appropriate when a separate decision can be deployed for each objective, but it can be unnecessarily costly when many objectives must be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

served under a limited deployment budget. The few-for-many principle instead selects a prescribed number of complementary solutions such that every objective is well served by at least one of them (Lin et al., 2025; Liu et al., 2025). However, existing studies apply this principle to deterministic many-objective optimization, whereas in the bandit setting both the objective-wise ideal values and the quality of any candidate representative set must be inferred from noisy observations.

To bridge this gap, we formulate a stochastic counterpart of few-for-many optimization, which we call *few-for-many best-arm identification*. The quality of a representative set is measured by its worst weighted degradation from the objective-wise ideal values. We consider two forms of prior structure. In the *known-group* setting, the objectives are partitioned in advance and one arm is selected for each group. In the *latent-group* setting, only a representative budget L is specified, and the selected arms implicitly induce groups according to which representative best serves each objective. In the latter setting, $L = 1$ recovers a single robust representative, whereas a sufficiently large L recovers objective-wise identification; the intermediate regime $1 < L \ll M$ captures the practically relevant case in which a few deployable arms must jointly serve many objectives. Our main contributions are as follows:

- We introduce a few-for-many bandit framework that covers both known and latent objective groups. It bridges MO-BAI (Chen et al., 2025) and few-for-many optimization (Lin et al., 2025), accommodating both noisy observations and a limited representative budget.
- We develop two phase-based elimination algorithms, GFM-ELIM for known objective groups and AFM-ELIM for latent groups, and establish finite-time sample-complexity guarantees for both. Unlike MO-BAI (Chen et al., 2025), our algorithms need not resolve objective-wise distinctions that do not affect the optimal representatives, which can substantially reduce the sampling cost in many-objective settings.
- For Gaussian rewards, we establish an information-theoretic lower bound for the latent-group setting that characterizes its fundamental instance-dependent sample complexity. We further show in Appendix C that this lower bound is asymptotically attainable with high probability by an oracle-based Track-and-Stop procedure.
- We evaluate the proposed algorithms on synthetic and real-world tasks. The results show that GFM-ELIM and AFM-ELIM reliably identify compact representative sets while requiring fewer samples than existing multi-objective identification baselines.

Related Work. Best-arm identification (BAI) aims to identify the optimal decision using as few samples as possible. Garivier & Kaufmann (2016) established the instance-dependent asymptotic lower bound for classical BAI and introduced Track-and-Stop, which has since been extended to structured pure-exploration problems, including combinatorial answer spaces and multiple correct answers (Chen et al., 2014; Degenne & Koolen, 2019; Jourdan et al., 2021). For vector-valued rewards, multi-objective bandits have been studied under Pareto-based and preference-based criteria (Drugan & Nowe, 2013; Auer et al., 2016; Shukla & Basu, 2024); in particular, Chen et al. (2025) identify the best arm for every objective and develop asymptotically optimal sampling strategies. In parallel, few-for-many optimization seeks a small set of complementary solutions that jointly serve many objectives, as exemplified by the many-objective cover problem (Liu et al., 2024), Tchebycheff set scalarization (Lin et al., 2025), and the general F4M framework (Liu et al., 2025; Shang et al., 2026). These works mainly consider deterministic optimization, where objective evaluations or gradients are directly accessible. Our work connects these two lines of research by studying few-for-many identification with unknown vector-valued means, where only the distinctions that can affect the optimal representatives need to be resolved from noisy sequential observations. A more detailed discussion is provided in Appendix A.

2 PRELIMINARIES

For any positive integer n, let $[n] = \{1, \dots, n\}$. For a finite set $\mathcal{X}$, $|\mathcal{X}|$ denotes its cardinality. For any $x \in \mathbb{R}$, let $[x]_+ = \max\{x, 0\}$. We use $\Delta_K = \{w \in \mathbb{R}_+^K : \sum_{a=1}^K w_a = 1\}$ to denote the probability simplex over K arms. We consider a stochastic multi-armed bandit problem with K arms and M objectives. At each time step t, the learner selects an arm $a_t \in [K]$ and observes a vector-valued reward $\boldsymbol{r}_t = (r_{t,1}, \dots, r_{t,M})^\top \in \mathbb{R}^M$. For each arm $a \in [K]$, let $\boldsymbol{\mu}_a = (\mu_{a,1}, \dots, \mu_{a,M})^\top$

denote its unknown mean reward vector, where

$$\mu_{a,m} = \mathbb{E}[r_{t,m} \mid \mathcal{F}_{t-1}, a_t = a], \quad \forall m \in [M],$$

and $\mathcal{F}_{t-1}$ denotes the history available before time step t. We assume that the reward noise is σ^2-sub-Gaussian; that is, for any $s \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(s\left(r_{t,m} - \mu_{a,m}\right)\right) | \mathcal{F}_{t-1}, a_t = a\right] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right). \tag{1}$$

The reward coordinates may be statistically dependent, since our finite-time analysis relies only on the marginal concentration property in Eq. (1). For ease of presentation, we assume $\mu_{a,m} \in [0, 1]$.

We further assume that each objective has a unique best arm. For every objective $m \in [M]$, we define its ideal value and best arm as

$$\mu_m^\star = \max_{a \in [K]} \mu_{a,m}, \quad b_m^\star = \arg\max_{a \in [K]} \mu_{a,m}.$$

Few-for-Many Identification. When the number of objectives is large, identifying a separate best arm for every objective may yield many arms, whereas requiring a single arm to perform well on all objectives can be overly restrictive. Motivated by the few-for-many paradigm (Lin et al., 2025), we seek a small number of representative arms that together perform well across many objectives, with each representative serving a group of objectives. We consider two complementary settings, depending on whether the objective groups are known in advance.

Known Objective Groups. In the *known-group* setting, objectives are partitioned in advance, and all objectives in a group share one representative arm. Let $1 \leq Q \ll M$ and let $\mathcal{G} = \{G_1, \ldots, G_Q\}$ be a partition of $[M]$, i.e., $[M] = \cup_{q=1}^{Q} G_q$, where G_q contains the objectives in the q-th group. Let $\lambda_m > 0$ be a known weight for objective $m \in [M]$, which reflects its relative importance.

For every group $q \in [Q]$ and arm $a \in [K]$, we define the group coverage loss as

$$R_q(a) = \max_{m \in G_q} \lambda_m \left(\mu_m^\star - \mu_{a,m}\right). \tag{2}$$

Thus, $R_q(a)$ measures the largest weighted degradation incurred when arm a represents all objectives in G_q. Accordingly, the identification target is

$$\boldsymbol{a}_{\mathcal{G}}^\star = (a_1^\star, \ldots, a_Q^\star), \qquad a_q^\star = \arg\min_{a \in [K]} R_q(a). \tag{3}$$

When $Q = 1$, the problem reduces to identifying a single robust representative for all objectives. When $Q = M$ and $G_q = \{q\}$, it reduces to identifying the best arm for every objective.

Latent Objective Groups. In the *latent-group* setting, the objective partition is not known in advance. Instead, the learner is given a representative budget $1 \leq L \ll M$ and seeks L arms that jointly provide good coverage of all objectives, with the objective groups implicitly induced by the selected representatives.

For any nonempty arm set $S \subseteq [K]$, we define its overall coverage loss as

$$\Phi(S) = \max_{m \in [M]} \lambda_m \left(\mu_m^\star - V_m(S)\right), \qquad V_m(S) = \max_{a \in S} \mu_{a,m}. \tag{4}$$

Hence, $\Phi(S)$ measures the largest weighted degradation incurred when each objective is served by its best arm in S rather than by its best arm in $[K]$. Accordingly, the identification target is

$$S_L^\star = \arg\min_{S \in \mathcal{S}_L} \Phi(S), \qquad \mathcal{S}_L = \{S \subseteq [K] : |S| = L\}. \tag{5}$$

The optimal set $S_L^\star$ naturally induces an assignment of objectives to representatives. Specifically, let $\pi_L^\star(m) = \arg\max_{a \in S_L^\star} \mu_{a,m}$. The objectives assigned to the same representative arm then form a latent group:

$$G_a^\star = \{m \in [M] : \pi_L^\star(m) = a\}, \quad a \in S_L^\star.$$

Thus, the objective partition is induced directly by $S_L^\star$ rather than identified separately. When $L = 1$, the problem reduces to identifying a single robust representative for all objectives, which coincides with the known-group setting with $Q = 1$. At the other extreme, once $L \geq |\{b_m^\star : m \in [M]\}|$ (in particular, whenever $L \geq M$), the representative set can contain all objective-wise best arms and attains zero coverage loss.

Algorithm 1 GFM-ELIM: Grouped Few-for-Many Elimination

Input: Confidence level δ, groups $\mathcal{G} = \{G_1, \ldots, G_Q\}$, and weights $\{\lambda_m\}_{m\in[M]}$

$\mathcal{B}_{m,1} \leftarrow [K]$ for all m; $\mathcal{C}_{q,1} \leftarrow [K]$ for all q
for $r = 1, 2, \ldots$ **do**
  $\mathcal{Q}_r \leftarrow \{q : |\mathcal{C}_{q,r}| > 1\}$
  if $\mathcal{Q}_r = \emptyset$ **then**
    return the unique arms in $(\mathcal{C}_{q,r})_{q\in[Q]}$
  end if
  $\mathcal{M}_r \leftarrow \bigcup_{q\in\mathcal{Q}_r} G_q$; $\mathcal{U}_r \leftarrow (\bigcup_{q\in\mathcal{Q}_r} \mathcal{C}_{q,r}) \cup (\bigcup_{m\in\mathcal{M}_r} \mathcal{B}_{m,r})$
  Sample every $a \in \mathcal{U}_r$ until it has n_r observations and compute $\widehat{\mu}^{(r)}_{a,m}$
  for $m \in \mathcal{M}_r$ **do**
    $\widehat{\mu}^{\star}_{m,r} = \max_{b\in\mathcal{B}_{m,r}} \widehat{\mu}^{(r)}_{b,m}$
    $\mathcal{B}_{m,r+1} \leftarrow \{a \in \mathcal{B}_{m,r} : \widehat{\mu}^{(r)}_{a,m} \geq \widehat{\mu}^{\star}_{m,r} - 2\epsilon_r\}$
  end for
  for each $q \in \mathcal{Q}_r$ **do**
    Compute $\underline{R}_{q,r}(a)$ and $\overline{R}_{q,r}(a)$ using Eq. (7)
    $\mathcal{C}_{q,r+1} \leftarrow \{a \in \mathcal{C}_{q,r} : \underline{R}_{q,r}(a) \leq \min_b \overline{R}_{q,r}(b)\}$
  end for
  Keep all inactive sets unchanged
end for

3 FINITE-TIME ELIMINATION ALGORITHMS

This section presents finite-time elimination algorithms for the two few-for-many settings: known objective groups and latent objective groups.

3.1 GROUPED FEW-FOR-MANY ELIMINATION

We begin with the known-group setting. Although the objective groups are given, the group coverage losses cannot be evaluated directly because the mean rewards are unknown. To address this, we propose GFM-ELIM, a phase-based elimination algorithm that jointly learns the mean rewards and the representative of each group. The complete procedure is summarized in Algorithm 1.

At phase r, GFM-ELIM maintains benchmark candidate sets $\mathcal{B}_{m,r}$ for the objective-wise ideal values and representative candidate sets $\mathcal{C}_{q,r}$ for the group representatives, all initialized to $[K]$. Let $\mathcal{Q}_r = \{q : |\mathcal{C}_{q,r}| > 1\}$ denote the set of unresolved groups and $\mathcal{M}_r = \cup_{q\in\mathcal{Q}_r} G_q$ be the associated objectives. The algorithm samples only the arms in $\mathcal{U}_r = (\cup_{q\in\mathcal{Q}_r} \mathcal{C}_{q,r}) \cup (\cup_{m\in\mathcal{M}_r} \mathcal{B}_{m,r})$, so that each observation is shared across all unresolved groups for which the arm remains relevant.

Each phase refines the candidates at a finer accuracy than the previous one. Specifically, let

$$\epsilon_r = 2^{-r}, \qquad \delta_r = \frac{6\delta}{\pi^2 r^2}, \qquad n_r = \left\lceil \frac{2\sigma^2}{\epsilon_r^2} \log\left(\frac{2KM}{\delta_r}\right) \right\rceil. \tag{6}$$

This choice of n_r ensures that, with probability at least $1 - \delta_r$, the empirical means of all arm–objective pairs are simultaneously accurate to within ϵ_r. Each active arm is therefore sampled until it has accumulated n_r observations, from which GFM-ELIM computes the empirical means $\widehat{\mu}^{(r)}_{a,m}$.

