WHEN MORE OBJECTIVES MAKE IDENTIFICATION EASIER: THE ROLE OF PREFERENCES AND FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Best-arm identification (BAI) aims to identify the optimal arm using as few samples as possible. While most existing work focuses on a single objective, many real-world applications involve multiple criteria, which motivates the study of BAI in multi-objective bandits. Intuitively, multi-objective identification seems harder, since the learner must identify optimal arms in a higher-dimensional reward space; however, the additional objectives also provide richer feedback that can reduce the required exploration. In this paper, we characterize when and why multiple objectives can make BAI easier. For Gaussian bandits, we establish a characteristic-time lower bound for fixed-confidence identification that holds uniformly across preference orders. Through explicit comparisons under Pareto, linear, and lexicographic preferences, we reveal three phenomena: (i) identifying a larger optimal set can require fewer samples than identifying a unique optimum; (ii) the advantage of vector feedback over coordinate feedback depends fundamentally on the preference order; and (iii) multi-objective identification can require fewer samples than even the easiest corresponding single-objective problem. Finally, we propose Order-Aware Track-and-Stop (OATS), an algorithm that is correct at any prescribed confidence level and asymptotically matches the lower bound.

1 INTRODUCTION

Best-arm identification (BAI) is a fundamental pure-exploration problem that aims to identify the optimal arm using as few samples as possible (Even-Dar et al., 2006; Bubeck et al., 2009; Garivier & Kaufmann, 2016). Classical BAI considers scalar rewards and defines the optimal arm as the one with the largest mean reward (Audibert et al., 2010; Gabillon et al., 2012; Karnin et al., 2013; Kaufmann et al., 2016). However, many applications involve multiple and potentially conflicting criteria, such as accuracy, cost, and latency in LLM evaluation (Liang et al., 2023; Chang et al., 2024). This motivates multi-objective BAI, in which each arm yields a reward vector and optimality is determined by a preference relation (Roijers et al., 2013; Auer et al., 2016; Ararat & Tekin, 2023; Shukla & Basu, 2024). For example, linear scalarization compares weighted rewards (Drugan & Nowe, 2013), lexicographic preferences rank objectives by priority (Hüyük & Tekin, 2021), and Pareto dominance may admit multiple maximal arms (Auer et al., 2016). Multi-objective BAI therefore comprises distinct identification tasks whose complexity depends on the preference relation.

Multi-objective identification may appear harder than its scalar counterpart, as its complexity depends not only on reward gaps but also on the geometry of the reward vectors (Karagözlü et al., 2024). Under Pareto preferences, the learner may even need to identify the entire set of nondominated arms (Kone et al., 2023). On the other hand, multiple objectives also provide richer feedback: each arm pull typically reveals all reward coordinates, and evidence accumulated across objectives can reduce the required exploration (Xue et al., 2026). This tension motivates our central question: *When do multiple objectives make best-arm identification harder, and when can they make it easier?*

Existing studies provide algorithms and complexity guarantees for Pareto identification (Auer et al., 2016), preference-cone identification (Das et al., 2025), and lexicographic identification with cross-objective constraints (Xue et al., 2026). However, these results are derived under different answer maps, model classes, and feedback models, which makes direct comparisons difficult. A unified analysis should therefore disentangle three factors: the preference order, which determines how arms are compared and hence what must be identified; the model structure, which specifies how

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.

rewards may vary across arms and objectives; and the feedback model, which determines what each sample reveals (Garivier & Kaufmann, 2016; Fiez et al., 2019; Degenne et al., 2020; Du et al., 2021). Their interaction, rather than the number of objectives alone, determines the sample complexity.

To address this gap, we adopt the characteristic time as a common measure of identification difficulty (Garivier & Kaufmann, 2016; Degenne & Koolen, 2019). For Gaussian bandits, it is governed by the optimal rate at which sampling can distinguish the true instance from feasible alternatives with a different correct answer. We construct matched comparisons in which only one of the three factors (preference order, model structure, or feedback model) varies at a time. This allows us to isolate the effect of each factor on sample complexity and to quantify when, and by how much, multiple objectives facilitate identification. Our main contributions are as follows.

- We establish a characteristic-time lower bound for fixed-confidence multi-objective BAI under general preference-induced answer maps. This provides a unified information-theoretic basis for comparing identification tasks across preference orders, model structures, and feedback models.
- Through explicit comparisons under Pareto and lexicographic preferences, we construct instances in which identifying multiple optimal arms is easier than identifying a unique optimum, as well as instances in which the reverse holds. Hence, neither the preference order nor the number of returned arms alone determines identification difficulty.
- We characterize the advantage of vector feedback over coordinate feedback and establish conditions under which multi-objective identification requires fewer samples than the best corresponding single-objective benchmark. These results clarify when observations of multiple objectives accumulate into useful evidence rather than being redundant.
- We propose Order-Aware Track-and-Stop (OATS) and prove its fixed-confidence correctness and asymptotic optimality. Hence, the information-theoretic gains revealed by our comparisons are achievable by an adaptive sampling algorithm.

2 Related Work

We review two closely related lines of research: best-arm identification and multi-objective bandits.

Best-arm identification. BAI is a pure-exploration problem studied in both the fixed-budget and the fixed-confidence settings. Early work developed elimination- and confidence-based methods with gap-dependent guarantees for identifying the best arm (Even-Dar et al., 2006; Bubeck et al., 2009; Audibert et al., 2010; Kalyanakrishnan et al., 2012; Gabillon et al., 2012; Karnin et al., 2013; Jamieson et al., 2014). A complementary information-theoretic line characterizes instance difficulty through alternative models with different correct answers, which yields fundamental lower bounds and asymptotically optimal frequentist or Bayesian sampling rules (Kaufmann et al., 2016; Garivier & Kaufmann, 2016; Russo, 2020). This framework has been extended to problems with multiple valid answers and general answer maps, where the optimal allocation of samples can be formulated as a sequential game (Degenne & Koolen, 2019; Degenne et al., 2019). Work on combinatorial, linear, and transductive pure exploration further exploits structural dependencies and shared observations through optimization oracles and experimental-design principles (Chen et al., 2014; Soare et al., 2014; Fiez et al., 2019; Degenne et al., 2020).

Multi-objective bandits. Multi-objective bandits associate each arm with a reward vector and define desirable decisions through a preference relation. Early regret-minimization studies considered Pareto dominance and scalarized utilities, and developed methods that learn nondominated arms or optimize aggregate objectives (Drugan & Nowe, 2013; Busa-Fekete et al., 2017). Subsequent work extended multi-objective regret analysis to contextual, adversarial, and structured linear settings, including dominant-objective and mixed Pareto–lexicographic preferences (Turğay et al., 2018; Tekin & Turğay, 2018; Xu & Klabjan, 2023; Cheng et al., 2024). A complementary pure-exploration line studies Pareto-set identification, with confidence-region, adaptive-sampling, and elimination methods for approximate, relaxed, fixed-budget, and exact fixed-confidence recovery (Auer et al., 2016; Kone et al., 2023; 2024; Crépon et al., 2024). More general formulations replace coordinatewise dominance with cone-induced, hierarchical, or lexicographic orders, and characterize how

preference geometry and structural relations shape identification complexity (Ararat & Tekin, 2023; Karagözlü et al., 2024; Shukla & Basu, 2024; Das et al., 2025; Xue et al., 2026).

Most existing studies analyze a single preference relation under a fixed feedback model. In contrast, we compare Pareto, linear, and lexicographic identification within a common characteristic-time framework and determine when multi-objective feedback reduces the required exploration.

3 PRELIMINARIES

Notation. For any integer $K \geq 1$, let $[K] = \{1, 2, \ldots, K\}$ and $\Delta_K = \{w \in \mathbb{R}_+^K : \sum_{i=1}^K w_i = 1\}$ denote the probability simplex. For any matrix $\boldsymbol{x} \in \mathbb{R}^{K \times M}$ with rows $x_1, \ldots, x_K$, its squared Frobenius norm is $\|\boldsymbol{x}\|_F^2 = \sum_{i=1}^K \|x_i\|_2^2$. For any set $S \subseteq \mathbb{R}^M$, the distance from $x \in \mathbb{R}^M$ to S is $\mathrm{dist}(x, S) = \inf_{z \in S} \|x - z\|_2$.

We consider a stochastic multi-armed bandit with $K \geq 2$ arms and $M \geq 1$ objectives, where larger rewards are preferred. Each arm $i \in [K]$ is associated with an unknown mean vector $\mu_i \in \mathbb{R}^M$, and the bandit instance is denoted by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^{K \times M}$. At round t, the learner selects an arm I_t based on the observations collected so far and observes

$$X_t = \mu_{I_t} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I_M),$$

where the noise vectors are independent across rounds. Accordingly, sample complexity counts vector observations (i.e., arm pulls) rather than individual scalar measurements.

The identification target is determined by a known preference relation $\preceq$ on $\mathbb{R}^M$, with strict part denoted by $\prec$. We consider three representative orders:

$$\begin{aligned}
x \prec_{\mathrm{P}} y &\iff x_m \leq y_m \text{ for every } m \text{ and } x_m < y_m \text{ for some } m, && \text{(Pareto)},\\
x \prec_v y &\iff v^\top x < v^\top y, \quad v \in \mathbb{R}_+^M \setminus \{0\}, && \text{(linear scalarization)},\\
x \prec_{\mathrm{lex}} y &\iff x_{r(x,y)} < y_{r(x,y)}, \quad r(x,y) = \min\{m : x_m \neq y_m\}, && \text{(lexicographic)}.
\end{aligned}$$

For an instance $\boldsymbol{\mu}$, the prescribed answer is the complete set of maximal arms:

$$A_\preceq(\boldsymbol{\mu}) = \{i \in [K] : \nexists j \in [K] \text{ such that } \mu_i \prec \mu_j\}. \tag{1}$$

Under linear scalarization or lexicographic preferences, this set is typically a singleton, whereas Pareto dominance may admit multiple maximal arms. Importantly, $A_\preceq(\boldsymbol{\mu})$ is a single set-valued answer: the learner must recover the entire set in Eq. (1), rather than return any one of several valid answers as in multiple-answer identification (Degenne & Koolen, 2019). More generally, our analysis applies to any prescribed answer map with a finite range.