GFM-ELIM first refines the benchmark candidates. For each active objective $m \in \mathcal{M}_r$, the empirical benchmark is defined as $\widehat{\mu}^{\star}_{m,r} = \max_{b\in\mathcal{B}_{m,r}} \widehat{\mu}^{(r)}_{b,m}$. On the confidence event, an arm whose empirical mean falls more than $2\epsilon_r$ below this benchmark cannot be the best arm for objective m, and is therefore eliminated, i.e., $\mathcal{B}_{m,r+1} = \{a \in \mathcal{B}_{m,r} : \widehat{\mu}^{(r)}_{a,m} \geq \widehat{\mu}^{\star}_{m,r} - 2\epsilon_r\}$.

The representative candidates are then updated according to their group coverage losses. Since both the empirical benchmark and the empirical mean of arm a are accurate to within ϵ_r, the coverage loss $R_q(a)$ is bracketed by

$$\underline{R}_{q,r}(a) = \max_{m\in G_q} \lambda_m \left[\widehat{\mu}^{\star}_{m,r} - \widehat{\mu}^{(r)}_{a,m} - 2\epsilon_r\right]_+, \; \overline{R}_{q,r}(a) = \max_{m\in G_q} \lambda_m \left[\widehat{\mu}^{\star}_{m,r} - \widehat{\mu}^{(r)}_{a,m} + 2\epsilon_r\right]_+. \tag{7}$$

An arm is eliminated as a representative candidate once its optimistic loss exceeds the smallest pessimistic loss in its group, i.e., $\mathcal{C}_{q,r+1} = \{a \in \mathcal{C}_{q,r} : \underline{R}_{q,r}(a) \le \min_b \overline{R}_{q,r}(b)\}$.

This procedure continues until every group has a unique representative candidate. We next state its finite-time guarantee, which reflects the arm-dependent stopping behavior described above. To characterize the complexity, we define the group-representative gap and the objective-wise gap as

$$\Delta_{q,a} = R_q(a) - R_q(a_q^\star), \quad g_{m,a} = \mu_m^\star - \mu_{a,m}.$$

For each $q \in [Q]$, let $\overline{\lambda}_q = \max_{m \in G_q} \lambda_m$ and define its identification radius as $\rho_q = \min_{a \neq a_q^\star} \frac{\Delta_{q,a}}{8\overline{\lambda}_q}$. The effective gap of arm a is then defined as

$$\Gamma_a = \min_{q \in [Q]} \left\{ \max \left\{ \rho_q, \frac{\Delta_{q,a}}{8\overline{\lambda}_q} \right\}, \min_{m \in G_q} \max \left\{ \rho_q, \frac{g_{m,a}}{4} \right\} \right\}. \tag{8}$$

The two terms in Eq. (8) correspond to the two roles of arm a: the first determines when it can be eliminated as a representative candidate for group q, while the second determines when it is no longer needed to estimate the objective-wise benchmarks within that group. The radius ρ_q reflects the fact that both roles terminate once group q is resolved. Hence, Γ_a captures the hardest role of arm a across all groups and determines the last phase in which it is sampled.

Theorem 1 (Instance-dependent sample complexity)**.** *With probability at least $1-\delta$, Algorithm 1 returns the optimal representative tuple $\boldsymbol{a}_{\mathcal{G}}^\star$ using at most $T_{\mathrm{GFM}}(\boldsymbol{\mu}, \delta)$ arm pulls, where*

$$T_{\mathrm{GFM}}(\boldsymbol{\mu}, \delta) = \widetilde{O}\left(\sigma^2 \sum_{a=1}^{K} \frac{1}{\Gamma_a^2} \log \frac{KM}{\delta} \right). \tag{9}$$

Theorem 1 shows that GFM-ELIM is instance-adaptive at the arm level: each arm is sampled only until all of its benchmark and representative roles have been resolved. Consequently, the complexity depends on the hardest remaining role of each arm through Γ_a, rather than accumulating a separate sampling cost for every group. This shared-feedback effect can improve the leading complexity term by a factor of up to Q relative to running an independent elimination procedure for each group.

3.2 ADAPTIVE FEW-FOR-MANY ELIMINATION

We next turn to the latent-group setting, where no objective partition is available and the learner must identify the optimal representative set directly. We propose AFM-ELIM, which lifts the elimination principle of GFM-ELIM from individual arms to candidate sets. Importantly, although elimination is performed at the set level, sampling remains at the base-arm level, so that observations are shared across candidate sets. The complete procedure is summarized in Algorithm 2.

At phase r, AFM-ELIM maintains a benchmark candidate set $\mathcal{B}_{m,r}$ for each objective $m \in [M]$ and a family $\mathcal{C}_r$ of candidate representative sets, initialized as $\mathcal{B}_{m,1} = [K]$ and $\mathcal{C}_1 = \mathcal{S}_L$, respectively. The algorithm samples only the base arms that either belong to an active candidate set or remain relevant for estimating an objective-wise benchmark, i.e., $\mathcal{U}_r = (\cup_{S \in \mathcal{C}_r} S) \cup \left(\cup_{m=1}^{M} \mathcal{B}_{m,r}\right)$. Thus, a single pull of an arm simultaneously refines the estimates of all active candidate sets containing that arm, as well as all objectives for which it remains a benchmark candidate.

AFM-ELIM uses the same phase schedule in Eq. (6) and the same benchmark-refinement rule as GFM-ELIM. The remaining candidate sets are then compared through their empirical coverage losses. For each $S \in \mathcal{C}_r$, its empirical value on objective m is $\widehat{V}_{m,r}(S) = \max_{a \in S} \widehat{\mu}_{a,m}^{(r)}$. Since both the empirical benchmark $\widehat{\mu}_{m,r}^\star$ and the empirical set value $\widehat{V}_{m,r}(S)$ are accurate to within ϵ_r on the confidence event, the coverage loss $\Phi(S)$ is bracketed by

$$\underline{\Phi}_r(S) = \max_{m \in [M]} \lambda_m \left[\widehat{\mu}_{m,r}^\star - \widehat{V}_{m,r}(S) - 2\epsilon_r \right]_+, \quad \overline{\Phi}_r(S) = \max_{m \in [M]} \lambda_m \left[\widehat{\mu}_{m,r}^\star - \widehat{V}_{m,r}(S) + 2\epsilon_r \right]_+. \tag{10}$$

A candidate set is eliminated once its optimistic loss exceeds the pessimistic loss of some other active set, i.e., $\mathcal{C}_{r+1} = \{S \in \mathcal{C}_r : \underline{\Phi}_r(S) \le \min_{S' \in \mathcal{C}_r} \overline{\Phi}_r(S')\}$.

This procedure continues until a unique candidate set remains. We next characterize its arm-dependent stopping behavior. For every candidate set $S \in \mathcal{S}_L$, we define its identification gap

Algorithm 2 AFM-ELIM: Adaptive Few-for-Many Elimination

Input: Confidence level δ, representative budget L, and weights $\{\lambda_m\}_{m\in[M]}$

$\mathcal{B}_{m,1} \leftarrow [K]$ for all m; $\mathcal{C}_1 \leftarrow \mathcal{S}_L$
for $r = 1, 2, \ldots$ **do**
  if $|\mathcal{C}_r| = 1$ **then**
    return the unique set in $\mathcal{C}_r$
  end if
  $\mathcal{U}_r \leftarrow (\bigcup_{S\in\mathcal{C}_r} S) \cup (\bigcup_{m=1}^{M} \mathcal{B}_{m,r})$
  Sample every $a \in \mathcal{U}_r$ until it has n_r observations and compute $\widehat{\mu}_{a,m}^{(r)}$
  for $m \in [M]$ **do**
    $\widehat{\mu}_{m,r}^{\star} = \max_{b\in\mathcal{B}_{m,r}} \widehat{\mu}_{b,m}^{(r)}$
    $\mathcal{B}_{m,r+1} \leftarrow \{a \in \mathcal{B}_{m,r} : \widehat{\mu}_{a,m}^{(r)} \geq \widehat{\mu}_{m,r}^{\star} - 2\epsilon_r\}$
  end for
  for $S \in \mathcal{C}_r$ **do**
    Compute $\underline{\Phi}_r(S)$ and $\overline{\Phi}_r(S)$ using Eq. (10)
  end for
  $\mathcal{C}_{r+1} \leftarrow \{S \in \mathcal{C}_r : \underline{\Phi}_r(S) \leq \min_{S'\in\mathcal{C}_r} \overline{\Phi}_r(S')\}$
end for

and the minimum set-level gap as

$$\Delta_L(S) = \Phi(S) - \Phi(S_L^\star), \qquad \Delta_{L,\min} = \min_{S\in\mathcal{S}_L\setminus\{S_L^\star\}} \Delta_L(S) > 0.$$

Let $\overline{\lambda} = \max_{m\in[M]} \lambda_m$ and define the set-identification radius as $\rho_L = \Delta_{L,\min}/(8\overline{\lambda})$. The effective gap of each base arm $a \in [K]$ is defined as

$$\Gamma_{L,a} = \min\left\{\max\left\{\rho_L, \min_{\substack{S\in\mathcal{S}_L\\ a\in S}} \frac{\Delta_L(S)}{8\overline{\lambda}}\right\}, \min_{m\in[M]} \max\left\{\rho_L, \frac{g_{m,a}}{4}\right\}\right\}. \tag{11}$$

The first term in Eq. (11) determines when all candidate sets containing arm a can be eliminated, whereas the second determines when arm a is no longer needed for benchmark estimation. The truncation at ρ_L reflects the fact that both roles become irrelevant once the optimal representative set has been identified. Hence, $\Gamma_{L,a}$ determines the last phase in which arm a is sampled.

Theorem 2 (Instance-dependent sample complexity)**.** *With probability at least $1-\delta$, Algorithm 2 returns the optimal representative set $S_L^\star$ using at most $T_{\mathrm{AFM}}(\boldsymbol{\mu}, \delta)$ arm pulls, where*

$$T_{\mathrm{AFM}}(\boldsymbol{\mu}, \delta) = \widetilde{O}\left(\sigma^2 \sum_{a=1}^{K} \frac{1}{\Gamma_{L,a}^2} \log\frac{KM}{\delta}\right). \tag{12}$$

Theorem 2 shows that AFM-ELIM remains adaptive at the base-arm level even though it operates over a combinatorial family of candidate sets. An arm that appears only in clearly suboptimal candidate sets is discarded early, whereas an arm involved in difficult set comparisons or in benchmark estimation is sampled for longer. Moreover, each base-arm observation is shared across all active candidate sets containing that arm, so no separate sampling cost is incurred for individual sets. The identified representative set can further be used to recover the latent objective groups; a confidence-based assignment procedure and its sample complexity are given in Appendix F.3.

Computationally Efficient Relaxation. Exact latent-set identification in AFM-ELIM requires optimizing over the combinatorial family $\mathcal{S}_L = \{S \subseteq [K] : |S| = L\}$, which is computationally intractable in general. To obtain a polynomial-time alternative that preserves the worst-objective coverage criterion, we consider a bicriteria relaxation that allows $O(L \log M)$ representative arms. Specifically, with probability at least $1-\delta$, the relaxed procedure returns a set $\widehat{S}$ satisfying

$$\Phi(\widehat{S}) \leq \Phi(S_L^\star) + \epsilon, \qquad |\widehat{S}| \leq \lceil H_M L \rceil, \qquad H_M = \sum_{j=1}^{M} \frac{1}{j} \leq 1 + \log M.$$

The key to computational efficiency is that, at any fixed loss threshold, feasibility reduces to a set-cover problem over the objectives. Here, a loss threshold specifies the largest weighted degradation from the objective-wise ideal value that we are willing to tolerate on each objective, and an arm covers exactly those objectives whose weighted degradation does not exceed the threshold. Building on this observation, we propose BIAFM-GREEDY, which first estimates the reward matrix by uniform sampling, then scans candidate thresholds in increasing order and applies the greedy set-cover rule at each one. It returns the first cover that satisfies the relaxed budget, thereby avoiding enumeration of all size-L candidate sets. Compared with AFM-ELIM, this procedure forgoes exact size-L identification and instance-adaptive sampling, but in return offers a polynomial-time bicriteria guarantee: with high probability, its coverage loss is within ϵ of the optimal size-L loss while using at most $O(L \log M)$ arms. The formal procedure and its theoretical guarantee are given in Appendix B.

4 INFORMATION-THEORETIC LOWER BOUND

The preceding algorithms come with explicit finite-time guarantees, but these guarantees do not characterize the fundamental difficulty. In this section, we establish an information-theoretic lower bound on the sample complexity of identifying the optimal representative set $S_L^\star$.