Fixed-confidence identification. For $\delta \in (0, 1/2)$, a fixed-confidence algorithm adaptively selects arms based on past observations, stops at a data-dependent time τ_δ, and outputs a recommendation $\widehat{A}_\delta \subseteq [K]$. It is δ-correct on a model class $\Theta \subseteq \mathbb{R}^{K \times M}$ if, for every $\boldsymbol{\mu} \in \Theta$,

$$\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty) = 1, \qquad \mathbb{P}_{\boldsymbol{\mu}}\Big(\widehat{A}_\delta = A_\preceq(\boldsymbol{\mu})\Big) \geq 1 - \delta.$$

Performance is measured by the expected stopping time $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$, which we aim to minimize subject to δ-correctness. The difficulty of this problem depends on how close $\boldsymbol{\mu}$ is to instances in Θ with a different answer: the closer these alternatives are, the more samples are needed to rule them out.

In classical BAI, a standard assumption is that the best arm is unique (Kaufmann et al., 2016; Garivier & Kaufmann, 2016). With finitely many arms, this implies a strictly positive gap between the best and second-best arms, so the best arm is preserved under sufficiently small perturbations of the means. We impose the analogous condition on the general, possibly set-valued, answer map $A_\preceq$.

Assumption 1 (Local stability)**.** *The true instance $\boldsymbol{\mu} \in \Theta$ is locally stable with respect to the answer map: there exists $r_{\boldsymbol{\mu}} > 0$ such that*

$$A_\preceq(\boldsymbol{\nu}) = A_\preceq(\boldsymbol{\mu}) \quad \textit{for every } \boldsymbol{\nu} \in \Theta \textit{ satisfying } \|\boldsymbol{\nu} - \boldsymbol{\mu}\|_F < r_{\boldsymbol{\mu}}. \tag{2}$$

Assumption 1 requires the answer $A_\preceq(\boldsymbol{\mu})$ to remain unchanged under sufficiently small feasible perturbations. Equivalently, $\boldsymbol{\mu}$ lies in the interior of the set of instances sharing its answer.

4 A Unified Lower Bound for Identification

In this section, we derive a characteristic-time lower bound that holds uniformly across preference relations. We first describe the answer-changing alternatives that any algorithm must rule out. For an instance $\boldsymbol{\mu} \in \Theta$, define its set of feasible alternatives as

$$\mathrm{Alt}_{\preceq}(\boldsymbol{\mu}) = \{\boldsymbol{\nu} \in \Theta : A_{\preceq}(\boldsymbol{\nu}) \neq A_{\preceq}(\boldsymbol{\mu})\}. \tag{3}$$

A δ-correct algorithm must distinguish $\boldsymbol{\mu}$ from every instance in $\mathrm{Alt}_{\preceq}(\boldsymbol{\mu})$. Under the Gaussian observation model, each sample from arm i contributes a KL divergence of $\|\mu_i - \nu_i\|_2^2/2$ between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. Hence, under an allocation $w \in \Delta_K$, the information rate against $\boldsymbol{\nu}$ is

$$\mathcal{I}(\boldsymbol{\mu}, \boldsymbol{\nu}; w) = \frac{1}{2}\sum_{i=1}^{K} w_i \|\mu_i - \nu_i\|_2^2.$$

Optimizing the allocation against the worst-case feasible alternative yields the characteristic information rate, whose reciprocal is the characteristic time:

$$\mathcal{I}^*_{\preceq}(\boldsymbol{\mu}) = \sup_{w \in \Delta_K} \inf_{\boldsymbol{\nu} \in \mathrm{Alt}_{\preceq}(\boldsymbol{\mu})} \mathcal{I}(\boldsymbol{\mu}, \boldsymbol{\nu}; w), \qquad T_{\preceq}(\boldsymbol{\mu}) = \frac{1}{\mathcal{I}^*_{\preceq}(\boldsymbol{\mu})}. \tag{4}$$

We restrict attention to instances with nonempty alternative sets. Together with local stability (Assumption 1), this guarantees $0 < \mathcal{I}^*_{\preceq}(\boldsymbol{\mu}) < \infty$.

Theorem 1 (Unified characteristic-time lower bound)**.** *For any $\delta \in (0, 1/2)$, any algorithm that is δ-correct on Θ, and any $\boldsymbol{\mu} \in \Theta$,*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T_{\preceq}(\boldsymbol{\mu})\,\mathrm{kl}(1-\delta, \delta) \quad \Rightarrow \quad \liminf_{\delta \downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geq T_{\preceq}(\boldsymbol{\mu}), \tag{5}$$

where $\mathrm{kl}(p, q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ is the binary relative entropy.

Theorem 1 applies uniformly to all preference relations because the order enters only through the answer-changing set $\mathrm{Alt}_{\preceq}(\boldsymbol{\mu})$ and the resulting characteristic time $T_{\preceq}(\boldsymbol{\mu})$. By a standard change-of-measure argument, any δ-correct algorithm must accumulate an expected divergence of at least $\mathrm{kl}(1-\delta, \delta)$ between $\boldsymbol{\mu}$ and every answer-changing alternative (Kaufmann et al., 2016; Garivier & Kaufmann, 2016). Under any sampling allocation, the worst-case information gained per sample is at most $\mathcal{I}^*_{\preceq}(\boldsymbol{\mu})$. Therefore, at least $T_{\preceq}(\boldsymbol{\mu})\,\mathrm{kl}(1-\delta, \delta)$ samples are needed in expectation, which yields Eq. (5). The asymptotic statement follows from $\mathrm{kl}(1-\delta, \delta) \sim \log(1/\delta)$ as $\delta \downarrow 0$.

Two-arm geometry. For $K = 2$, suppose that both Θ and the answer depend only on the difference between the two arm means. Define the set of feasible differences as

$$\mathcal{D} = \{\mu_1 - \mu_2 : \boldsymbol{\mu} = (\mu_1, \mu_2) \in \Theta\} \subseteq \mathbb{R}^M.$$

For a fixed instance $\boldsymbol{\mu} = (\mu_1, \mu_2) \in \Theta$, let $d = \mu_1 - \mu_2 \in \mathcal{D}$. With a slight abuse of notation, define

$$T_{\preceq}(d) = T_{\preceq}(\boldsymbol{\mu}), \qquad \mathrm{Alt}_{\preceq}(d) = \{\nu_1 - \nu_2 : \boldsymbol{\nu} = (\nu_1, \nu_2) \in \mathrm{Alt}_{\preceq}(\boldsymbol{\mu})\} \subseteq \mathcal{D}.$$

Both quantities are well defined and invariant under a common translation of the two arm means.

Lemma 1 (Two-arm distance formula)**.** *If $d \in \mathcal{D}$ and $0 < \mathrm{dist}(d, \mathrm{Alt}_{\preceq}(d)) < \infty$, then*

$$T_{\preceq}(d) = \frac{8}{\mathrm{dist}^2(d, \mathrm{Alt}_{\preceq}(d))}. \tag{6}$$

Lemma 1 gives a geometric interpretation: identification complexity is determined by the distance from d to the nearest answer-changing difference. When several objectives must be perturbed to change the answer, their contributions add up in the squared distance, reducing the characteristic time. We next use this formula to compare preference relations.

5 Multi-Objective Identification Gains

In this section, we use the characteristic time to determine when multiple objectives facilitate identification. Specifically, we address three questions: (i) Can a larger optimal set be easier to identify? (ii) How much can vector feedback reduce the complexity of a fixed identification task? (iii) When can a multi-objective task be easier than the best corresponding single-objective benchmark?

5.1 Larger Optimal Sets Can Be Easier to Identify

We first compare Pareto and lexicographic identification on the same two-arm instances. The following result shows that identification difficulty is not determined by the size of the optimal set.

Proposition 1 (A larger optimal set can be easier to identify)**.** *Consider the unrestricted two-arm model (Appendix D). Let $M \geq 4$ be even and $\Delta > 0$. For $d = (\Delta \mathbf{1}_{M/2}, -\Delta \mathbf{1}_{M/2})$, the Pareto-optimal set is $\{1, 2\}$ and the lexicographic optimum is $\{1\}$. Nevertheless,*

$$T_{\mathrm{P}}(d) = \frac{16}{M\Delta^2}, \qquad T_{\mathrm{lex}}(d) = \frac{8}{\Delta^2} \quad \Rightarrow \quad \frac{T_{\mathrm{lex}}(d)}{T_{\mathrm{P}}(d)} = \frac{M}{2}. \tag{7}$$

Conversely, for $d = (1, \ldots, 1, \epsilon)$ with $0 < \epsilon < 1$, both preferences identify arm 1 as optimal, but

$$T_{\mathrm{P}}(d) = \frac{8}{\epsilon^2}, \qquad T_{\mathrm{lex}}(d) = 8 \quad \Rightarrow \quad \frac{T_{\mathrm{lex}}(d)}{T_{\mathrm{P}}(d)} = \epsilon^2. \tag{8}$$

In the first instance, Eq. (7), each arm outperforms the other on half of the objectives. Under Pareto dominance, changing the answer requires removing one arm from the Pareto-optimal set, which in turn requires closing all $M/2$ coordinate gaps on which that arm is superior. Hence, the nearest Pareto alternative lies at squared distance $M\Delta^2/2$. By contrast, the lexicographic answer is determined solely by the first coordinate, so the nearest alternative lies at squared distance Δ^2. Lemma 1 therefore yields the factor-$M/2$ separation in Eq. (7).

The second instance, Eq. (8), shows that this ordering does not hold uniformly. Although both preferences identify arm 1 as optimal, the Pareto answer changes as soon as the smallest coordinate gap ϵ is reversed, whereas the lexicographic answer depends only on the fixed gap in the first coordinate. Consequently, $T_{\mathrm{lex}}(d)/T_{\mathrm{P}}(d) = \epsilon^2$, which can be made arbitrarily small as $\epsilon \downarrow 0$. Together, these two constructions show that neither preference relation is uniformly easier than the other, and that identification complexity cannot be inferred from the size of the optimal set alone.

5.2 The Value of Feedback Depends on the Preference Order

We next compare vector and coordinate feedback for a fixed identification task. The following result shows that the gain from vector feedback can range from none to a factor of M.