We consider Gaussian rewards, where each pull of arm $a \in [K]$ yields an independent observation $\boldsymbol{r}_t \sim \mathcal{N}(\boldsymbol{\mu}_a, \sigma^2 \mathbf{I}_M)$. The Kullback–Leibler (KL) divergence between the reward distributions with mean vectors $\boldsymbol{\mu}_a$ and $\boldsymbol{\nu}_a$ is $d(\boldsymbol{\mu}_a, \boldsymbol{\nu}_a) = \frac{\|\boldsymbol{\mu}_a - \boldsymbol{\nu}_a\|_2^2}{2\sigma^2}$. Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ denote a bandit instance, and let $S_L^\star(\boldsymbol{\mu})$ denote its optimal representative set, which we assume to be unique. For a fixed instance $\boldsymbol{\mu}$, the set of alternative instances is

$$\mathrm{Alt}_L(\boldsymbol{\mu}) = \{\boldsymbol{\nu} : S_L^\star(\boldsymbol{\nu}) \neq S_L^\star(\boldsymbol{\mu})\}. \tag{13}$$

Each $\boldsymbol{\nu} \in \mathrm{Alt}_L(\boldsymbol{\mu})$ induces a different optimal representative set and must therefore be distinguished from $\boldsymbol{\mu}$ by any correct identification algorithm. The corresponding characteristic time is defined by

$$(T_L^\star(\boldsymbol{\mu}))^{-1} = \sup_{\boldsymbol{w} \in \Delta_K} \inf_{\boldsymbol{\nu} \in \mathrm{Alt}_L(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\boldsymbol{\mu}_a, \boldsymbol{\nu}_a). \tag{14}$$

Here, the outer optimization determines the asymptotic sampling proportions, while the inner optimization selects the hardest alternative under the chosen allocation.

Theorem 3 (Information-theoretic lower bound)**.** *Let $\delta \in (0, 1/2)$. Any fixed-confidence algorithm that identifies the optimal representative set with probability at least $1-\delta$ on every Gaussian instance with a unique optimal representative set satisfies*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T_L^\star(\boldsymbol{\mu}) \, \mathrm{kl}(1-\delta, \delta), \tag{15}$$

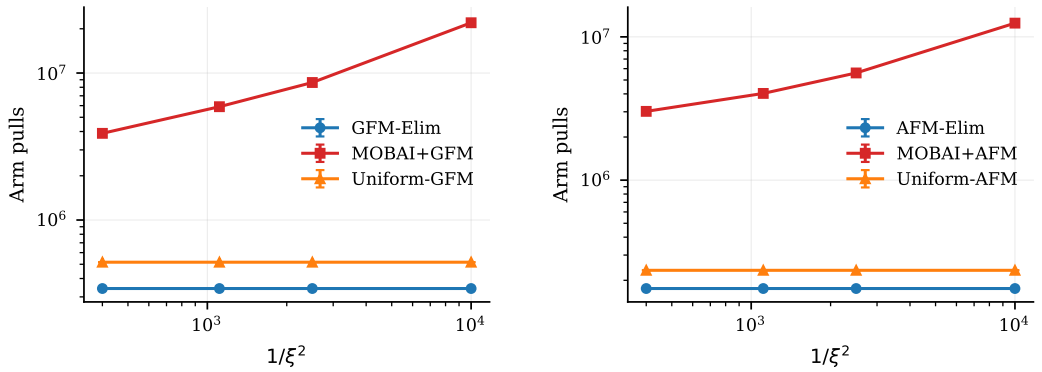
where $\mathrm{kl}(\cdot, \cdot)$ denotes the binary relative entropy. Consequently,

$$\liminf_{\delta \downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(1/\delta)} \geq T_L^\star(\boldsymbol{\mu}). \tag{16}$$

Theorem 3 shows that any δ-PAC algorithm requires, in expectation, at least order $\log(1/\delta)$ samples, with $T_L^\star(\boldsymbol{\mu})$ quantifying the intrinsic difficulty of the instance. The max–min form in Eq. (14) admits the standard fixed-confidence interpretation: the learner chooses asymptotic sampling proportions $\boldsymbol{w}$, and the inner optimization selects the hardest alternative instance that changes the optimal representative set. The resulting weighted KL divergence gives the maximal information rate, whose inverse is the leading constant of the asymptotic sample complexity. Unlike in objective-wise MO-BAI (Chen et al., 2025), the lower bound only requires distinguishing alternatives that change $S_L^\star$; objective-wise near-ties are therefore irrelevant unless they alter the optimal representative set. An oracle-based Track-and-Stop algorithm that asymptotically attains this bound with high probability is presented in Appendix C.

5 EXPERIMENTS

In this section, we evaluate the proposed few-for-many identification framework from both statistical and practical perspectives. We first examine whether it improves sample efficiency by avoiding unnecessary objective-wise distinctions and by sharing vector-valued observations across objectives. We then assess its practical effectiveness on a real-world task.



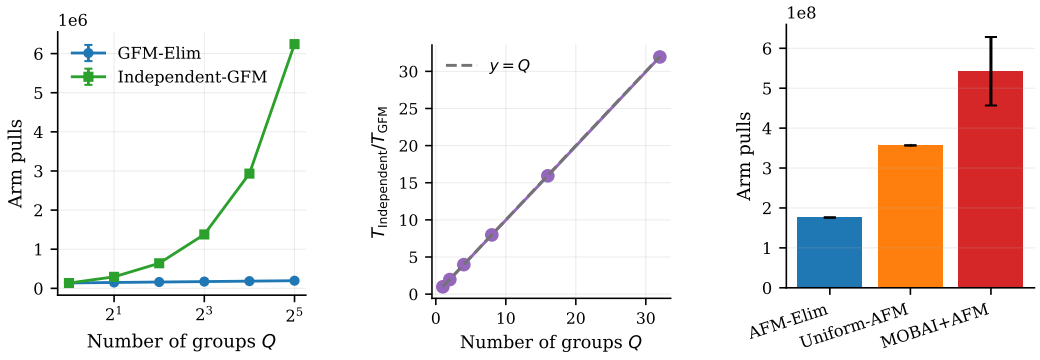
(a) Known groups (b) Latent groups

Figure 1: Sample cost on synthetic instances with irrelevant objective-wise near-ties and dominated arms. Points show means over 100 trials and error bars indicate 95% CIs; the horizontal axis is $1/\xi^2$.

Synthetic setup. All synthetic experiments use Gaussian rewards with $\sigma = 1$ and 100 independent trials, with random seeds shared across methods. We report the mean number of arm pulls with 95% confidence intervals, and set $\delta = 0.05$ unless otherwise specified. We compare against three types of baselines: uniform-sampling variants of our algorithms; INDEPENDENT-GFM, which solves each group separately with confidence level δ/Q; and the two-stage MOBAI+GFM/AFM, which first identifies the objective-wise best arms using the method of Chen et al. (2025) and then performs few-for-many identification. The two stages split the confidence budget equally, and the second stage reuses the first-stage samples. For Figure 1, the known-group instance has $Q = 4$ groups with four objectives each, so $M = 16$; the latent-group instance has $M = 9$ objectives generated from three planted blocks of size three and uses the representative budget $L = 3$, but the partition is not revealed to the algorithm. These instances are constructed so that objective-wise identification becomes harder as ξ decreases, while the optimal representatives remain unchanged. Each group or planted block has a common representative with mean 0.8, each objective has a specialist with mean 1 and a near-specialist with mean $1 - \xi$, and the remaining arms are dominated with mean 0.05. Decreasing ξ therefore isolates distinctions that are difficult for MO-BAI but irrelevant to few-for-many identification. To evaluate the benefit of sharing vector observations across groups, we further use a two-arm construction with $M = Q$ singleton groups and vary $Q \in \{1, 2, 4, 8, 16, 32\}$.

Synthetic results. All methods identify the correct representatives in every run. In Figure 1, the two-stage MO-BAI baselines require substantially more samples as ξ decreases, because distinguishing each specialist from its near-specialist becomes increasingly difficult. In contrast, GFM-ELIM and AFM-ELIM are nearly insensitive to ξ, since these near-ties do not affect the optimal representatives. This illustrates the advantage of targeting the few-for-many solution directly and avoiding unnecessary objective-wise distinctions. Figure 1 also highlights the benefit of adaptive elimination: unlike the uniform variants, which keep sampling all arms, the proposed methods discard arms once they become irrelevant to benchmark or representative identification, and consistently require fewer samples. Figures 2(a)–(b) further demonstrate the benefit of shared vector feedback. Because a single pull reveals all objective coordinates, GFM-ELIM reuses the same observations across groups, whereas INDEPENDENT-GFM solves each group separately. As a result, the relative gain grows approximately linearly in Q and reaches about $32\times$ at $Q = 32$, consistent with our analysis.

Real-world setup. We evaluate the proposed methods on the Adult dataset (Kohavi, 1996). We construct 12 arms by combining three classifiers (logistic regression, random forest, and histogram gradient boosting) and four decision thresholds in $\{0.35, 0.45, 0.55, 0.65\}$. The 15 objectives are test accuracies on demographic subgroups defined by sex, age, race, and sex–age combinations, after excluding two subgroups whose best arms are tied. Each pull returns subgroup-level correctness indicators, which are sub-Gaussian with $\sigma = 1/2$. We set $L = 2$, for which the optimal representative set consists of the two histogram-gradient-boosting models with thresholds 0.45 and 0.55.



(a) Sample cost (b) Sharing gain (c) Adult, $L = 2$

Figure 2: (a)–(b) Sample cost and sharing gain on the shared-feedback synthetic instance; the dashed line in (b) marks the linear reference $y = Q$. (c) Sample cost on the Adult model-portfolio task; all methods identify the same optimal pair in every trial.

We run 100 independent trials and compare AFM-ELIM with Uniform-AFM and MOBAI+AFM, using the same MO-BAI implementation as in the synthetic experiments.

Real-world results. As shown in Figure 2(c), all methods identify the optimal two-model portfolio in every trial, but their sampling costs differ substantially. Uniform-AFM samples every model throughout the elimination process and thus cannot exploit the fact that several classifiers are quickly revealed to be irrelevant to the best size-two portfolio. AFM-ELIM eliminates such arms and candidate sets earlier, reducing the cost by roughly a factor of two. The MO-BAI-based baseline is even more expensive, because its first stage attempts to identify the best model for every subgroup before solving the portfolio problem. On Adult, several subgroup accuracies are close across neighboring thresholds, so resolving every subgroup-wise winner requires many additional samples. The few-for-many objective, in contrast, only needs to determine which pair jointly covers the subgroups well, and can therefore stop without resolving all objective-wise winners. Consequently, AFM-ELIM uses only about one third of the pulls of MOBAI+AFM. Overall, the real-world experiment supports the same conclusion as the synthetic study: directly targeting the deployable representative set can be much cheaper than solving the stronger objective-wise identification problem.

Appendix D provides further empirical results, including confidence calibration, objective scaling, representative-gap sensitivity, and a computational study of the bicriteria approximation algorithm.

6 CONCLUSION AND FUTURE WORK

We formulated few-for-many best-arm identification as a fixed-confidence problem whose goal is to find a small set of arms such that every objective is well served by at least one selected arm. We considered two settings: known objective groups, where one representative is identified for each group, and latent objective groups, where a budgeted representative set is selected without a predefined partition. For these settings, we developed GFM-ELIM and AFM-ELIM, which exploit vector-valued feedback and eliminate arms as soon as they become irrelevant to the final decision. We established instance-dependent sample-complexity guarantees for both algorithms and an information-theoretic lower bound for the latent-group setting, and further showed that this lower bound is asymptotically attainable with high probability by an oracle-based Track-and-Stop procedure. Our experiments demonstrate that directly targeting the few-for-many solution can substantially reduce the sample complexity by sharing observations and avoiding unnecessary objective-wise distinctions.

An important direction for future work is to better reconcile statistical efficiency with computational scalability in the latent-group setting, where exact identification involves combinatorial optimization. It would also be of interest to extend the framework beyond the current setting, for example, to adaptive selection of the representative budget and to partial feedback.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, LaTeX consistency checks, and drafting of experiment code. All mathematical definitions, theorem statements, proofs, algorithmic claims, experimental designs, and final interpretations were reviewed and validated by the authors, who take full responsibility for the content of this work, including any text, code, or figures prepared with AI assistance.

REPRODUCIBILITY STATEMENT

The problem setup, assumptions, algorithms, and main theoretical guarantees are stated in the main text, and complete proofs and additional theoretical details are provided in the appendix. The experimental section specifies the synthetic instance constructions, objective-group structures, reward models, confidence levels, evaluation metrics, and baselines. For the real-world experiment, it also describes the Adult model-portfolio construction, the fixed train–test split, and the preprocessing and model configurations needed to reproduce the reported results. The code and complete experimental materials, including instance generators, implementations of all compared methods, plotting scripts, raw trial outputs, processed summaries, and validation checks, will be made publicly available upon acceptance.