Under vector feedback, sampling an arm reveals all M objectives, whereas coordinate feedback reveals only one chosen objective of one chosen arm. For an allocation $q \in \Delta_{KM}$ over arm–objective pairs, define the coordinate-feedback information rate against an alternative $\boldsymbol{\nu}$ as

$$\mathcal{I}^{\mathrm{co}}(\boldsymbol{\mu}, \boldsymbol{\nu}; q) = \frac{1}{2} \sum_{i=1}^{K} \sum_{m=1}^{M} q_{i,m} (\mu_{i,m} - \nu_{i,m})^2.$$

The coordinate-feedback characteristic time $T^{\mathrm{co}}_{\preceq}$ is defined analogously to (4), with q and $\mathcal{I}^{\mathrm{co}}$ replacing w and $\mathcal{I}$, respectively.

Proposition 2 (The value of feedback depends on the preference order)**.** *Every instance $\boldsymbol{\mu}$ with $\mathrm{Alt}_{\preceq}(\boldsymbol{\mu}) \neq \emptyset$ satisfies*

$$T_{\preceq}(\boldsymbol{\mu}) \leq T^{\mathrm{co}}_{\preceq}(\boldsymbol{\mu}) \leq M T_{\preceq}(\boldsymbol{\mu}). \tag{9}$$

In the unrestricted two-arm model with $d = \Delta \mathbf{1}_M$ and $\Delta > 0$, the lexicographic and Pareto orders attain the lower and upper endpoints, respectively:

$$T_{\mathrm{lex}}(d) = T_{\mathrm{P}}(d) = \frac{8}{\Delta^2}, \qquad T^{\mathrm{co}}_{\mathrm{lex}}(d) = \frac{8}{\Delta^2}, \qquad T^{\mathrm{co}}_{\mathrm{P}}(d) = \frac{8M}{\Delta^2}. \tag{10}$$

Proposition 2 shows that the benefit of vector feedback depends fundamentally on the preference order, ranging from no gain to an M-fold reduction in characteristic time. Under Pareto dominance, every objective can change the optimal set, so coordinate feedback must spread samples across all M objectives, whereas vector feedback observes them simultaneously and achieves the full factor-M improvement. Under the lexicographic order, only the first objective matters, so coordinate feedback can concentrate entirely on that objective and match vector feedback. Hence, richer feedback helps only when the preference order makes multiple objectives informative for identification.

5.3 When Multiple Objectives Outperform Single-Objective Benchmarks

We now compare multi-objective identification with the best single-objective benchmark. Each objective, considered alone, defines a scalar BAI problem, and we ask whether multi-objective identification can require fewer samples than even the easiest of these scalar problems.

For the two-arm gap vector $d = \mu_1 - \mu_2 \in \mathbb{R}^M$, assume that $d_m \neq 0$ for all m. Objective m alone then has characteristic time $8/d_m^2$. With $\Delta_{\max}(d) = \max_{m\in[M]} |d_m|$, we define the best single-objective benchmark and the corresponding multi-objective gain as

$$T_{\text{best}}(d) = \min_{m\in[M]} \frac{8}{d_m^2} = \frac{8}{\Delta_{\max}^2(d)}, \quad \mathsf{Gain}_{\preceq}(d) = \frac{T_{\text{best}}(d)}{T_{\preceq}(d)}.$$

A gain greater than one means that the multi-objective task has a smaller characteristic time than every single-objective task. Since the multi-objective and single-objective tasks may have different answers, this is a comparison between tasks, not between feedback models for the same task.

Proposition 3 (Acceleration under three preference orders). *Let $d \in \mathbb{R}^M$ have no zero coordinates, and define $S_+(d) = \sum_{m:d_m>0} d_m^2$ and $S_-(d) = \sum_{m:d_m<0} d_m^2$. Under vector feedback, strict improvement over the best single-objective benchmark occurs in the following regimes:*

$$\begin{aligned}
\mathsf{Gain}_{\mathrm{P}}(d) &= \frac{\min\{S_+(d), S_-(d)\}}{\Delta_{\max}^2(d)} > 1, && \text{if } \min\{S_+(d), S_-(d)\} > \Delta_{\max}^2(d), \\
\mathsf{Gain}_{v}(d) &= \frac{(v^\top d)^2}{\|v\|_2^2 \Delta_{\max}^2(d)} > 1, && \text{if } |v^\top d| > \|v\|_2 \Delta_{\max}(d), \\
\mathsf{Gain}_{\text{lex}}(d) &= \frac{\|d\|_2^2}{\Delta_{\max}^2(d)} > 1, && \text{if } d \in \mathcal{D} = \mathbb{R}^M_{++} \cup (-\mathbb{R}^M_{++}),\ M \geq 2.
\end{aligned} \tag{11}$$

Proposition 3 shows that multi-objective acceleration arises when changing the answer requires coordinated changes across several objectives, so that their information contributions accumulate under vector feedback. Under Pareto dominance, this occurs when the two arms are incomparable and both sign groups have sufficiently large aggregate gaps, i.e., $\min\{S_+(d), S_-(d)\} > \Delta_{\max}^2(d)$. The gain is largest when the positive and negative gaps are well balanced across objectives, and it can reach $\lfloor M/2 \rfloor$. Under linear scalarization, acceleration occurs when the weighted gaps reinforce one another, i.e., $|v^\top d| > \|v\|_2 \Delta_{\max}(d)$. Thus, aligned gaps promote acceleration, whereas cancellation across objectives reduces or eliminates it; the maximum gain is M.

Under lexicographic preferences, acceleration stems from the common-ranking constraint $\mathcal{D} = \mathbb{R}^M_{++} \cup (-\mathbb{R}^M_{++})$. This constraint forces every answer-changing alternative to reverse the ordering on all objectives simultaneously, so the relevant separation scales with $\|d\|_2^2$. Consequently, strict acceleration occurs for every $M \geq 2$, and the gain reaches M when all coordinate gaps have equal magnitude. Without this constraint, reversing the first objective alone suffices to change the lexicographic answer, and the multi-objective acceleration disappears (Appendix G).

6 Attaining the Characteristic Time

In this section, we show that the characteristic-time lower bound in Theorem 1 can be attained asymptotically. Building on the Track-and-Stop framework (Garivier & Kaufmann, 2016; Degenne & Koolen, 2019), we propose OATS, an order-aware procedure that repeatedly estimates the current answer, allocates samples to distinguish it from its closest alternatives, and stops once sufficient evidence has been collected. The complete procedure is given in Algorithm 1.

We first introduce some notation. Let $\mathcal{A} = \{A_{\preceq}(\boldsymbol{\mu}) : \boldsymbol{\mu} \in \Theta\}$ denote the set of possible answers, and let $\Theta_A = \{\boldsymbol{\nu} \in \Theta : A_{\preceq}(\boldsymbol{\nu}) = A\}$ denote the region of instances with answer A. For any estimate $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_K) \in \mathbb{R}^{K\times M}$ and allocation $w \in \Delta_K$, define

$$g_A(\widehat{\boldsymbol{\mu}}, w) = \inf_{\boldsymbol{\nu}\in\Theta\setminus\Theta_A} \mathcal{I}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\nu}; w) = \inf_{\boldsymbol{\nu}\in\Theta\setminus\Theta_A} \frac{1}{2} \sum_{i=1}^K w_i \|\widehat{\mu}_i - \nu_i\|_2^2,$$

which is the smallest information rate that w provides against instances with a different answer.

Algorithm 1 OATS: Order-Aware Track-and-Stop

Input: Preference relation $\preceq$, model class Θ, confidence δ

Pull each arm once; set $t \leftarrow K$ and $S \leftarrow \mathbf{1}_K$
while true **do**
  Compute the empirical mean matrix $\widehat{\boldsymbol{\mu}}_t$
  Fit the candidate answer $\widehat{A}_t$ using Eq. (12)
  Evaluate the evidence Z_t against competing answers using Eq. (13)
  if $Z_t > \beta(t, \delta)$ **then**
    return $\widehat{A}_t$
  end if
  Compute the most informative allocation: $w_t \in \arg\max_{w \in \Delta_K} g_{\widehat{A}_t}(\widehat{\boldsymbol{\mu}}_t, w)$
  Mix with forced exploration: $q_t \leftarrow (1 - t^{-1/2})w_t + t^{-1/2}\mathbf{1}_K/K$
  Update the cumulative target: $S \leftarrow S + q_t$
  Select the arm with the largest deficit: $I_{t+1} \leftarrow \arg\max_{i \in [K]}\{S_i - N_i(t)\}$
  Pull I_{t+1} and update the counts and empirical means
  $t \leftarrow t + 1$
end while

Given a confidence level $\delta \in (0, 1)$, OATS first samples every arm once. At each round t, it computes the empirical mean matrix $\widehat{\boldsymbol{\mu}}_t$ and selects the answer whose region is closest to this estimate:

$$\widehat{A}_t \in \arg\min_{A \in \mathcal{A}} \inf_{\boldsymbol{\nu} \in \Theta_A} \|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\nu}\|_F^2. \tag{12}$$

Let $N_i(t) = \sum_{s=1}^{t} \mathbb{I}\{I_s = i\}$ denote the number of pulls of arm $i \in [K]$ up to round t, and let $N(t) = (N_1(t), \ldots, N_K(t))$. OATS then evaluates the evidence in favor of $\widehat{A}_t$ through

$$Z_t = t \times g_{\widehat{A}_t}\left(\widehat{\boldsymbol{\mu}}_t, \frac{N(t)}{t}\right) = \inf_{\boldsymbol{\nu} \in \Theta \setminus \Theta_{\widehat{A}_t}} \frac{1}{2} \sum_{i=1}^{K} N_i(t) \|\widehat{\mu}_i(t) - \nu_i\|_2^2. \tag{13}$$

Thus, Z_t measures the information accumulated against the closest instance with a different answer. We use the stopping threshold

$$\beta(t, \delta) = \frac{KM}{2} + \sqrt{KM \log \frac{2(t+1)^{K+2}}{\delta}} + \log \frac{2(t+1)^{K+2}}{\delta}, \tag{14}$$

which controls the estimation error uniformly over time and over all possible sample-count vectors.

If $Z_t > \beta(t, \delta)$, the algorithm stops and returns $\widehat{A}_t$. Otherwise, OATS computes the most informative allocation

$$w_t \in \arg\max_{w \in \Delta_K} g_{\widehat{A}_t}(\widehat{\boldsymbol{\mu}}_t, w), \tag{15}$$

which maximizes the estimated information rate against alternatives to the current answer. To ensure sufficient exploration, it then sets

$$q_t = (1 - t^{-1/2})w_t + t^{-1/2}\frac{\mathbf{1}_K}{K}.$$

The uniform component vanishes asymptotically but guarantees that every arm receives $\Omega(\sqrt{t})$ observations. Next, OATS updates the cumulative target and selects the arm with the largest deficit:

$$S \leftarrow S + q_t, \qquad I_{t+1} \in \arg\max_{i \in [K]}\{S_i - N_i(t)\}.$$

Since $S_i - N_i(t)$ is the sampling deficit of arm i relative to its cumulative target, this tracking rule keeps the realized sample counts within a bounded distance of their targets. Finally, OATS pulls arm I_{t+1} and updates the counts and empirical means.

The following theorem establishes the correctness and asymptotic optimality of OATS.

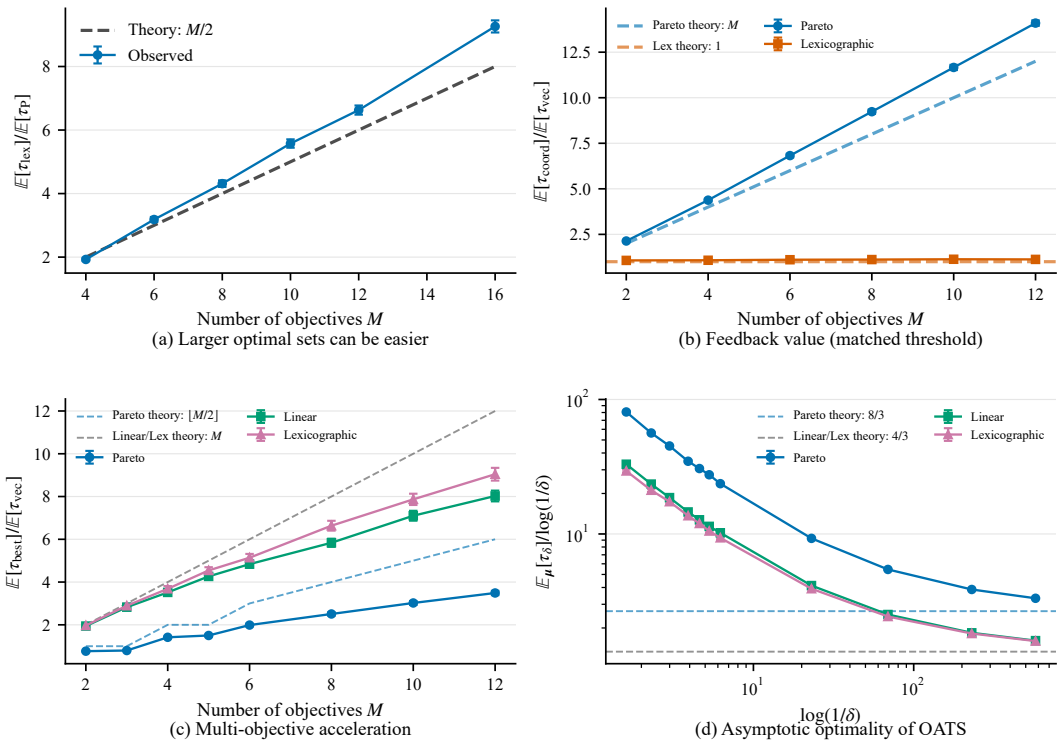


Figure 1: Synthetic results for Propositions 1–3 and Theorem 2. (a) The ratio $\mathbb{E}[\tau_{\mathrm{lex}}]/\mathbb{E}[\tau_{\mathrm{P}}]$ grows with M: Pareto-set identification becomes increasingly easier than lexicographic identification. (b) The feedback gain $\mathbb{E}[\tau_{\mathrm{coord}}]/\mathbb{E}[\tau_{\mathrm{vec}}]$ grows with M under the Pareto order but remains near one under the lexicographic order. (c) The gain $\mathbb{E}[\tau_{\mathrm{best}}]/\mathbb{E}[\tau_{\mathrm{vec}}]$ increases with M under all three orders and stays above one, confirming multi-objective acceleration. (d) The normalized stopping time $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]/\log(1/\delta)$ decreases toward the corresponding characteristic time as δ decreases. Markers show empirical means with 95% confidence intervals, and dashed lines show theoretical predictions.

Theorem 2 (Asymptotic optimality of OATS)**.** *Suppose every instance in Θ satisfies Assumption 1 and has a nonempty alternative set. Then* OATS *is δ-correct, and for every $\boldsymbol{\mu} \in \Theta$,*

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T^*_{\preceq}(\boldsymbol{\mu}). \tag{16}$$

Theorem 2 shows that OATS is not only δ-correct but also asymptotically instance-optimal: for any fixed instance, its expected stopping time matches the characteristic-time lower bound of Theorem 1 to first order as $\delta \downarrow 0$. This extends classical Track-and-Stop for scalar BAI (Garivier & Kaufmann, 2016) to preference-induced, possibly set-valued identification problems. Whereas previous extensions have considered multiple valid answers (Degenne & Koolen, 2019) or specific multi-objective targets such as Pareto-front identification (Crépon et al., 2024), OATS handles different preference orders and model structures through a single answer-changing formulation. Consequently, the three phenomena of Section 5 are attained by an adaptive procedure and are not mere properties of the lower bound. In particular, whenever the preference relation or structural constraints enlarge the information distance to the nearest answer-changing alternative, OATS asymptotically exploits this additional information. Applying the same analysis to a single known coordinate also shows that the oracle single-objective benchmark is attainable.

7 Experiments

In this section, we run synthetic experiments to test whether OATS exhibits the identification phenomena of Propositions 1–3 and the asymptotic optimality of Theorem 2. Appendix A reports supplementary results on instances derived from the HELM Instruct ratings (Zhang et al., 2024).

Experimental setup. All experiments use independent observations $X_{i,t} \sim \mathcal{N}(\mu_i, I_M)$. For two-arm instances, we set $\mu_1 = d/2$ and $\mu_2 = -d/2$. Unless otherwise specified, we use $\Delta = 1$ and $\delta = 0.05$. Each configuration is evaluated over 500 independent repetitions with deterministic seeds.

Figure 1(a) evaluates Proposition 1 for even $M \in \{4, 6, 8, 10, 12, 16\}$ under the mixed-sign gap $d = (\mathbf{1}_{M/2}, -\mathbf{1}_{M/2})$, for which the proposition predicts $T_{\mathrm{lex}}/T_{\mathrm{P}} = M/2$. The empirical stopping-time ratio grows approximately linearly in M and closely tracks this prediction, lying slightly above it for larger M. These results confirm that the advantage of Pareto-set identification over lexicographic identification scales linearly with the number of objectives.

Figure 1(b) validates Proposition 2 with $d = \mathbf{1}_M$ and $M \in \{2, 4, 6, 8, 10, 12\}$, comparing vector and coordinate feedback under the Pareto and lexicographic orders. The proposition predicts coordinate-to-vector characteristic-time ratios of M and 1, respectively. The observed Pareto ratio grows approximately linearly in M, whereas the lexicographic ratio stays close to one. Both trends closely follow the predictions, up to a modest finite-confidence gap. These results confirm that the benefit of vector feedback depends critically on the preference order.

Figure 1(c) tests Proposition 3 for $M \in \{2, 3, 4, 5, 6, 8, 10, 12\}$ by comparing vector-feedback identification with a scalar oracle that knows $m^* \in \arg\max_m |d_m|$. We use $d = (\mathbf{1}_{M/2}, -\mathbf{1}_{M/2})$ for the Pareto order and $d = \mathbf{1}_M$ for the linear and common-ranking lexicographic orders, with $v = \mathbf{1}_M/\sqrt{M}$ for linear scalarization. The stopping-time gain $\mathbb{E}[\tau_{\mathrm{best}}]/\mathbb{E}[\tau_{\mathrm{vec}}]$ grows nearly linearly in M for the linear and lexicographic orders and more slowly for the Pareto order, consistent with the theory. At finite confidence, the empirical gains remain below the characteristic-time predictions, particularly for larger M, because the thresholds in Eq. (14) for the scalar and vector procedures use different dimensions, namely $KM = 2$ and $KM = 2M$, respectively.

Figure 1(d) evaluates the asymptotic optimality of OATS established in Theorem 2. We use $M = 6$, with $d = (1, 1, 1, -1, -1, -1)$ for the Pareto order and $d = \mathbf{1}_6$ for the linear and lexicographic orders. We plot the normalized stopping time $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]/\log(1/\delta)$ for confidence levels δ ranging from 0.2 down to 10^{-250}. The theorem predicts that this quantity converges to $T_{\preceq}(\boldsymbol{\mu}) = 8/3$ for Pareto identification and to $T_{\preceq}(\boldsymbol{\mu}) = 4/3$ for both linear and lexicographic identification. Consistent with this prediction, all three curves decrease monotonically toward their respective limits as $\log(1/\delta)$ grows. The linear and lexicographic curves nearly coincide and approach their common limit of $4/3$, whereas the Pareto curve converges to the higher limit of $8/3$. The empirical results thus agree with the asymptotic characterization in Theorem 2.

Summary. Overall, the synthetic experiments confirm that larger optimal sets can be easier to identify, the benefit of vector feedback depends on the preference order, and multiple objectives can accelerate identification; moreover, the normalized stopping time of OATS approaches the characteristic time. Appendix A complements these synthetic results with implementation details, HELM-derived Gaussian instances, allocation baselines against uniform sampling and Oracle-Track, and additional robustness and scaling checks.

8 Conclusion and Future Work

In this paper, we studied when multiple objectives make fixed-confidence identification easier rather than harder. By casting different preference orders in a common characteristic-time framework, we disentangled the effects of the preference relation, the feasible model class, and the feedback model on identification complexity. Our results show that identification difficulty is governed by the geometry of answer-changing alternatives rather than by the number of objectives or the size of the optimal set alone. In particular, Pareto and lexicographic identification admit no universal hardness ordering, the benefit of vector feedback ranges from no gain to an M-fold improvement, and multi-objective identification can be strictly easier than the best corresponding single-objective problem. We further proposed OATS, which asymptotically matches the lower bound and thereby shows that these gains are achievable.

Future directions include extending the analysis beyond Gaussian observations to more general reward and feedback models, such as heavy-tailed noise or partially observed objectives, and developing scalable approximations of the optimal allocation for problems with many arms or objectives.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, LaTeX consistency checks, and drafting 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, algorithm, 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, reward models, confidence levels, evaluation metrics, and baselines. The experimental appendix describes the HELM-derived evaluation, including the released evaluator ratings, the prompt-alignment rule, the global scaling factor, and the Gaussian simulation model. The code and complete experimental materials, including instance generators, implementations, plotting scripts, raw trial outputs, processed summaries, and validation checks, will be made publicly available upon acceptance.