REFERENCES

Cagin Ararat and Cem Tekin. Vector optimization with stochastic bandit feedback. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 2165–2190, 2023.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pp. 41–53, 2010.

Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina Drugan. Pareto front identification from stochastic bandit feedback. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 939–947, 2016.

Lijie Chen, Anupam Gupta, and Jian Li. Pure exploration of multi-armed bandit under matroid constraints. In *Proceedings of the 29th Annual Conference on Learning Theory*, pp. 647–669, 2016.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems 27*, pp. 379–387, 2014.

Zhirui Chen, P. N. Karthik, Yeow Meng Chee, and Vincent Y. F. Tan. Optimal multi-objective best arm identification with fixed confidence. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, pp. 1162–1170, 2025.

Ji Cheng, Bo Xue, Jiaxiang Yi, and Qingfu Zhang. Hierarchize pareto dominance in multi-objective stochastic linear bandits. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 11489–11497, 2024.

Élise Crepon, Aurélien Garivier, and Wouter M Koolen. Sequential learning of the Pareto front for multi-objective bandits. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 3583–3591, 2024.

Udvas Das, Apurv Shukla, and Debabrota Basu. FraPPE: Fast and efficient preference-based pure exploration. In *Advances in Neural Information Processing Systems 38*, pp. 10190–10236, 2025.

Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems 32*, pp. 14564–14573, 2019.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2432–2442, 2020.

Madalina M. Drugan and Ann Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint Conference on Neural Networks*, pp. 1–8, 2013.

Tanner Fiez, Lalit Jain, Kevin G. Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems 32*, pp. 10667–10677, 2019.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems 25*, pp. 3212 – 3220, 2012.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory*, pp. 998–1027, 2016.

Tong Geng, Xiliang Lin, and Harikesh S. Nair. Online evaluation of audiences for targeted advertising via bandit experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pp. 13273–13279, 2020.

Ping Guo, Tiantian Zhang, Xi Lin, Xiang Li, Zhi-Ri Tang, and Qingfu Zhang. Few-for-many personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17515–17524, 2026.

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

Marc Jourdan, Mojmír Mutný, Johannes Kirschner, and Andreas Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 805–849, 2021.

Xynthia Kavelaars, Joris Mulder, and Maurits Kaptein. Decision-making with multiple correlated binary outcomes in clinical trials. *Statistical Methods in Medical Research*, 29(11):3265–3277, 2020.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 202–207, 1996.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Adaptive algorithms for relaxed pareto set identification. In *Advances in Neural Information Processing Systems 36*, 2023.

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

Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth tchebycheff scalarization for multi-objective optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 30479–30509, 2024.

Xi Lin, Yilu Liu, Xiaoyuan Zhang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Few for many: Tchebycheff set scalarization for many-objective optimization. In *The 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=O4N9kWwV6R.

Yilu Liu, Chengyu Lu, Xi Lin, and Qingfu Zhang. Many-objective cover problem: Discovering few solutions to cover many objectives. In *Parallel Problem Solving from Nature*, pp. 68–82, 2024.

Yilu Liu, Xi Lin, Liang Zhao, and Qingfu Zhang. Few for many: Towards efficient and flexible many-objective optimization. *IEEE Transactions on Evolutionary Computation*, 2025.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. A flexible framework for multi-objective bayesian optimization using random scalarizations. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, pp. 766–776, 2020.

Somangchan Park, Heesang Ann, and Min hwan Oh. Thompson sampling for multi-objective linear contextual bandit. In *Advances in Neural Information Processing Systems 38*, pp. 84523–84555, 2025.

Eric Schwartz, Eric Bradlow, and Peter Fader. Customer acquisition via display advertising using multi-armed bandit experiments. In *Marketing Science*, pp. 500–522, 2017.

Ke Shang, Hisao Ishibuchi, Zexuan Zhu, and Qingfu Zhang. An efficient evolutionary algorithm for few-for-many optimization. *IEEE/CAA Journal of Automatica Sinica*, 13(6):1362–1377, 2026.

Apurv Shukla and Debabrota Basu. Preference-based pure exploration. In *Advances in Neural Information Processing Systems 37*, pp. 17313–17347, 2024.

Eralp Turgay, Doruk Oner, and Cem Tekin. Multi-objective contextual bandit problem with similarity information. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pp. 1673–1681, 2018.

Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2):199 – 215, 2015.

Mengfan Xu and Diego Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *Proceedings of the 40th International Conference on International Conference on Machine Learning*, pp. 38499–38517, 2023.

Bo Xue, Ji Cheng, Fei Liu, Yimu Wang, and Qingfu Zhang. Multiobjective lipschitz bandits under lexicographic ordering. *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 16238 – 16246, 2024.

Qiuyi (Richard) Zhang. Optimal scalarizations for sublinear hypervolume regret. In *Advances in Neural Information Processing Systems 37*, pp. 39963–39999, 2024.

A Extended Related Work

In this section, we review three lines of research most relevant to our work: fixed-confidence best-arm identification, multi-objective bandits, and few-for-many optimization.

Fixed-confidence best-arm identification. Fixed-confidence best-arm identification (BAI) seeks to identify the optimal arm with a prescribed confidence level while minimizing the number of samples (Audibert & Bubeck, 2010; Gabillon et al., 2012). A major line of work studies its instance-dependent sample complexity and develops sampling strategies that attain the corresponding information-theoretic limits. In the classical setting, Garivier & Kaufmann (2016) established an asymptotically tight lower bound and proposed Track-and-Stop, which adaptively tracks the optimal sampling proportions. This framework has since been extended to more structured pure-exploration problems, including combinatorial identification (Chen et al., 2014; 2016; Jourdan et al., 2021), identification with multiple correct answers (Degenne & Koolen, 2019), and linear and transductive bandits (Fiez et al., 2019; Degenne et al., 2020; Jedra & Proutiere, 2020). These studies show that the structure of the answer space and of the feedback model plays a central role in determining both the optimal sampling allocation and the resulting sample complexity.

Multi-objective bandits. Multi-objective bandits extend classical bandit models to vector-valued rewards and have been studied extensively in both online decision making and pure exploration (Drugan & Nowe, 2013). In the online setting, one line of work targets Pareto-efficient decisions without imposing an explicit scalar preference, including contextual Pareto regret (Turgay et al., 2018), Pareto regret under stochastic and adversarial feedback (Xu & Klabjan, 2023), and Pareto-regret minimization in linear contextual bandits (Park et al., 2025). Another line incorporates user preferences through scalarizations or structured orderings. Representative examples include nonlinear scalarizations for exploring the Pareto front with hypervolume guarantees (Zhang, 2024), as well as lexicographic and Pareto–lexicographic bandits that optimize objectives according to prescribed priority structures (Xue et al., 2024; Cheng et al., 2024).

In pure exploration, most existing work focuses on identifying the Pareto-optimal set under fixed-budget or fixed-confidence criteria (Auer et al., 2016; Ararat & Tekin, 2023; Kone et al., 2023; Crepon et al., 2024). More recent studies incorporate preference information through ordering cones, which induce partial orders over vector-valued rewards and encode structured trade-offs among objectives (Shukla & Basu, 2024; Das et al., 2025). Most closely related to our work, Chen et al. (2025) study fixed-confidence multi-objective best-arm identification, in which the learner identifies the best arm for each objective, and characterize asymptotically optimal sampling. In contrast, few-for-many identification seeks only a small set of representative arms that collectively serve all objectives. Consequently, it need not recover the entire Pareto set or resolve every objective-wise best arm, but only those distinctions that can affect the optimal representatives.

Few-for-many optimization. The few-for-many paradigm has emerged as an alternative to approximating a large Pareto set when only a small number of solutions can ultimately be deployed. Liu et al. (2024) formulate the many-objective cover problem (MaCP), which selects $k \ll M$ solutions such that each objective is well served by at least one of them, and propose a clustering-based method for the resulting NP-hard problem. Building on this idea, Lin et al. (2025) introduce Tchebycheff set scalarization, which evaluates a set through its worst objective-wise performance, together with a smooth variant amenable to gradient-based optimization. Liu et al. (2025) further develop a conditional set-optimization framework that supports different scalarizations and both continuous and discrete many-objective problems. Subsequent studies improve the computational efficiency of few-for-many optimization (Shang et al., 2026) and extend the paradigm to personalized federated learning, where a small collection of shared models is optimized to serve heterogeneous clients (Guo et al., 2026).

These works primarily address deterministic optimization, where objective evaluations or optimization feedback are directly available. In contrast, we study few-for-many identification under statistical uncertainty, where the vector-valued means of the candidate arms are unknown and must be inferred from noisy sequential observations.

Algorithm 3 GREEDYCOVER

Input: Coverage sets $\{C_a\}_{a\in[K]}$ and universe $[M]$

$S \leftarrow \emptyset, \mathcal{U} \leftarrow [M]$, and $\mathcal{A} \leftarrow [K]$
while $\mathcal{U} \neq \emptyset$ **do**
  Choose $a' \in \arg\max_{a\in\mathcal{A}} |C_a \cap \mathcal{U}|$
  if $|C_{a'} \cap \mathcal{U}| = 0$ **then**
    return FAIL
  end if
  $S \leftarrow S \cup \{a'\}, \mathcal{U} \leftarrow \mathcal{U} \setminus C_{a'}$, and $\mathcal{A} \leftarrow \mathcal{A} \setminus \{a'\}$
end while
return S

Algorithm 4 BIAFM-GREEDY: Bicriteria Approximate Few-for-Many Identification

Input: Accuracy $\epsilon > 0$, confidence $\delta \in (0,1)$, budget L, and weights $\{\lambda_m\}_{m\in[M]}$

$\overline{\lambda} \leftarrow \max_{m\in[M]} \lambda_m$ and $n \leftarrow \left\lceil \frac{32\sigma^2\overline{\lambda}^2}{\epsilon^2} \log\left(\frac{2KM}{\delta}\right) \right\rceil$
Pull every arm n times and compute $\widehat{\mu}_{a,m}$ for all a, m
$\widehat{\mu}^\star_m \leftarrow \max_{a\in[K]} \widehat{\mu}_{a,m}$ for every $m \in [M]$
$\widehat{d}_{a,m} \leftarrow \lambda_m(\widehat{\mu}^\star_m - \widehat{\mu}_{a,m})$ for all a, m
$\widehat{\mathcal{T}} \leftarrow \{\widehat{d}_{a,m} : a \in [K], m \in [M]\}$, sorted increasingly after removing duplicates
$B_L \leftarrow \min\{K, \lceil H_M L \rceil\}$
for each $\gamma \in \widehat{\mathcal{T}}$ in increasing order **do**
  $\widehat{C}_a(\gamma) \leftarrow \{m \in [M] : \widehat{d}_{a,m} \leq \gamma\}$ for every $a \in [K]$
  $S_\gamma \leftarrow$ GREEDYCOVER$(\{\widehat{C}_a(\gamma)\}_{a\in[K]}, [M])$
  if $S_\gamma \neq$ FAIL and $|S_\gamma| \leq B_L$ **then**
    return $\widehat{S} \leftarrow S_\gamma$
  end if
end for

B POLYNOMIAL-TIME BICRITERIA APPROXIMATE IDENTIFICATION

The exact AFM-ELIM algorithm identifies $S^\star_L$ with an instance-dependent sample complexity, but it maintains a combinatorial family of candidate sets. This section presents BIAFM-GREEDY, a complementary polynomial-time procedure. The price of computational tractability is a bicriteria relaxation: the returned set may contain $O(L \log M)$ arms, while its coverage loss remains within ϵ of the optimal size-L loss.

The key observation is that, for a fixed target loss, the latent few-for-many problem reduces to a set-cover problem over the objectives. For each arm $a \in [K]$ and objective $m \in [M]$, define the weighted deficit $d_{a,m} = \lambda_m(\mu^\star_m - \mu_{a,m})$. The latent-set loss can then be written as

$$\Phi(S) = \max_{m\in[M]} \min_{a\in S} d_{a,m}. \tag{17}$$

For any threshold $\gamma \geq 0$, define the set of objectives covered by arm a as

$$C_a(\gamma) = \{m \in [M] : d_{a,m} \leq \gamma\}. \tag{18}$$

It follows that

$$\Phi(S) \leq \gamma \quad \Longleftrightarrow \quad \bigcup_{a\in S} C_a(\gamma) = [M]. \tag{19}$$

Hence, for a fixed threshold γ, finding a minimum-cardinality representative set with loss at most γ is equivalent to finding a minimum set cover of the objectives.