REFERENCES

Çağın 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, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd 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.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pp. 23–37, 2009.

Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. Multi-objective bandits: Optimizing the generalized Gini index. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 625–634, 2017.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 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.

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.

Elise Crépon, 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, Wouter M. Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems 32*, pp. 14465–14474, 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 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.

Yihan Du, Yuko Kuroki, and Wei Chen. Combinatorial pure exploration with full-bandit or partial linear feedback. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 7262–7270, 2021.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006.

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. 10666–10676, 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 Conference on Learning Theory*, pp. 998–1027, 2016.

Alihan Hüyük and Cem Tekin. Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives. *Machine Learning*, 110:1233–1266, 2021.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Conference on Learning Theory*, pp. 423–439, 2014.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 655–662, 2012.

Efe Mert Karagözlü, Yaşar Cahit Yıldırım, Çağın Ararat, and Cem Tekin. Learning the Pareto set under incomplete preferences: Pure exploration in vector bandits. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 3070–3078, 2024.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1238–1246, 2013.

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

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Adaptive algorithms for relaxed pareto set identification. In *Advances in Neural Information Processing Systems 36*, pp. 35190–35201, 2023.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit Pareto set identification: The fixed-budget setting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 2548–2556, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, and et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2 edition, 2006.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48(1):67–113, 2013.

Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6): 1625–1647, 2020.

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

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems 27*, pp. 828–836, 2014.

Cem Tekin and Eralp Turğay. Multi-objective contextual multi-armed bandit with a dominant objective. *IEEE Transactions on Signal Processing*, 66(14):3799–3813, 2018.

Eralp Turğay, Doruk Öner, 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.

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

Bo Xue, Yuanyu Wan, Zhichao Lu, and Qingfu Zhang. Beyond the lower bound: Bridging regret minimization and best arm identification in lexicographic bandits. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, pp. 27414–27422, 2026.

Yian Zhang, Yifan Mai, Josselin Somerville Roberts, Rishi Bommasani, Yann Dubois, and Percy Liang. HELM Instruct: A multidimensional instruction following evaluation framework with absolute ratings. Stanford Center for Research on Foundation Models, 2024. URL https://crfm.stanford.edu/2024/02/18/helm-instruct.html.

A Experimental Details and Additional Results

Implementation and numerical checks. We implement OATS following Algorithm 1, including the empirical answer, the stopping statistic, the forced-exploration mixture, and the cumulative-deficit tracking rule. For two-arm experiments under vector feedback, Lemma 1 and its proof reduce the closest-alternative problem to (18), which yields the balanced allocation $w = (1/2, 1/2)$; the corresponding Pareto, linear, and lexicographic distances are derived in Proposition 4 and its proof. For coordinate feedback, we use the information formulas and optimal allocations derived in the proof of Proposition 2. For linear identification with general K, we project each mean vector onto $v/\|v\|_2$, which reduces the allocation problem to scalar Gaussian best-arm identification. For a candidate singleton answer $A = \{i_A\}$, let $c_j = (v^\top(\mu_{i_A} - \mu_j))^2/(2\|v\|_2^2)$ for $j \neq i_A$, so that

$$g_A(\boldsymbol{\mu}, w) = \min_{j \neq i_A} c_j \frac{w_{i_A} w_j}{w_{i_A} + w_j}.$$

The maximizing weights satisfy $w_j = w_{i_A}\lambda/(c_j - \lambda)$ for $j \neq i_A$, where $\lambda \in (0, \min_{j \neq i_A} c_j)$ is the unique solution of $\sum_{j \neq i_A}(\lambda/(c_j - \lambda))^2 = 1$, and w_{i_A} is then determined by the normalization $\sum_k w_k = 1$. Since the left-hand side is strictly increasing in λ, we solve this equation by bisection on $(0, \min_{j \neq i_A} c_j)$ (Boyd & Vandenberghe, 2004, Section 4.2.5); after 45 iterations, the bracketing interval has width at most $2^{-45} \min_{j \neq i_A} c_j$. We cross-check the resulting allocation and information value against a constrained epigraph formulation solved by sequential quadratic programming (Nocedal & Wright, 2006, Chapter 18), and the two solutions agree to seven decimal places on all test instances. We further verify numerically the two-arm distance formulas, the scalar-acceleration gains, the vector- and coordinate-feedback endpoints, and the predicted scaling laws, and we confirm that instances outside the common-ranking model are correctly rejected. Every simulated trial stopped before reaching the prescribed sample cap, so all reported stopping times are finite.

Seeds, intervals, and baselines. For reproducibility, each experimental condition is assigned a fixed base seed, from which independent per-trial seeds are derived. We report 95% confidence intervals for mean stopping times as the sample mean ± 1.96 standard errors, intervals for ratios of mean stopping times via the delta method for independent samples, and intervals for error rates via the exact two-sided Clopper–Pearson method. The three allocation rules (OATS, Uniform, and Oracle-Track) differ only in their target allocations; they share the same observations, candidate answer, evidence statistic, and stopping threshold. Oracle-Track computes its target allocation from the true mean matrix; it is therefore not implementable in practice and serves only as a reference. In the two-arm vector-feedback setting, all three rules target the same allocation $(1/2, 1/2)$, so any observed differences among them are due solely to Monte Carlo variability.

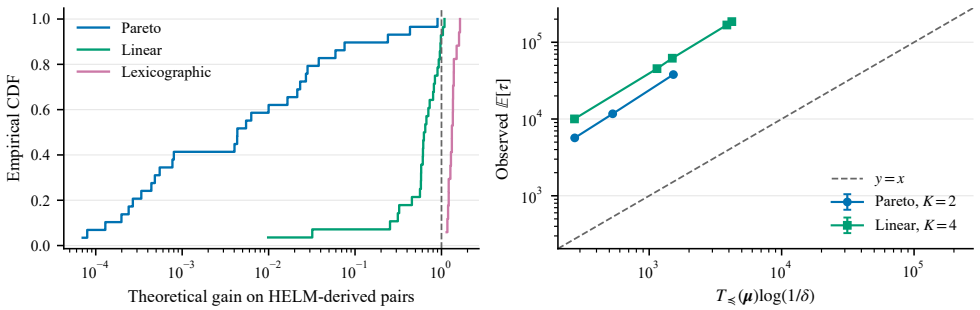
Stopping threshold and coordinate-feedback ablation. All simulations stop as soon as Z_t exceeds the threshold

$$\beta(t, \delta) = \frac{D}{2} + \sqrt{D \log\big(2(t+1)^{K_{\exp}+2}/\delta\big)} + \log\big(2(t+1)^{K_{\exp}+2}/\delta\big),$$

where $K_{\exp}$ is the number of selectable experiments and D is the total dimension of their mean parameters. The values dictated by our proofs are $(K_{\exp}, D) = (K, KM)$ for vector feedback and (KM, KM) for coordinate feedback, and we use them throughout, with one exception. In Figure 1(b), we also set $K_{\exp} = K$ for coordinate feedback, so that both feedback models use the same threshold and the comparison isolates the effect of the feedback model itself.

A.1 HELM-Derived Model Evaluation

Data and Gaussian construction. We use the publicly released GPT-4 evaluator ratings from CRFM HELM Instruct v1.0.0 (Zhang et al., 2024). In the HELM-derived Gaussian instances, each arm corresponds to a language model and each reward vector contains $M = 5$ objectives. Pairwise comparisons use $K = 2$ arms, and the linear experiments use all $K = 4$ models. HELM groups prompts into evaluation scenarios, each corresponding to a fixed prompt source or task family. For each scenario s, model i, and objective m, we let $\widehat{\mu}_{i,m}^{(s)}$ denote the average released rating on objective m over the prompts shared by all four models in that scenario, and we collect these five averages into the mean vector $\widehat{\mu}_i^{(s)} \in \mathbb{R}^5$. We also construct a pooled instance by averaging each model's



(a) Gain distributions (b) Stopping-time scaling

Figure 2: HELM-derived Gaussian instances. (a) Empirical CDF of the theoretical gain across within-scenario model pairs; a gain above 1 indicates a multi-objective speedup over the best scalar benchmark. (b) Observed stopping time versus $T_{\preceq}(\boldsymbol{\mu})\log(1/\delta)$ for eligible Pareto and linear instances, showing that $T_{\preceq}(\boldsymbol{\mu})$ captures instance difficulty on HELM-derived geometries.

ratings over all 678 prompts retained across the seven scenarios. To place these means on the scale of the unit-covariance Gaussian model assumed in our theory, we rescale them as

$$\widetilde{\mu}_i^{(s)} = \gamma\, \widehat{\mu}_i^{(s)}, \qquad \gamma = 1.3226021498,$$

where γ is the inverse of the median nonzero within-scenario model-pair Euclidean gap on the raw mean scores. We then simulate fresh observations $X_{i,t} \sim \mathcal{N}(\widetilde{\mu}_i^{(s)}, I_5)$. Because $\gamma > 0$ is a single global constant, this rescaling preserves all Pareto, linear, and lexicographic comparisons.

Across the seven scenarios and the pooled instance, the six model pairs yield 48 two-arm instances with $M = 5$. Figure 2(a) shows the empirical distribution of the theoretical gain over the pairs to which each notion applies, namely Pareto, linear, and common-ranking lexicographic pairs. Most Pareto and linear gains fall below one, whereas the lexicographic distribution is shifted to the right. The large accelerations in our synthetic examples are therefore not typical of HELM mean gaps. Real model pairs rarely exhibit the balanced, mixed-sign gap structure that favors Pareto identification, while common-ranking lexicographic pairs can still aggregate evidence across objectives. Figure 2(b) examines whether the characteristic time predicts stopping times on HELM-derived instances. We fix $\delta = 0.05$ and consider the eligible two-arm Pareto instances and the four-arm linear instances. Observed stopping times increase with $T_{\preceq}(\boldsymbol{\mu})\log(1/\delta)$, indicating that $T_{\preceq}(\boldsymbol{\mu})$ remains a reliable indicator of relative instance difficulty. The points lie well above the $y = x$ reference line because, at this practical confidence level, the finite-confidence threshold is conservative and its lower-order terms are not negligible.

Allocation baselines. We compare three allocation rules on a four-arm linear HELM instance. OATS estimates the answer and tracks the allocation returned by Eq. (15). Uniform samples all arms equally with the same stopping rule. Oracle-Track, an unattainable reference, computes the optimal allocation from the true mean matrix with the same observations and stopping rule.

Table 1: A four-arm HELM-derived linear instance (Best ChatGPT Prompts scenario, $\delta = 0.05$). All methods share the same evidence statistic and stopping threshold. We report the mean stopping time, its 95% confidence interval, and the number of incorrect answers over 500 trials. Oracle-Track uses the true mean matrix and is not implementable.

Allocation	Mean τ	95% CI	Errors
OATS	**61,797**	[61,119, 62,474]	0/500
Uniform	87,600	[86,663, 88,538]	0/500
Oracle-Track	61,938	[61,289, 62,588]	0/500

As shown in Table 1, OATS reduces the mean stopping time by about 29.5% relative to uniform allocation. Its performance is statistically indistinguishable from Oracle-Track, as the two 95% confidence intervals overlap. All three methods return the correct answer in every trial.

A.2 Additional Synthetic Experiments

This subsection reports two additional sets of synthetic experiments. The first varies the gap size, the scalarization weights, and the balance of Pareto coordinates. The second extends linear identification to more than two arms. Unless otherwise stated, all settings follow those of the main experiments.

Table 2: Additional two-arm synthetic checks ($\delta = 0.05$, 500 repetitions).

Check	Setting	$T_{\preceq}(\boldsymbol{\mu})$	Mean τ
Gap scaling	$\Delta = 0.5$	10.667	631.2
Gap scaling	$\Delta = 1$	2.667	133.4
Gap scaling	$\Delta = 2$	0.667	28.3
Unequal linear weights	$M = 4$	2.400	94.5
Unequal linear weights	$M = 8$	1.259	58.3
Unequal linear weights	$M = 12$	0.855	45.3
Pareto imbalance	$p = 1$	8.000	447.1
Pareto imbalance	$p = 2$	4.000	208.5
Pareto imbalance	$p = 3$	2.667	131.0
Pareto imbalance	$p = 4$	2.000	106.7

Table 2 confirms that empirical stopping times respond to the instance geometry in the same direction as the characteristic time. First, each doubling of the gap scale Δ divides $T_{\preceq}(\boldsymbol{\mu})$ by four, consistent with its inverse-square dependence on the gap. The mean stopping time decreases by a similar factor of about 4.7, with the difference attributable to lower-order terms in the threshold. Second, with unequal linear weights, increasing the number of objectives enlarges the projected gap and hence reduces both the characteristic time and the stopping time. Third, for Pareto identification with $M = 8$ objectives split into sign groups of sizes p and $8 - p$, making the two groups more balanced increases the distance to the closest alternative. As a result, $T_{\preceq}(\boldsymbol{\mu})$ scales as $1/\min\{p, 8-p\}$ and the stopping time decreases accordingly.

Table 3: Linear identification with larger K ($\delta = 0.05$, 500 repetitions). The last three columns report empirical mean stopping times. Oracle-Track uses the true instance and is not implementable.

K	M	$T_{\preceq}(\boldsymbol{\mu})$	OATS	Uniform	Oracle-Track
5	3	3.568	302	397	312
5	5	3.568	354	470	367
5	8	3.568	420	561	438
10	3	4.765	792	1586	813
10	5	4.765	933	1876	958
10	8	4.765	1118	2250	1127
20	3	5.792	1941	6726	1967
20	5	5.792	2283	7801	2298
20	8	5.792	2712	9287	2736

Table 3 extends linear identification beyond two arms. For $K \in \{5, 10, 20\}$ and $M \in \{3, 5, 8\}$, we generate instances with a unique best arm after projection and a fixed projected gap between the best arm and its closest competitor, which avoids near-ties. This comparison isolates the value of adaptive allocation. Uniform uses the same stopping rule but samples all arms equally, whereas Oracle-Track tracks the true optimal allocation and is not implementable. For fixed K, $T_{\preceq}(\boldsymbol{\mu})$ does not depend on M because the projected gaps are held fixed. The increase in stopping time with M instead comes from the threshold, whose dimension term $D = KM$ grows with M. OATS performs on par with Oracle-Track in every setting and is faster than uniform allocation. Its advantage grows with the number of arms, from roughly $1.3\times$ at $K = 5$ to $2.0\times$ at $K = 10$ and about $3.4\times$ at $K = 20$.

B Proof of Theorem 1

Proof. Fix $\boldsymbol{\mu} \in \Theta$. The claim is trivial if $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = \infty$, so assume that this expectation is finite. For any $\boldsymbol{\nu} \in \mathrm{Alt}_\preceq(\boldsymbol{\mu})$, consider the event

$$E = \{\widehat{A}_\delta = A_\preceq(\boldsymbol{\mu})\}.$$

Since the algorithm is δ-correct on the entire model class and $A_\preceq(\boldsymbol{\nu}) \neq A_\preceq(\boldsymbol{\mu})$,

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

Applying the change-of-measure inequality (Kaufmann et al., 2016, Lemma 1) to the stopped history gives

$$\sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \frac{\|\mu_i - \nu_i\|_2^2}{2} \geq \mathrm{kl}\big(\mathbb{P}_{\boldsymbol{\mu}}(E), \mathbb{P}_{\boldsymbol{\nu}}(E)\big) \geq \mathrm{kl}(1-\delta, \delta). \tag{17}$$

For completeness, the first inequality follows from the data-processing inequality applied to the indicator of E in the stopped history. The relative entropy of that history equals the left-hand side, because the sampling rule is the same under both instances and each observation from arm i contributes the Gaussian divergence $\|\mu_i - \nu_i\|_2^2/2$. This identity is justified by first stopping at $\tau_\delta \wedge n$ and then using the integrability of the stopped Gaussian log-likelihood increments when $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] < \infty$. The second inequality uses the monotonicity of the binary relative entropy $\mathrm{kl}(p, q)$ for $p \geq 1-\delta > \delta \geq q$.

Define the expected sampling proportions as

$$\overline{w}_i = \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}.$$

The denominator is positive by Eq. (17), and $\sum_i N_i(\tau_\delta) = \tau_\delta$ implies $\overline{w} \in \Delta_K$. Taking the infimum of Eq. (17) over $\boldsymbol{\nu} \in \mathrm{Alt}_\preceq(\boldsymbol{\mu})$ yields

$$\begin{aligned} \mathrm{kl}(1-\delta, \delta) &\leq \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \inf_{\boldsymbol{\nu} \in \mathrm{Alt}_\preceq(\boldsymbol{\mu})} \mathcal{I}(\boldsymbol{\mu}, \boldsymbol{\nu}; \overline{w}) \\ &\leq \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]\, \mathcal{I}^*_\preceq(\boldsymbol{\mu}). \end{aligned}$$

Using $T_\preceq(\boldsymbol{\mu}) = 1/\mathcal{I}^*_\preceq(\boldsymbol{\mu})$ proves Eq. (5). Finally,

$$\mathrm{kl}(1-\delta, \delta) = (1-2\delta)\log\frac{1-\delta}{\delta}, \qquad \frac{\mathrm{kl}(1-\delta, \delta)}{\log(1/\delta)} \longrightarrow 1,$$

which proves the asymptotic lower bound. □

C Proof of Lemma 1

For a feasible-difference set $\mathcal{D} \subseteq \mathbb{R}^M$, write

$$\Theta(\mathcal{D}) = \{(\mu_1, \mu_2) \in \mathbb{R}^{2\times M} : \mu_1 - \mu_2 \in \mathcal{D}\}.$$

The assumption of the lemma that the model depends only on the mean difference means precisely that $\Theta = \Theta(\mathcal{D})$; in particular, the common location of the two means is unrestricted.

Proof. Fix an allocation $(\alpha, 1-\alpha) \in \Delta_2$ and an alternative difference $z \in \mathrm{Alt}_\preceq(d)$. Since the model $\Theta(\mathcal{D})$ leaves the common location of the means unrestricted, every pair with difference z is feasible. Writing $a = \mu_1 - \nu_1$ and $b = \mu_2 - \nu_2$, the constraint is $a - b = d - z$. The identity

$$\alpha\|a\|_2^2 + (1-\alpha)\|b\|_2^2 = \alpha(1-\alpha)\|a-b\|_2^2 + \|\alpha a + (1-\alpha)b\|_2^2$$

therefore gives

$$\inf_{\nu_1 - \nu_2 = z} \mathcal{I}(\boldsymbol{\mu}, \boldsymbol{\nu}; (\alpha, 1-\alpha)) = \frac{\alpha(1-\alpha)}{2}\|d - z\|_2^2. \tag{18}$$

The infimum is attained at $a = (1-\alpha)(d-z)$ and $b = -\alpha(d-z)$, including when $\alpha \in \{0, 1\}$. Consequently,

$$\mathcal{I}^*_\preceq(d) = \max_{\alpha \in [0,1]} \frac{\alpha(1-\alpha)}{2}\,\mathrm{dist}^2(d, \mathrm{Alt}_\preceq(d)) = \frac{1}{8}\,\mathrm{dist}^2(d, \mathrm{Alt}_\preceq(d)).$$

Taking reciprocals proves Eq. (6). □

D AUXILIARY RESULTS FOR THE TWO-ARM COMPARISONS

We now specify the feasible-difference sets used in the two-arm comparisons. Define

$$\begin{aligned} \mathcal{D}_{\text{free}} &= (\mathbb{R} \setminus \{0\})^M, \\ \mathcal{D}_v &= \{z \in \mathcal{D}_{\text{free}} : v^\top z \neq 0\}, \\ \mathcal{D}_{\text{cmp}} &= \mathbb{R}^M_{++} \cup (-\mathbb{R}^M_{++}). \end{aligned} \tag{19}$$

The unrestricted Pareto and lexicographic comparisons use $\Theta(\mathcal{D}_{\text{free}})$, and linear scalarization uses $\Theta(\mathcal{D}_v)$. Excluding ties ensures local stability and does not change the distances to the decision boundaries computed below. In the common-ranking comparison in Proposition 3, the feasible-difference set $\mathcal{D}$ is $\mathcal{D}_{\text{cmp}}$ and the model is $\Theta(\mathcal{D}_{\text{cmp}})$. This restriction is known to the learner and applies to the entire model, i.e., to both the true instance and every feasible alternative in Eq. (3).