We therefore use the standard greedy set-cover procedure in Algorithm 3. Starting with all objectives uncovered, it repeatedly selects the arm that covers the largest number of currently uncovered objectives. Whenever a feasible cover exists, the classical greedy guarantee gives

$$|S_{\text{greedy}}| \leq H_M \ \text{OPT}_{\text{SC}}, \qquad H_M = \sum_{j=1}^{M} \frac{1}{j} \leq 1 + \log M, \tag{20}$$

where $\mathrm{OPT}_{\mathrm{SC}}$ denotes the minimum cover size.

The complete BIAFM-GREEDY procedure is summarized in Algorithm 4. It first samples all arms uniformly to estimate the unknown means, and then computes the corresponding empirical ideal values and weighted deficits. Instead of enumerating the candidate family $\mathcal{S}_L$, the algorithm uses the distinct empirical deficits as candidate thresholds. For each threshold, it constructs the induced empirical coverage sets and applies GREEDYCOVER; the first threshold whose greedy cover satisfies the relaxed budget $B_L = \min\{K, \lceil H_M L \rceil\}$ determines the returned set. Algorithm 4 thus examines at most KM thresholds and never enumerates the $\binom{K}{L}$ candidate sets.

Theorem 4 (Polynomial-time bicriteria identification)**.** *Suppose that the reward noise is σ^2-sub-Gaussian, and let $\overline{\lambda} = \max_{m\in[M]} \lambda_m$. For any $\epsilon > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, Algorithm 4 returns a set $\widehat{S}$ such that*

$$|\widehat{S}| \le \min\{K, \lceil H_M L \rceil\}, \qquad \Phi(\widehat{S}) \le \Phi(S_L^\star) + \epsilon, \tag{21}$$

where $H_M = \sum_{j=1}^{M} 1/j \le 1 + \log M$. The total number of arm pulls satisfies

$$T_{\mathrm{BiAFM}} \le K \left\lceil \frac{32\sigma^2 \overline{\lambda}^{-2}}{\epsilon^2} \log\left(\frac{2KM}{\delta}\right) \right\rceil. \tag{22}$$

Moreover, the computational complexity of the algorithm is polynomial in K, M, L, $1/\epsilon$, and $\log(1/\delta)$.

Theorem 4 provides a finite-time bicriteria guarantee for the polynomial-time relaxation. Ignoring integer rounding, it is an (H_M, ϵ, δ)-PAC result: with probability at least $1-\delta$, the algorithm returns a set whose loss is within ϵ of that of the best size-L representative set, at the cost of a multiplicative relaxation of at most $H_M \le 1 + \log M$ in set size. Thus, BIAFM-GREEDY complements AFM-ELIM by trading exact cardinality for polynomial-time computation.

C ORACLE-BASED TRACK-AND-STOP

This section presents AFM-TAS, an oracle-based, asymptotically optimal counterpart of AFM-ELIM. Rather than providing a practical finite-time procedure, its purpose is to show that, under Gaussian rewards and given the required optimization oracles, the information-theoretic lower bound in Theorem 3 is asymptotically attainable with high probability. Following the Track-and-Stop framework (Garivier & Kaufmann, 2016; Chen et al., 2025), AFM-TAS repeatedly estimates the optimal representative set, tracks the corresponding optimal sampling allocation, and stops once sufficient evidence separates the current answer from all alternatives.

The optimal sampling proportions are characterized by the max–min problem in Eq. (14). For an instance $\boldsymbol{\mu}$, define

$$\mathcal{W}_L^\star(\boldsymbol{\mu}) = \arg \max_{\boldsymbol{w}\in\Delta_K} \inf_{\boldsymbol{\nu}\in \mathrm{Alt}_L(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\boldsymbol{\mu}_a, \boldsymbol{\nu}_a). \tag{23}$$

At round t, let $\widehat{\boldsymbol{\mu}}(t)$ denote the empirical mean instance. The current empirical answer is

$$\widehat{S}_t \in \arg\min_{S\in\mathcal{S}_L} \Phi_{\widehat{\boldsymbol{\mu}}(t)}(S), \tag{24}$$

and the statistical evidence in its favor is quantified by the generalized likelihood-ratio (GLR) statistic

$$Z_L(t) = \inf_{\boldsymbol{\nu}:\, S_L^\star(\boldsymbol{\nu}) \neq \widehat{S}_t} \sum_{a=1}^{K} N_a(t) d\left(\widehat{\boldsymbol{\mu}}_a(t), \boldsymbol{\nu}_a\right). \tag{25}$$

Thus, $Z_L(t)$ measures the minimum accumulated evidence needed to move from the empirical instance to an alternative under which $\widehat{S}_t$ is no longer optimal.

We instantiate the stopping threshold as

$$\beta(t,\delta) = \log\left(\frac{\pi^2 t^2 |\mathcal{S}_L|}{6\delta}\right), \tag{26}$$

Algorithm 5 AFM-TAS: Oracle-Based Adaptive Few-for-Many Track-and-Stop

Input: Confidence level δ and representative budget L

Pull every arm once
for $t = K, K+1, \ldots$ **do**
  Compute $\widehat{S}_t$ and $Z_L(t)$ using Eqs. (24)–(25)
  if $Z_L(t) \geq \beta(t, \delta)$ **then**
    return $\widehat{S}_t$
  end if
  Compute $\boldsymbol{w}_t \in \mathcal{W}_L^\star(\widehat{\boldsymbol{\mu}}(t))$
  if $\min_{a \in [K]} N_a(t) < \sqrt{t} - K/2$ **then**
    Pull an arm with the smallest number of observations
  else
    Pull $A_{t+1} \in \arg\min_{a \in [K]} \left[N_a(t) - \sum_{s=K}^{t} w_{s,a} \right]$
  end if
end for

where the factor $|\mathcal{S}_L|$ accounts for the finite answer space. By the standard GLR maximal inequality and a union bound over the candidate answers and over time (using $\sum_{t \geq 1} 6/(\pi^2 t^2) = 1$), this threshold satisfies

$$\mathbb{P}_{\boldsymbol{\mu}} \left(\exists t \geq 1 : \widehat{S}_t \neq S_L^\star(\boldsymbol{\mu}),\ Z_L(t) \geq \beta(t, \delta) \right) \leq \delta, \tag{27}$$

and, for fixed problem dimensions,

$$\beta(t, \delta) = \log \frac{1}{\delta} + o\left(\log \frac{1}{\delta} \right) \tag{28}$$

uniformly over $t = O(\log(1/\delta))$.

AFM-TAS first samples every arm once. At each subsequent round, it stops and returns $\widehat{S}_t$ if $Z_L(t) \geq \beta(t, \delta)$. Otherwise, it computes an allocation in $\mathcal{W}_L^\star(\widehat{\boldsymbol{\mu}}(t))$ and tracks it while enforcing sufficient exploration of every arm. This tracking rule drives the empirical sampling proportions toward the information-theoretically optimal allocation while ensuring consistent estimation of the underlying instance.

Theorem 5 (Oracle-based asymptotic upper bound)**.** *Suppose that the rewards are Gaussian and $T_L^\star(\boldsymbol{\mu}) < \infty$, and that the allocation oracle in Step 7 is exact. Then Algorithm 5 is δ-PAC and, for every $\eta > 0$, satisfies*

$$\lim_{\delta \downarrow 0} \mathbb{P}_{\boldsymbol{\mu}} \left(\tau_\delta \leq (1 + \eta) T_L^\star(\boldsymbol{\mu}) \log \frac{1}{\delta} \right) = 1. \tag{29}$$

Thus, Algorithm 5 attains the characteristic time $T_L^\star(\boldsymbol{\mu})$ with asymptotically high probability.

Theorem 5 shows that the oracle allocation is not merely a heuristic for focusing samples: with asymptotically high probability, the stopping time of AFM-TAS matches the characteristic time $T_L^\star(\boldsymbol{\mu})$ appearing in the lower bound of Theorem 3. By contrast, AFM-ELIM trades this asymptotic optimality for a simpler elimination rule and an explicit non-asymptotic guarantee. The two algorithms therefore play different roles: AFM-ELIM is the implementable finite-time procedure, whereas AFM-TAS certifies the information-theoretic benchmark.

D ADDITIONAL EXPERIMENTS

This appendix provides additional experiments that complement the main results in Section 5. We first examine confidence calibration, scaling with the number of objectives, and sensitivity to the gaps that directly determine the few-for-many answer. We then study the computational behavior of the polynomial-time bicriteria algorithm in Appendix B. We do not include AFM-TAS, since it is an oracle-based construction that requires solving the characteristic-time allocation problem exactly and repeatedly computing GLR stopping statistics against all alternative representative sets.

D.1 Confidence Calibration

We first examine whether the empirical error rate is consistent with the prescribed confidence level. We consider one known-group instance and one latent-group instance from the synthetic setting in Section 5, fix $\xi = 0.1$, and vary $\delta \in \{0.2, 0.1, 0.05, 0.02\}$. For each δ, we run 200 independent trials, using GFM-Elim on the known-group instance and AFM-Elim on the latent-group instance.

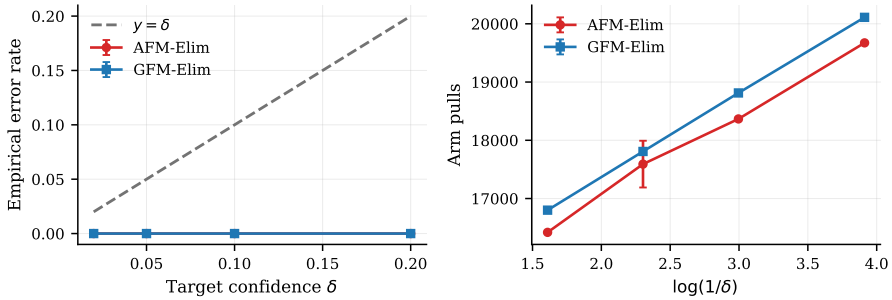


Figure 3: Confidence calibration. The left panel compares the empirical error rate with the target confidence level, and the right panel reports the corresponding sample complexity.

As shown in Figure 3, both algorithms return the correct answer in all trials at every tested confidence level, which is consistent with the conservative finite-time confidence allocation used by the elimination procedures. The sample complexity increases only mildly as δ decreases, in line with the logarithmic dependence on $1/\delta$ in the theoretical bounds. Hence, the algorithms remain reliable across confidence levels, and the sample budget grows gradually as the confidence requirement becomes stricter.

D.2 Scaling with the Number of Objectives

We next isolate the many-objective effect in the known-group setting. We fix $Q = 4$ groups and increase the number of objectives per group from 2 to 16, so that M ranges from 8 to 64. The optimal representative of each group remains unchanged, while every objective has its own specialist and near-specialist. Consequently, methods that first identify all objective-wise best arms must resolve a growing number of distinctions that are unnecessary for the few-for-many answer.

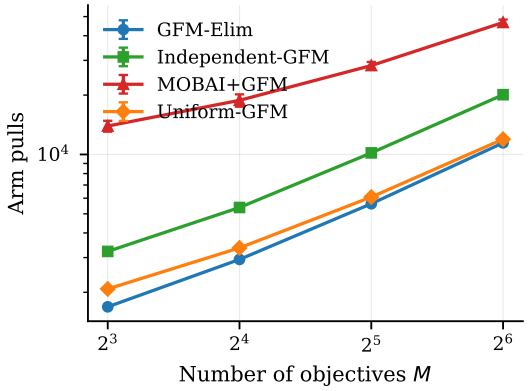


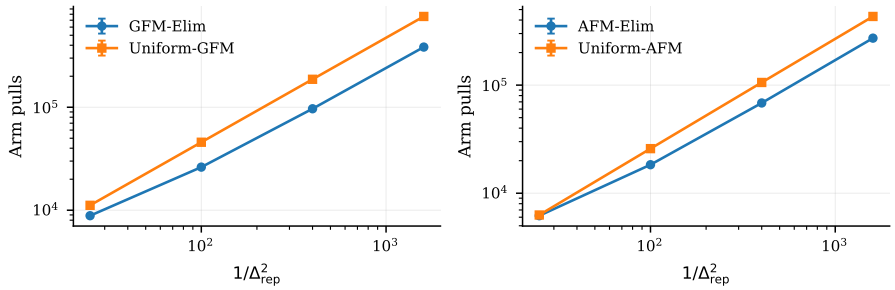
Figure 4: Scaling with the number of objectives in the known-group setting.