Proposition 4 (Two-arm characteristic times)**.** *Let $d \in \mathcal{D}_{\text{free}}$, and define*

$$P(d) = \{m : d_m > 0\}, \qquad N(d) = \{m : d_m < 0\},$$

$$S_+(d) = \sum_{m \in P(d)} d_m^2, \qquad S_-(d) = \sum_{m \in N(d)} d_m^2.$$

Under vector feedback, Pareto identification on $\Theta(\mathcal{D}_{\text{free}})$ has characteristic time

$$T_{\text{P}}(d) = \begin{cases} \dfrac{8}{\min_{m \in [M]} d_m^2}, & P(d) = \emptyset \text{ or } N(d) = \emptyset, \\ \dfrac{8}{\min\{S_+(d), S_-(d)\}}, & P(d) \neq \emptyset \text{ and } N(d) \neq \emptyset. \end{cases} \tag{20}$$

For $v \in \mathbb{R}^M_+ \setminus \{0\}$ and $d \in \mathcal{D}_v$, linear scalarization on $\Theta(\mathcal{D}_v)$ satisfies

$$T_v(d) = \frac{8\|v\|_2^2}{(v^\top d)^2}. \tag{21}$$

For lexicographic identification, distinguish the two model classes by writing

$$T^{\text{free}}_{\text{lex}}(d) = \frac{8}{d_1^2} \quad (d \in \mathcal{D}_{\text{free}}), \qquad T^{\text{cmp}}_{\text{lex}}(d) = \frac{8}{\|d\|_2^2} \quad (d \in \mathcal{D}_{\text{cmp}}). \tag{22}$$

The superscripts in Eq. (22) indicate the model class, not the feedback model.

Proof. By Lemma 1, it suffices to compute the squared distance from d to the corresponding set of feasible answer-changing differences. We use $\text{dist}(d, S) = \text{dist}(d, \overline{S})$, so boundary points need not belong to the model to determine this infimum.

Pareto preferences. Suppose first that $d \in \mathbb{R}^M_{++}$. The prescribed answer is $\{1\}$. Any alternative in $\mathcal{D}_{\text{free}}$ must have a negative coordinate, so its squared distance from d is at least $\min_m d_m^2$. Conversely, choose $j \in \arg\min_m d_m^2$ and replace d_j by $-\eta$, leaving the other coordinates unchanged. This yields a feasible answer-changing difference for every $\eta > 0$, and letting $\eta \downarrow 0$ shows that the squared distance equals $\min_m d_m^2$. The case $d \in -\mathbb{R}^M_{++}$ is symmetric.

If $P(d)$ and $N(d)$ are both nonempty, neither arm dominates the other, and the answer is $\{1, 2\}$. An alternative changes this answer if and only if one arm dominates the other. Since no coordinate is zero in $\mathcal{D}_{\text{free}}$, this is equivalent to all coordinate differences having the same sign. Hence

$$\overline{\text{Alt}_{\text{P}}(d)} = \mathbb{R}^M_+ \cup (-\mathbb{R}^M_+).$$

Coordinatewise projection gives

$$\text{dist}^2(d, \mathbb{R}^M_+) = S_-(d), \qquad \text{dist}^2(d, -\mathbb{R}^M_+) = S_+(d).$$

Both projections can be approached by strictly positive or strictly negative differences, respectively. Therefore, the squared distance to the alternative set is $\min\{S_+(d), S_-(d)\}$, which proves Eq. (20).

Linear scalarization. Write $a = v^\top d \neq 0$. The closure of the answer-changing differences is the halfspace

$$H_a = \{z \in \mathbb{R}^M : a\, v^\top z \leq 0\}.$$

Indeed, differences whose scalarized gap has the opposite sign are answer-changing, and excluding coordinate ties does not affect the closure of this open halfspace. For any $z \in H_a$, Cauchy–Schwarz implies

$$\|d - z\|_2 \geq \frac{|v^\top (d - z)|}{\|v\|_2} \geq \frac{|a|}{\|v\|_2}.$$

The projection

$$z^* = d - \frac{a}{\|v\|_2^2} v$$

attains this distance to H_a. To approach z^* by feasible alternatives, take $z^* - \eta\, \mathrm{sign}(a) v$ with $\eta > 0$ and, if necessary, apply an additional perturbation of norm $o(\eta)$ to avoid coordinate ties. Such perturbations preserve the strictly opposite sign of the scalarized gap. Thus,

$$\mathrm{dist}^2(d, \mathrm{Alt}_v(d)) = \frac{(v^\top d)^2}{\|v\|_2^2},$$

which proves Eq. (21).

Unrestricted lexicographic preferences. Suppose $d_1 > 0$. Every answer-changing difference in $\mathcal{D}_{\mathrm{free}}$ has a negative first coordinate, so its squared distance from d is at least d_1^2. Replacing only d_1 by $-\eta$ gives feasible alternatives approaching this lower bound as $\eta \downarrow 0$. Exchanging the arms covers $d_1 < 0$ and proves the first identity in Eq. (22).

Lexicographic preferences under common ranking. Suppose $d \in \mathbb{R}^M_{++}$. The answer is $\{1\}$, and the feasible answer-changing differences are exactly $-\mathbb{R}^M_{++}$. Consequently,

$$\inf_{z \in -\mathbb{R}^M_{++}} \|d - z\|_2^2 = \sum_{m=1}^{M} d_m^2.$$

For the lower bound, each summand $(d_m - z_m)^2$ is at least d_m^2. For the upper bound, set $z = -\eta \mathbf{1}_M$ and let $\eta \downarrow 0$. The case $d \in -\mathbb{R}^M_{++}$ is symmetric. The second identity in Eq. (22) now follows from Lemma 1. □

The role of the feasible model. Both lexicographic identities can be evaluated at the same gap vector with strictly same-sign coordinates, but they refer to different sets of feasible alternatives. Merely observing that the true difference has a common sign does not justify $T_{\mathrm{lex}}^{\mathrm{cmp}}(d)$: the learner must know that all feasible instances obey the common-ranking restriction. This distinction explains why the unrestricted lexicographic examples in Propositions 1 and 2 are consistent with the common-ranking acceleration in Proposition 3.

E Proof of Proposition 1

Proof. Both orders are evaluated on $\Theta(\mathcal{D}_{\mathrm{free}})$. For $d = (\Delta \mathbf{1}_{M/2}, -\Delta \mathbf{1}_{M/2})$, the Pareto answer is $\{1, 2\}$, whereas the lexicographic answer is $\{1\}$. Since

$$S_+(d) = S_-(d) = \frac{M}{2}\Delta^2, \qquad d_1^2 = \Delta^2,$$

Proposition 4 gives

$$T_{\mathrm{P}}(d) = \frac{16}{M\Delta^2}, \qquad T_{\mathrm{lex}}(d) = \frac{8}{\Delta^2}, \qquad \frac{T_{\mathrm{lex}}(d)}{T_{\mathrm{P}}(d)} = \frac{M}{2}.$$

This proves Eq. (7).

For $d = (1, \ldots, 1, \epsilon)$ with $0 < \epsilon < 1$, both orders identify arm 1. Now $\min_m d_m^2 = \epsilon^2$ and $d_1^2 = 1$, so the same proposition yields

$$T_{\mathrm{P}}(d) = \frac{8}{\epsilon^2}, \qquad T_{\mathrm{lex}}(d) = 8, \qquad \frac{T_{\mathrm{lex}}(d)}{T_{\mathrm{P}}(d)} = \epsilon^2.$$

This proves Eq. (8). □

F Proof of Proposition 2

Proof. For coordinate feedback, write $q = (q_{i,m}) \in \Delta_{KM}$ for the sampling proportions and define

$$\mathcal{I}^{*,\mathrm{co}}_{\preceq}(\boldsymbol{\mu}) = \sup_{q \in \Delta_{KM}} \inf_{\boldsymbol{\nu} \in \mathrm{Alt}_{\preceq}(\boldsymbol{\mu})} \frac{1}{2} \sum_{i=1}^{K} \sum_{m=1}^{M} q_{i,m} (\mu_{i,m} - \nu_{i,m})^2. \tag{23}$$

Its reciprocal is $T^{\mathrm{co}}_{\preceq}(\boldsymbol{\mu})$.

General feedback bounds. For any $q \in \Delta_{KM}$, define $w_i = \sum_m q_{i,m}$. Then $w \in \Delta_K$ and, against every feasible alternative,

$$\frac{1}{2} \sum_{i,m} w_i (\mu_{i,m} - \nu_{i,m})^2 \geq \frac{1}{2} \sum_{i,m} q_{i,m} (\mu_{i,m} - \nu_{i,m})^2.$$

Taking infima over alternatives and optimizing over allocations shows that the characteristic information under vector feedback is at least that under coordinate feedback. Conversely, any vector allocation w induces the coordinate allocation $q_{i,m} = w_i/M$, which collects exactly a $1/M$ fraction of its information against every alternative. Thus,

$$\frac{1}{M} \mathcal{I}^{*}_{\preceq}(\boldsymbol{\mu}) \leq \mathcal{I}^{*,\mathrm{co}}_{\preceq}(\boldsymbol{\mu}) \leq \mathcal{I}^{*}_{\preceq}(\boldsymbol{\mu}).$$

Local stability and a nonempty alternative set ensure that the vector-feedback characteristic information is positive and finite, so taking reciprocals proves Eq. (9). We next verify the two endpoint examples on $\Theta(\mathcal{D}_{\mathrm{free}})$.

Information for one coordinate comparison. For two arms, let $p_m = q_{1,m} + q_{2,m}$. When $d_m > 0$, the minimal information needed to reverse the comparison on coordinate m is

$$c_m(q) = \inf_{a \leq b} \left\{ \frac{q_{1,m}}{2} (\mu_{1,m} - a)^2 + \frac{q_{2,m}}{2} (\mu_{2,m} - b)^2 \right\} = \frac{q_{1,m} q_{2,m}}{2p_m} d_m^2,$$

where the right-hand side is defined as zero if $p_m = 0$. If both weights are positive, the minimizing values coincide at $(q_{1,m}\mu_{1,m} + q_{2,m}\mu_{2,m})/p_m$. If a weight vanishes, the corresponding arm mean can be moved at no cost, which gives the same formula. Strict reversals approach this infimum. Since $q_{1,m} q_{2,m} \leq p_m^2/4$,

$$c_m(q) \leq \frac{p_m d_m^2}{8}, \tag{24}$$

with equality at $q_{1,m} = q_{2,m} = p_m/2$. Balancing the arms within each coordinate achieves these equalities simultaneously without changing p.