As shown in Figure 4, GFM-Elim consistently identifies the group representatives using substantially fewer pulls than the two-stage MOBAI+GFM baseline. The gap widens as M grows, because MOBAI+GFM spends many samples resolving objective-wise specialists before solving the

grouped problem. GFM-ELIM also improves over INDEPENDENT-GFM by sharing each vector observation across group decisions. Its curve approaches that of Uniform-GFM for larger M, because this experiment adds a specialist and a near-specialist arm for each new objective while keeping the number of easily discarded dominated arms fixed; hence, a growing fraction of arms remains relevant to benchmark estimation. These results confirm the main message of the synthetic study: as the number of objectives grows, directly identifying representatives is more sample-efficient than recovering every objective-wise winner.

D.3 SENSITIVITY TO REPRESENTATIVE GAPS

The main experiment in Figure 1 varies an objective-wise gap that is irrelevant to the answer. Here, we instead vary the gap between the optimal representative and its closest competitor, while keeping the near-specialist gap fixed at $\xi = 0.1$. This directly changes the difficulty of the few-for-many decision.



(a) Known groups (b) Latent groups

Figure 5: Sensitivity to the representative gap. Both known-group and latent-group identification become harder as the representative gap shrinks.

Figure 5 exhibits the expected quadratic growth in sample complexity as the representative gap decreases. In the known-group case, the average sample complexity of GFM-ELIM increases from the order of thousands to the order of hundreds of thousands as the representative gap shrinks from 0.2 to 0.025; the latent-group results for AFM-ELIM follow the same pattern. This contrasts with Figure 1: the proposed methods are insensitive to gaps that do not change the few-for-many answer, yet correctly spend more samples on gaps that determine the representative decision. Together, the two experiments indicate that the algorithms adapt to the right notion of difficulty.

D.4 COMPUTATIONAL SCALABILITY OF THE BICRITERIA ALGORITHM

We further evaluate the polynomial-time BIAFM-GREEDY algorithm. This experiment focuses on computational efficiency and complements, rather than replaces, the sample-complexity comparisons in Section 5. Unlike AFM-ELIM, which maintains the combinatorial candidate family to identify the exact size-L representative set, BIAFM-GREEDY uses uniform sampling and returns a representative set under the relaxed budget.

We use the latent version of the synthetic family from Section 5, with $Q = 4$ planted blocks of three objectives each ($M = 12$), $L = 4$, $\xi = 0.05$, $\sigma = 0.1$, and $\delta = 0.05$. We vary the number of dominated arms, which increases K and hence the number of exact candidate sets $|\mathcal{S}_L| = \binom{K}{L}$. For BIAFM-GREEDY, we use $\epsilon = 0.1$, which gives the relaxed budget $\lceil H_{12} L \rceil = 13$. Each configuration is averaged over 30 independent trials with common random seeds. Reported runtimes exclude the post-hoc computation of the true optimum, which is used only for evaluation.

K	Method	$\lvert\mathcal{S}_L\rvert$	Pulls	Time (ms)	$\lvert\widehat{S}\rvert$	$\Phi(\widehat{S})$
32	AFM-ELIM	35,960	2,144	42.9	4	0.20
32	BIAFM-GREEDY	35,960	9,888	0.35	12	0.00
48	AFM-ELIM	194,580	2,480	238.5	4	0.20
48	BIAFM-GREEDY	194,580	15,456	0.43	12	0.00
64	AFM-ELIM	635,376	2,784	782.8	4	0.20
64	BIAFM-GREEDY	635,376	21,184	0.51	12	0.00

Table 1: Supplementary latent-group experiment for the polynomial-time bicriteria procedure. AFM-ELIM returns the exact size-L optimum, whereas BIAFM-GREEDY avoids enumerating $\mathcal{S}_L$ and returns a bicriteria set within the relaxed budget.

Table 1 summarizes the statistical–computational trade-off. On this instance, exact AFM-ELIM is more sample-efficient: it respects the strict budget $L = 4$ and uses only the samples needed to certify the exact optimal representative set. However, its runtime grows quickly with K, because the candidate family expands from tens of thousands to hundreds of thousands of sets. In contrast, BIAFM-GREEDY samples all arms uniformly and therefore uses more pulls, but it never enumerates $\mathcal{S}_L$, and its runtime stays below one millisecond in all tested configurations. The returned set sizes and coverage losses reflect the bicriteria nature of the guarantee: BIAFM-GREEDY returns a larger set, still within the relaxed logarithmic budget, which attains zero true coverage loss in this construction, whereas AFM-ELIM returns the best strict size-L set, whose coverage loss is nonzero.

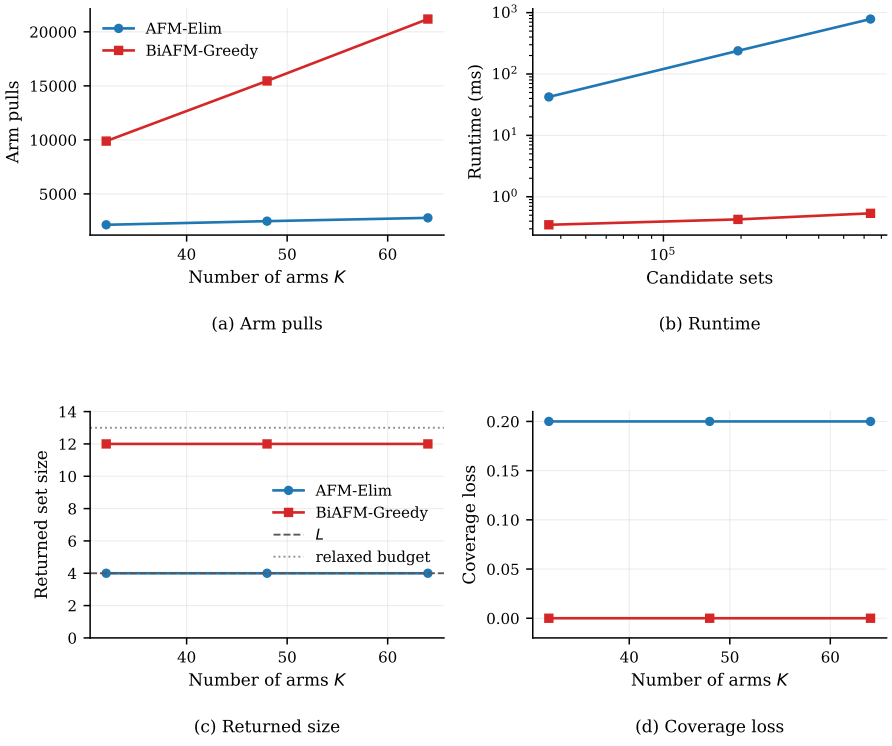


Figure 6: BIAFM-GREEDY versus AFM-ELIM. The four panels show arm pulls, runtime, returned set size, and true coverage loss as the exact candidate family grows.

Figure 6 visualizes these trends across the four measured quantities. The sample panel shows the price of the uniform exploration in BIAFM-GREEDY, while the runtime panel shows the benefit of replacing exact enumeration of candidate sets with a greedy cover computation. The set-size and coverage-loss panels make the approximation trade-off explicit: the bicriteria method returns more

arms, and this enlarged set covers all objectives better than the strict size-L optimum on this instance. Hence, BIAFM-GREEDY is useful when exact enumeration of candidate sets is the computational bottleneck and a modest relaxation of the budget is acceptable.

E ANALYSIS OF THE GROUPED METHOD: GFM-ELIM

E.1 CONFIDENCE BOUNDS AND CORRECTNESS

For every arm a, objective m, and phase r, let $\widehat{\mu}_{a,m}(n_r)$ denote the empirical mean of the first n_r samples of arm a on objective m. Consider the event

$$\mathcal{E} = \bigcap_{r=1}^{\infty} \bigcap_{a\in[K]} \bigcap_{m\in[M]} \{|\widehat{\mu}_{a,m}(n_r) - \mu_{a,m}| \le \epsilon_r\}. \tag{30}$$

Lemma 1. *Under Eq. (1), $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) \ge 1-\delta$.*

Proof. For every fixed r, a, and m, the sub-Gaussian concentration inequality gives

$$\mathbb{P}_{\boldsymbol{\mu}}\left(|\widehat{\mu}_{a,m}(n_r) - \mu_{a,m}| > \epsilon_r\right) \le 2\exp\left(-\frac{n_r\epsilon_r^2}{2\sigma^2}\right) \le \frac{\delta_r}{KM}.$$

A union bound over all arms and objectives yields

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\exists a\in[K], m\in[M] : |\widehat{\mu}_{a,m}(n_r) - \mu_{a,m}| > \epsilon_r\right) \le \delta_r.$$

Finally,

$$\sum_{r=1}^{\infty}\delta_r = \frac{6\delta}{\pi^2}\sum_{r=1}^{\infty}\frac{1}{r^2} = \delta,$$

and a union bound over phases completes the proof. □

Lemma 2. *On the event $\mathcal{E}$, the following statements hold at every phase r.*

1. *The best arm of every objective remains in the corresponding benchmark set:*
$$b_m^\star \in \mathcal{B}_{m,r}, \qquad \forall m\in[M].$$
2. *For every unresolved group q and every $a\in\mathcal{C}_{q,r}$,*
$$\underline{R}_{q,r}(a) \le R_q(a) \le \overline{R}_{q,r}(a). \tag{31}$$
3. *The loss bounds satisfy*
$$R_q(a) - 4\overline{\lambda}_q\epsilon_r \le \underline{R}_{q,r}(a) \le \overline{R}_{q,r}(a) \le R_q(a) + 4\overline{\lambda}_q\epsilon_r. \tag{32}$$

Proof. We first prove the benchmark-set property by induction. Initially, $\mathcal{B}_{m,1} = [K]$, so the claim holds. Suppose that $b_m^\star \in \mathcal{B}_{m,r}$. On $\mathcal{E}$,

$$\widehat{\mu}^{(r)}_{b_m^\star,m} \ge \mu_m^\star - \epsilon_r, \qquad \widehat{\mu}^\star_{m,r} \le \mu_m^\star + \epsilon_r.$$

Hence, $\widehat{\mu}^{(r)}_{b_m^\star,m} \ge \widehat{\mu}^\star_{m,r} - 2\epsilon_r$, and thus $b_m^\star \in \mathcal{B}_{m,r+1}$.

Since $b_m^\star \in \mathcal{B}_{m,r}$, the definition of the maximum and the event $\mathcal{E}$ imply

$$\left|\widehat{\mu}^\star_{m,r} - \mu_m^\star\right| \le \epsilon_r. \tag{33}$$

Consequently, for every arm a,

$$\left|\left(\widehat{\mu}^\star_{m,r} - \widehat{\mu}^{(r)}_{a,m}\right) - (\mu_m^\star - \mu_{a,m})\right| \le 2\epsilon_r.$$

Since the positive-part operator is nondecreasing and 1-Lipschitz, and $\mu_m^\star - \mu_{a,m} \ge 0$, subtracting and adding $2\epsilon_r$ as in Eq. (7) and taking the weighted maximum over $m\in G_q$ proves Eq. (31). The same display also yields

$$\underline{R}_{q,r}(a) \ge R_q(a) - 4\overline{\lambda}_q\epsilon_r, \qquad \overline{R}_{q,r}(a) \le R_q(a) + 4\overline{\lambda}_q\epsilon_r,$$

which proves Eq. (32). □

Lemma 3 (Elimination correctness)**.** *Algorithm 1 satisfies*

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\tau < \infty, \widehat{\boldsymbol{a}}_{\mathcal{G}} = \boldsymbol{a}^{\star}_{\mathcal{G}}\right) \geq 1 - \delta.$$

More precisely, on the event $\mathcal{E}$:

1. *$a_q^\star$ is never eliminated from $\mathcal{C}_{q,r}$;*
2. *every $a \neq a_q^\star$ is eliminated from $\mathcal{C}_{q,r}$ at the first phase r at which*

$$\epsilon_r < \frac{\Delta_{q,a}}{8\overline{\lambda}_q}$$

and a is still active in group q.

Proof. Fix a group q. By Lemma 2,

$$\underline{R}_{q,r}(a_q^\star) \leq R_q(a_q^\star).$$

Moreover, for every $b \in \mathcal{C}_{q,r}$,

$$\overline{R}_{q,r}(b) \geq R_q(b) \geq R_q(a_q^\star).$$

Therefore,

$$\underline{R}_{q,r}(a_q^\star) \leq \min_{b \in \mathcal{C}_{q,r}} \overline{R}_{q,r}(b),$$

so $a_q^\star$ is never eliminated.

Now consider an arm $a \neq a_q^\star$. By Eq. (32),

$$\underline{R}_{q,r}(a) \geq R_q(a) - 4\overline{\lambda}_q \epsilon_r$$

and, since $a_q^\star \in \mathcal{C}_{q,r}$,

$$\min_{b \in \mathcal{C}_{q,r}} \overline{R}_{q,r}(b) \leq \overline{R}_{q,r}(a_q^\star) \leq R_q(a_q^\star) + 4\overline{\lambda}_q \epsilon_r.$$

Hence,

$$\underline{R}_{q,r}(a) - \min_{b \in \mathcal{C}_{q,r}} \overline{R}_{q,r}(b) \geq \Delta_{q,a} - 8\overline{\lambda}_q \epsilon_r.$$

If $\epsilon_r < \frac{\Delta_{q,a}}{8\overline{\lambda}_q}$, the right-hand side is strictly positive, and arm a is eliminated. Since all suboptimal group gaps are positive and $\epsilon_r \to 0$, every group eventually contains only $a_q^\star$ on $\mathcal{E}$. Combined with $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) \geq 1 - \delta$ from Lemma 1, this proves the lemma. □

E.2 Proof of Theorem 1

Proof. We work on the event $\mathcal{E}$ throughout. Fix a group q. By Lemma 3, every suboptimal representative candidate is eliminated once

$$\epsilon_r < \frac{\Delta_{q,a}}{8\overline{\lambda}_q}.$$

In particular, all suboptimal candidates are eliminated once $\epsilon_r < \rho_q$, so group q is resolved no later than the first phase satisfying $\epsilon_r < \rho_q$.

Consider first the representative role of arm a in group q. If $a \neq a_q^\star$, it is eliminated once

$$\epsilon_r < \frac{\Delta_{q,a}}{8\overline{\lambda}_q}.$$

If $a = a_q^\star$, it remains active until group q is resolved. In both cases, this role has ended by the first phase satisfying

$$\epsilon_r < \max\left\{\rho_q, \frac{\Delta_{q,a}}{8\overline{\lambda}_q}\right\}.$$

Next, consider the benchmark role of arm a for an objective $m \in G_q$. If $g_{m,a} > 4\epsilon_r$, then on $\mathcal{E}$,

$$\widehat{\mu}_{a,m}^{(r)} \leq \mu_{a,m} + \epsilon_r < \mu_m^\star - 3\epsilon_r \leq \widehat{\mu}_{m,r}^\star - 2\epsilon_r,$$

so a is eliminated from $\mathcal{B}_{m,r+1}$. Otherwise, the benchmark role still ends once group q is resolved, since m then leaves $\mathcal{M}_r$. Therefore, this role has ended by the first phase satisfying

$$\epsilon_r < \max\left\{\rho_q, \frac{g_{m,a}}{4}\right\}.$$

By Eq. (8), all roles of arm a have ended by the phase

$$r_a = \min\{r \geq 1 : \epsilon_r < \Gamma_a\},$$

so arm a is pulled at most n_{r_a} times.

By the minimality of r_a, $\epsilon_{r_a} \geq \frac{\Gamma_a}{2}$ and $r_a \leq 1 + \left\lceil \log_2 \frac{1}{\Gamma_a} \right\rceil$. Using Eq. (6),

$$\begin{aligned} n_{r_a} &\leq 1 + \frac{2\sigma^2}{\epsilon_{r_a}^2} \log\left(\frac{\pi^2 K M r_a^2}{3\delta}\right) \\ &\leq 1 + \frac{8\sigma^2}{\Gamma_a^2} \log\left(\frac{\pi^2 K M}{3\delta}\left(1 + \left\lceil \log_2 \frac{1}{\Gamma_a} \right\rceil\right)^2\right). \end{aligned}$$

Consequently, on $\mathcal{E}$, the total number of pulls is at most

$$\sum_{a=1}^{K}\left[1 + \frac{8\sigma^2}{\Gamma_a^2} \log\left(\frac{\pi^2 K M}{3\delta}\left(1 + \left\lceil \log_2 \frac{1}{\Gamma_a} \right\rceil\right)^2\right)\right],$$

which is of the order stated in Eq. (9). Since $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) \geq 1 - \delta$ by Lemma 1, the high-probability guarantee follows. □

F Analysis of the Adaptive Method: AFM-Elim

F.1 Set-level confidence bounds and correctness

Let $\mathcal{E}$ be the simultaneous confidence event defined in Eq. (30). We first show that the loss bounds in Eq. (10) are valid uniformly over all candidate sets. Importantly, this result does not require a union bound over the $\binom{K}{L}$ candidate sets.

Lemma 4 (Set-level confidence bounds)**.** *On the event $\mathcal{E}$, the following statements hold at every phase r.*

1. *The best arm of every objective remains in its benchmark set:*
$$b_m^\star \in \mathcal{B}_{m,r}, \qquad \forall m \in [M].$$
2. *For every $S \in \mathcal{C}_r$ and $m \in [M]$, the empirical objective-wise value satisfies*
$$\left|\widehat{V}_{m,r}(S) - V_m(S)\right| \leq \epsilon_r. \tag{34}$$
3. *For every $S \in \mathcal{C}_r$,*
$$\underline{\Phi}_r(S) \leq \Phi(S) \leq \overline{\Phi}_r(S). \tag{35}$$
4. *The set-level loss bounds satisfy*
$$\Phi(S) - 4\overline{\lambda}\epsilon_r \leq \underline{\Phi}_r(S) \leq \overline{\Phi}_r(S) \leq \Phi(S) + 4\overline{\lambda}\epsilon_r. \tag{36}$$

Proof. The benchmark-set statement follows from the same induction argument as in Lemma 2, so $b_m^\star$ is never removed from $\mathcal{B}_{m,r}$. On $\mathcal{E}$, the nonexpansiveness of the maximum operator gives

$$\left|\widehat{V}_{m,r}(S) - V_m(S)\right| \leq \max_{a \in S}\left|\widehat{\mu}_{a,m}^{(r)} - \mu_{a,m}\right| \leq \epsilon_r,$$

which proves Eq. (34). Because the best arm remains in $\mathcal{B}_{m,r}$, Eq. (33) also holds. Therefore,

$$\left|\left(\widehat{\mu}_{m,r}^\star - \widehat{V}_{m,r}(S)\right) - (\mu_m^\star - V_m(S))\right| \leq 2\epsilon_r.$$

The monotonicity and 1-Lipschitz continuity of the positive-part operator in Eq. (10) then yield Eq. (35), as well as

$$\underline{\Phi}_r(S) \geq \Phi(S) - 4\overline{\lambda}\epsilon_r, \qquad \overline{\Phi}_r(S) \leq \Phi(S) + 4\overline{\lambda}\epsilon_r,$$

which proves Eq. (36). □

The set-level confidence event depends only on the KM base-arm means. Consequently, the confidence radius scales logarithmically in KM rather than in the exponentially large number of candidate sets.

Lemma 5 (Set-elimination correctness)**.** *Algorithm 2 is δ-PAC. More precisely, on the event $\mathcal{E}$:*

1. *the optimal representative set $S_L^\star$ is never eliminated;*
2. *every $S \neq S_L^\star$ that is still active is eliminated at the first phase satisfying*

$$\epsilon_r < \frac{\Delta_L(S)}{8\overline{\lambda}}. \tag{37}$$

Therefore,

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\tau < \infty, \widehat{S}_L = S_L^\star\right) \geq 1 - \delta.$$

Proof. Let $B_r = \min_{S' \in \mathcal{C}_r} \overline{\Phi}_r(S')$ denote the elimination threshold at phase r. By Lemma 4,

$$\underline{\Phi}_r(S_L^\star) \leq \Phi(S_L^\star),$$

and, for every $S \in \mathcal{C}_r$,

$$\overline{\Phi}_r(S) \geq \Phi(S) \geq \Phi(S_L^\star).$$

It follows that

$$\underline{\Phi}_r(S_L^\star) \leq B_r,$$

so $S_L^\star$ is never eliminated.

Now fix a suboptimal set $S \neq S_L^\star$. By Eq. (36),

$$\underline{\Phi}_r(S) \geq \Phi(S) - 4\overline{\lambda}\epsilon_r,$$

and, since $S_L^\star \in \mathcal{C}_r$,

$$B_r \leq \overline{\Phi}_r(S_L^\star) \leq \Phi(S_L^\star) + 4\overline{\lambda}\epsilon_r.$$

Therefore,

$$\underline{\Phi}_r(S) - B_r \geq \Delta_L(S) - 8\overline{\lambda}\epsilon_r.$$

Whenever Eq. (37) holds, the right-hand side is strictly positive, and hence S is removed from $\mathcal{C}_{r+1}$.

Since $\Delta_{L,\min} > 0$ and $\epsilon_r \to 0$, every suboptimal candidate set is eventually eliminated on $\mathcal{E}$. Combined with $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) \geq 1 - \delta$ from Lemma 1, this proves the result. □

F.2 Proof of Theorem 2

Proof. We work on the event $\mathcal{E}$ throughout. Fix a base arm a, and consider first its role in the active candidate family.

If $a \notin S_L^\star$, every candidate set containing a is suboptimal. By Lemma 5, each such set S is eliminated once

$$\epsilon_r < \frac{\Delta_L(S)}{8\overline{\lambda}}.$$

Therefore, all candidate sets containing a have been eliminated once

$$\epsilon_r < \min_{\substack{S \in \mathcal{S}_L \\ a \in S}} \frac{\Delta_L(S)}{8\overline{\lambda}}.$$

If $a \in S_L^\star$, then a remains in the union of active candidate sets until the algorithm terminates. Since every suboptimal set is eliminated once $\epsilon_r < \rho_L$, this role ends once

$$\epsilon_r < \rho_L.$$

Next, consider the benchmark role of arm a for objective m. If

$$g_{m,a} > 4\epsilon_r,$$

then, on $\mathcal{E}$,

$$\widehat{\mu}_{a,m}^{(r)} \le \mu_{a,m} + \epsilon_r < \mu_m^\star - 3\epsilon_r \le \widehat{\mu}_{m,r}^\star - 2\epsilon_r,$$

so a is removed from $\mathcal{B}_{m,r+1}$. Otherwise, this role ends when the algorithm terminates. Hence, the benchmark role has ended by the first phase satisfying

$$\epsilon_r < \max\left\{\rho_L, \frac{g_{m,a}}{4}\right\}.$$

By Eq. (11), all roles of arm a have ended by the phase

$$r_a = \min\left\{r \ge 1 : \epsilon_r < \Gamma_{L,a}\right\},$$

so the arm is pulled at most n_{r_a} times. By the minimality of r_a,

$$\epsilon_{r_a} \ge \frac{\Gamma_{L,a}}{2}, \qquad r_a \le 1 + \left\lceil \log_2 \frac{1}{\Gamma_{L,a}} \right\rceil.$$

Substituting these inequalities into the definition of n_r gives

$$n_{r_a} \le 1 + \frac{8\sigma^2}{\Gamma_{L,a}^2} \log\left(\frac{\pi^2 KM}{3\delta}\left(1 + \left\lceil \log_2 \frac{1}{\Gamma_{L,a}} \right\rceil\right)^2\right).$$

Consequently, on $\mathcal{E}$, the total number of pulls is at most

$$\sum_{a=1}^{K}\left[1 + \frac{8\sigma^2}{\Gamma_{L,a}^2} \log\left(\frac{\pi^2 KM}{3\delta}\left(1 + \left\lceil \log_2 \frac{1}{\Gamma_{L,a}} \right\rceil\right)^2\right)\right],$$

which is of the order stated in Eq. (12). Since $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) \ge 1 - \delta$ by Lemma 1, the high-probability guarantee follows. □

F.3 Recovery of the latent objective groups

Identifying $S_L^\star$ does not by itself determine a unique objective partition. To recover a unique assignment, we impose the following condition.

Assumption 1. *For every objective $m \in [M]$, there exists a unique arm $\pi_L^\star(m) \in S_L^\star$ satisfying*

$$\pi_L^\star(m) = \arg\max_{a \in S_L^\star} \mu_{a,m}.$$

The latent group represented by $a \in S_L^\star$ is then

$$G_a^\star = \left\{m \in [M] : \pi_L^\star(m) = a\right\}. \tag{38}$$

When $L = 1$, the assignment is trivial and requires no additional samples. We therefore assume $L \ge 2$ below and define the assignment gap

$$\zeta_m = \mu_{\pi_L^\star(m),m} - \max_{a \in S_L^\star \setminus \{\pi_L^\star(m)\}} \mu_{a,m}, \tag{39}$$

together with

$$\zeta_{\min} = \min_{m \in [M]} \zeta_m.$$

After $S_L^\star$ has been identified, the learner can continue sampling its L arms until, for every objective m, there exists an arm $\widehat{\pi}(m) \in S_L^\star$ such that

$$\widehat{\mu}_{\widehat{\pi}(m),m}^{(r)} - \max_{a \in S_L^\star \setminus \{\widehat{\pi}(m)\}} \widehat{\mu}_{a,m}^{(r)} > 2\epsilon_r. \tag{40}$$

The output groups are

$$\widehat{G}_a = \left\{m \in [M] : \widehat{\pi}(m) = a\right\}, \qquad a \in S_L^\star.$$

Corollary 1 (Latent-group recovery)**.** *Suppose that $L \geq 2$ and that Assumption 1 holds. On the event $\mathcal{E}$, condition Eq. (40) is satisfied for all objectives once*

$$\epsilon_r < \frac{\zeta_{\min}}{4},$$

and the resulting assignments and groups satisfy

$$\widehat{\pi}(m) = \pi_L^\star(m), \qquad \widehat{G}_a = G_a^\star.$$

The additional sample complexity is at most

$$\widetilde{O}\left(\frac{\sigma^2 L}{\zeta_{\min}^2} \log \frac{KM}{\delta}\right). \tag{41}$$

Proof. For every objective m, on $\mathcal{E}$,

$$\widehat{\mu}^{(r)}_{\pi_L^\star(m),m} - \max_{a \in S_L^\star \setminus \{\pi_L^\star(m)\}} \widehat{\mu}^{(r)}_{a,m} \geq \zeta_m - 2\epsilon_r.$$

When $\epsilon_r < \zeta_{\min}/4$, the right-hand side exceeds $2\epsilon_r$, so Eq. (40) holds with $\widehat{\pi}(m) = \pi_L^\star(m)$. Conversely, on $\mathcal{E}$, any arm $\widehat{\pi}(m)$ satisfying Eq. (40) has a true mean strictly larger than those of all other arms in $S_L^\star$, so the certified assignment is correct. At the first phase with $\epsilon_r < \zeta_{\min}/4$, the same calculation as in the proof of Theorem 1 gives $n_r = \widetilde{O}(\sigma^2 \zeta_{\min}^{-2} \log(KM/\delta))$. Sampling the L selected arms to this precision proves Eq. (41). □

G Proof of Theorem 3

Proof. If $\mathbb{E}_{\boldsymbol{\mu}}[\tau] = +\infty$, the lower bound holds trivially. We therefore assume $\mathbb{E}_{\boldsymbol{\mu}}[\tau] < +\infty$.

Let

$$E = \left\{\widehat{S}_L = S_L^\star(\boldsymbol{\mu})\right\}.$$

Since the algorithm is δ-PAC,

$$\mathbb{P}_{\boldsymbol{\mu}}(E) \geq 1 - \delta.$$

For every $\boldsymbol{\nu} \in \mathrm{Alt}_L(\boldsymbol{\mu})$, the set $S_L^\star(\boldsymbol{\mu})$ is not optimal under $\boldsymbol{\nu}$, and hence

$$\mathbb{P}_{\boldsymbol{\nu}}(E) \leq \delta.$$

The change-of-measure inequality for adaptive sampling gives

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)] d(\boldsymbol{\mu}_a, \boldsymbol{\nu}_a) \geq \mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\mu}}(E), \mathbb{P}_{\boldsymbol{\nu}}(E)\right) \geq \mathrm{kl}(1-\delta, \delta),$$

where the last step uses the monotonicity of kl and $\delta < 1/2$. Let

$$w_a = \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}.$$

Then $\boldsymbol{w} \in \Delta_K$, and since the previous inequality holds for every alternative,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \inf_{\boldsymbol{\nu} \in \mathrm{Alt}_L(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\boldsymbol{\mu}_a, \boldsymbol{\nu}_a) \geq \mathrm{kl}(1-\delta, \delta).$$

By the definition of $T_L^\star(\boldsymbol{\mu})$,

$$\inf_{\boldsymbol{\nu} \in \mathrm{Alt}_L(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\boldsymbol{\mu}_a, \boldsymbol{\nu}_a) \leq \left(T_L^\star(\boldsymbol{\mu})\right)^{-1}.$$

Combining the last two inequalities proves Eq. (15). Finally, since

$$\mathrm{kl}(1-\delta, \delta) \sim \log \frac{1}{\delta}$$

as $\delta \downarrow 0$, Eq. (16) follows. □

H Proof of Theorem 4

Proof. Set $\eta = \epsilon/(4\overline{\lambda})$ and define

$$\mathcal{E}_{\mathrm{Bi}} = \{|\widehat{\mu}_{a,m} - \mu_{a,m}| \le \eta,\ \forall a \in [K], m \in [M]\}.$$

For each arm–objective pair, the sub-Gaussian concentration inequality gives

$$\mathbb{P}_{\boldsymbol{\mu}}\left(|\widehat{\mu}_{a,m} - \mu_{a,m}| > \eta\right) \le 2\exp\left(-\frac{n\eta^2}{2\sigma^2}\right).$$

The prescribed value of n and $\eta^2 = \epsilon^2/(16\overline{\lambda}^2)$ imply

$$\frac{n\eta^2}{2\sigma^2} \ge \log\left(\frac{2KM}{\delta}\right).$$

Hence, each failure probability is at most $\delta/(KM)$, and a union bound yields $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_{\mathrm{Bi}}) \ge 1 - \delta$.

On $\mathcal{E}_{\mathrm{Bi}}$, the nonexpansiveness of the maximum operator gives, for every m,

$$|\widehat{\mu}^\star_m - \mu^\star_m| = \left|\max_a \widehat{\mu}_{a,m} - \max_a \mu_{a,m}\right| \le \max_a |\widehat{\mu}_{a,m} - \mu_{a,m}| \le \eta.$$

Consequently, for every a and m,

$$|\widehat{d}_{a,m} - d_{a,m}| \le \lambda_m\left(|\widehat{\mu}^\star_m - \mu^\star_m| + |\widehat{\mu}_{a,m} - \mu_{a,m}|\right) \le 2\overline{\lambda}\eta = \frac{\epsilon}{2}. \tag{42}$$

Define the empirical loss

$$\widehat{\Phi}(S) = \max_{m\in[M]} \min_{a\in S} \widehat{d}_{a,m}. \tag{43}$$

Since both the minimum and the maximum operators are nonexpansive in the sup norm, Eq. (42) implies, simultaneously for every nonempty $S \subseteq [K]$,

$$|\widehat{\Phi}(S) - \Phi(S)| \le \frac{\epsilon}{2}. \tag{44}$$

Let $\gamma^\star_{\mathrm{emp}} = \widehat{\Phi}(S^\star_L)$. For each objective m, some arm in $S^\star_L$ has an empirical deficit of at most $\gamma^\star_{\mathrm{emp}}$, so the L arms in $S^\star_L$ cover $[M]$ at this threshold. Thus,

$$\mathrm{OPT}_{\mathrm{SC}}(\gamma^\star_{\mathrm{emp}}) \le L, \qquad |S_{\gamma^\star_{\mathrm{emp}}}| \le H_M L. \tag{45}$$

Moreover, because both extrema in Eq. (43) are taken over finite sets, there exist $m' \in [M]$ and $a' \in S^\star_L$ such that $\gamma^\star_{\mathrm{emp}} = \widehat{d}_{a',m'}$; hence $\gamma^\star_{\mathrm{emp}} \in \widehat{\mathcal{T}}$. Algorithm 4 scans the thresholds in increasing order, and Eq. (45), together with the fact that a greedy cover uses at most K distinct arms, shows that the acceptance test is passed no later than at $\gamma^\star_{\mathrm{emp}}$. The returned threshold $\widehat{\gamma}$ therefore satisfies $\widehat{\gamma} \le \gamma^\star_{\mathrm{emp}}$.

Because $\widehat{S}$ covers every objective at threshold $\widehat{\gamma}$, we have $\widehat{\Phi}(\widehat{S}) \le \widehat{\gamma}$. Combining this with Eq. (44) gives

$$\Phi(\widehat{S}) \le \widehat{\Phi}(\widehat{S}) + \frac{\epsilon}{2} \le \widehat{\gamma} + \frac{\epsilon}{2} \le \widehat{\Phi}(S^\star_L) + \frac{\epsilon}{2} \le \Phi(S^\star_L) + \epsilon.$$

The budget bound follows directly from the acceptance test, and the preceding argument holds on an event of probability at least $1 - \delta$.

The algorithm uses exactly Kn arm pulls, which gives Eq. (22). There are at most KM distinct thresholds. Under a direct implementation, computing the empirical means costs $O(KMn)$ scalar operations, sorting the thresholds costs $O(KM\log(KM))$, and constructing all coverage sets and running the greedy cover at every threshold costs at most $O(K^3M^2)$. Hence, the overall computation is polynomial in the stated parameters. □

I PROOF OF THEOREM 5

Proof. The δ-PAC property follows directly from Eq. (27): an incorrect recommendation can occur only if, for some t,

$$\widehat{S}_t \neq S_L^\star(\boldsymbol{\mu}) \quad \text{and} \quad Z_L(t) \geq \beta(t,\delta),$$

and this event has probability at most δ.

Forced exploration guarantees that

$$N_a(t) \longrightarrow \infty, \qquad \forall a \in [K].$$

Therefore, by the strong law of large numbers,

$$\widehat{\boldsymbol{\mu}}_a(t) \longrightarrow \boldsymbol{\mu}_a \quad \text{almost surely.}$$

Since $\Delta_{L,\min} > 0$, the empirical optimal set eventually coincides with the true optimal set, i.e.,

$$\widehat{S}_t = S_L^\star$$

for all sufficiently large t, almost surely.

For $\boldsymbol{w} \in \Delta_K$ and an instance $\boldsymbol{\theta}$, define the information rate

$$I_L(\boldsymbol{w}, \boldsymbol{\theta}) = \inf_{\boldsymbol{\nu} \in \mathrm{Alt}_L(\boldsymbol{\theta})} \sum_{a=1}^{K} w_a d(\boldsymbol{\theta}_a, \boldsymbol{\nu}_a).$$

Let $\boldsymbol{w}(t) = \mathbf{N}(t)/t$. Once the empirical answer has stabilized, Eq. (25) gives

$$\frac{Z_L(t)}{t} = I_L\big(\boldsymbol{w}(t), \widehat{\boldsymbol{\mu}}(t)\big).$$

The exact allocation oracle and the tracking step of Algorithm 5 constitute precisely the Track-and-Stop tracking rule applied to the latent answer space. The standard tracking argument therefore implies that, along the realized trajectory, the empirical sampling proportions approach the optimal allocation set of the true instance. Consequently, we may choose $\boldsymbol{w}_t^\star \in \mathcal{W}_L^\star(\boldsymbol{\mu})$ such that $\|\boldsymbol{w}(t) - \boldsymbol{w}_t^\star\| \to 0$. Together with the strong law and the standard continuity argument for Track-and-Stop statistics, this yields

$$\big|I_L\big(\boldsymbol{w}(t), \widehat{\boldsymbol{\mu}}(t)\big) - I_L(\boldsymbol{w}_t^\star, \boldsymbol{\mu})\big| \longrightarrow 0 \quad \text{almost surely.}$$

Since every $\boldsymbol{w}_t^\star$ maximizes the information rate, $I_L(\boldsymbol{w}_t^\star, \boldsymbol{\mu}) = 1/T_L^\star(\boldsymbol{\mu})$. Therefore,

$$\frac{Z_L(t)}{t} \longrightarrow \frac{1}{T_L^\star(\boldsymbol{\mu})} \quad \text{almost surely.}$$

Hence, for every $\eta \in (0,1)$, there exists an almost surely finite random time t_η such that

$$Z_L(t) \geq \frac{(1-\eta)t}{T_L^\star(\boldsymbol{\mu})}$$

for all $t \geq t_\eta$.

Let

$$t_\delta = \left\lceil \frac{1+2\eta}{1-\eta} T_L^\star(\boldsymbol{\mu}) \log \frac{1}{\delta} \right\rceil.$$

By Eq. (28),

$$\beta(t_\delta, \delta) \leq (1+\eta) \log \frac{1}{\delta}$$

for all sufficiently small δ. Therefore, on the event $\{t_\eta \leq t_\delta\}$,

$$Z_L(t_\delta) \geq (1+2\eta) \log \frac{1}{\delta} > \beta(t_\delta, \delta)$$

for all sufficiently small δ. Since $t_\delta \to \infty$ and $t_\eta < \infty$ almost surely, $\mathbb{P}_{\boldsymbol{\mu}}(t_\eta \leq t_\delta) \to 1$. It follows that

$$\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta \leq t_\delta) \longrightarrow 1.$$

Since $(1+2\eta)/(1-\eta)$ can be made arbitrarily close to one by choosing η sufficiently small, relabeling the constant proves Eq. (29). □