Pareto identification. At $d = \Delta \mathbf{1}_M$, the answer is $\{1\}$. Every alternative reverses at least one coordinate (in the closure), and reversing a single coordinate suffices to change the answer. Hence, for fixed q, the information infimum is $\min_m c_m(q)$. Using Eq. (24),

$$\mathcal{I}^{*,\mathrm{co}}_{\mathrm{P}}(d) = \frac{\Delta^2}{8} \max_{p \in \Delta_M} \min_m p_m = \frac{\Delta^2}{8M}.$$

The last equality follows from $\min_m p_m \leq 1/M$, with equality at the uniform allocation. Thus, $T^{\mathrm{co}}_{\mathrm{P}}(d) = 8M/\Delta^2$.

Lexicographic identification. An answer-changing difference must have $z_1 \leq 0$, and a strict reversal of the first coordinate alone is feasible. The information infimum is therefore $c_1(q)$, giving

$$\mathcal{I}^{*,\mathrm{co}}_{\mathrm{lex}}(d) = \frac{\Delta^2}{8} \max_{p \in \Delta_M} p_1 = \frac{\Delta^2}{8}.$$

This value is attained by assigning mass $1/2$ to each of the two first-coordinate experiments. Hence, $T^{\mathrm{co}}_{\mathrm{lex}}(d) = 8/\Delta^2$.

Vector feedback and endpoint attainment. At $d = \Delta \mathbf{1}_M$, Proposition 4 gives $T_{\mathrm{P}}(d) = T_{\mathrm{lex}}(d) = 8/\Delta^2$ on $\Theta(\mathcal{D}_{\mathrm{free}})$. Combining these identities with the coordinate-feedback characteristic times proves Eq. (10) and shows that both bounds in Eq. (9) are attained. □

G Proof of Proposition 3

Proof. We use the model-specific characteristic times from Proposition 4. In particular, the third line of Eq. (11) uses the common-ranking model $\Theta(\mathcal{D}_{\mathrm{cmp}})$ specified in Eq. (19). The scalar benchmark is

$$T_{\mathrm{best}}(d) = \frac{8}{\Delta^2_{\max}(d)}.$$

Together with Lemma 1, this gives the identity

$$\mathsf{Gain}_{\preceq}(d) = \frac{\mathrm{dist}^2(d, \mathrm{Alt}_{\preceq}(d))}{\Delta^2_{\max}(d)},$$

where the alternative set is taken in the relevant model.

Pareto preferences. The condition $\min\{S_+(d), S_-(d)\} > \Delta^2_{\max}(d)$ implies that both $P(d)$ and $N(d)$ are nonempty. Equation (20) therefore yields

$$\mathsf{Gain}_{\mathrm{P}}(d) = \frac{8/\Delta^2_{\max}(d)}{8/\min\{S_+(d), S_-(d)\}} = \frac{\min\{S_+(d), S_-(d)\}}{\Delta^2_{\max}(d)} > 1.$$

This proves the first line of Eq. (11). For incomparable arms, $d_m^2 \le \Delta^2_{\max}(d)$ gives

$$\mathsf{Gain}_{\mathrm{P}}(d) \le \min\{|P(d)|, |N(d)|\} \le \left\lfloor \frac{M}{2} \right\rfloor.$$

The upper bound is attained by $d = \Delta(\mathbf{1}_{\lfloor M/2 \rfloor}, -\mathbf{1}_{\lceil M/2 \rceil})$ for any $\Delta > 0$ and $M \ge 2$. For comparable arms, Eq. (20) instead gives

$$\mathsf{Gain}_{\mathrm{P}}(d) = \frac{\min_m d_m^2}{\Delta^2_{\max}(d)} \le 1.$$

Hence, the condition in the proposition characterizes strict Pareto acceleration on the unrestricted comparison model, and such acceleration requires $M \ge 4$.

Linear scalarization. The condition $|v^\top d| > \|v\|_2 \Delta_{\max}(d)$ implies $v^\top d \neq 0$. Equation (21) gives

$$\mathsf{Gain}_v(d) = \frac{8/\Delta^2_{\max}(d)}{8\|v\|_2^2/(v^\top d)^2} = \frac{(v^\top d)^2}{\|v\|_2^2 \Delta^2_{\max}(d)} > 1,$$

proving the second line of Eq. (11). Moreover, since $v \ge 0$,

$$|v^\top d| \le \|v\|_1 \Delta_{\max}(d), \qquad \mathsf{Gain}_v(d) \le \frac{\|v\|_1^2}{\|v\|_2^2} \le M.$$

For a fixed v, the first bound is attained at $d = \Delta \mathbf{1}_M$; if, in addition, the weights are equal and positive, the gain equals M. Thus, the maximum gain for a fixed weight vector is $\|v\|_1^2/\|v\|_2^2$, and the global factor-M bound is attainable.

Lexicographic preferences under common ranking. For $d \in \mathcal{D}_{\mathrm{cmp}}$, the second identity in Eq. (22) yields

$$\mathsf{Gain}_{\mathrm{lex}}(d) = \frac{8/\Delta^2_{\max}(d)}{8/\|d\|_2^2} = \frac{\|d\|_2^2}{\Delta^2_{\max}(d)}.$$

All coordinates are nonzero, so for $M \ge 2$,

$$1 < \frac{\|d\|_2^2}{\Delta^2_{\max}(d)} \le M.$$

The upper bound is attained when all coordinate gaps have the same magnitude. This proves the third line of Eq. (11) and the stated factor-M gain. By contrast, the first identity in Eq. (22) gives $\mathsf{Gain}_{\mathrm{lex}}(d) = d_1^2/\Delta_{\max}^2(d) \le 1$ on the unrestricted model. Thus, the lexicographic acceleration is a consequence of the common-ranking constraint on feasible alternatives, not merely of the signs of the true gaps.

Finally, the common-ranking restriction does not change the single-coordinate benchmark: every two-arm scalar mean pair with a nonzero gap can be extended to an instance in $\Theta(\mathcal{D}_{\mathrm{cmp}})$. Hence, for each objective, the induced scalar model remains unrestricted apart from the exclusion of ties. The minimum in $T_{\mathrm{best}}(d)$ is an instance-wise oracle benchmark; we do not claim that an unknown easiest coordinate can be selected without additional sampling. □

H Proof of Theorem 2

Proof. Fix $\boldsymbol{\mu} \in \Theta$, and write $A^* = A_{\preceq}(\boldsymbol{\mu})$, $g = g_{A^*}$, and $\Gamma = \mathcal{I}^*_{\preceq}(\boldsymbol{\mu})$. We analyze the sampling process continued indefinitely, ignoring the stopping rule; for each confidence level, the actual algorithm coincides with this process up to its first threshold crossing. We use the exact allocation maximizer in Algorithm 1 and resolve all ties by fixed measurable rules that do not depend on δ, so that the confidence level enters only through the stopping threshold. Local stability is used in the metric form of Eq. (2).

Step 1: continuity and positivity of the information. Let $S \subseteq \mathbb{R}^{K\times M}$ be a fixed nonempty set, and define

$$h(\boldsymbol{x}, w) = \inf_{\boldsymbol{\nu}\in S} \frac{1}{2}\sum_i w_i \|x_i - \nu_i\|_2^2, \qquad \boldsymbol{x} \in \mathbb{R}^{K\times M}, \quad w \in \Delta_K.$$

This function is nonnegative and finite, since it is bounded above by the cost of any fixed $\boldsymbol{\nu}^0 \in S$. It is concave in w, as an infimum of linear functions.

For fixed $\boldsymbol{x}$, it is upper semicontinuous in w, as an infimum of continuous linear functions. To prove lower semicontinuity, let $w^{(n)} \to w$ and set

$$\eta_n = \max_{i:w_i>0} \left(1 - \frac{w_i^{(n)}}{w_i}\right)_+ .$$

Then $\eta_n \to 0$, and $w_i^{(n)} \ge (1-\eta_n)w_i$ for all i, including those with $w_i = 0$. Since each per-arm cost is nonnegative,

$$h(\boldsymbol{x}, w^{(n)}) \ge (1-\eta_n)h(\boldsymbol{x}, w).$$

Thus, $h(\boldsymbol{x}, \cdot)$ is continuous on the closed simplex. Moreover, the weighted seminorm $\|z\|_w = (\sum_i w_i \|z_i\|_2^2)^{1/2}$ and the triangle inequality give

$$\left|\sqrt{2h(\boldsymbol{x}, w)} - \sqrt{2h(\boldsymbol{y}, w)}\right| \le \|\boldsymbol{x} - \boldsymbol{y}\|_w \le \|\boldsymbol{x} - \boldsymbol{y}\|_F.$$

Combining this uniform bound with continuity in w yields joint continuity in $(\boldsymbol{x}, w)$. Note that no closedness, convexity, or boundedness of S is required. For any $A \in \mathcal{A}$, take $S = \Theta \setminus \Theta_A$ and write the answer-specific information as

$$g_A(\boldsymbol{x}, w) = \inf_{\boldsymbol{\nu}\in\Theta\setminus\Theta_A} \frac{1}{2}\sum_i w_i \|x_i - \nu_i\|_2^2.$$

At the empirical input $\boldsymbol{x} = \widehat{\boldsymbol{\mu}}_t$, its ith row is $x_i = \widehat{\mu}_i(t)$. The preceding argument establishes the required continuity and concavity of g_A, so its maximum over the compact simplex is attained. Each answer-region distance is also continuous in $\boldsymbol{x}$, so the minimization over the finitely many answers is well defined. Consequently, fixed measurable selectors can be used for both optimization steps.

Local stability and the nonempty alternative set give

$$\rho = \inf_{\boldsymbol{\nu}\in\Theta\setminus\Theta_{A^*}} \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_F \in (0, \infty).$$

Evaluating the information at the uniform allocation, and at any fixed alternative $\boldsymbol{\nu}^0$, gives

$$0 < \frac{\rho^2}{2K} \le \Gamma \le \frac{1}{2}\max_i \|\mu_i - \nu_i^0\|_2^2 < \infty.$$

Step 2: a confidence event valid under adaptive sampling. For each arm i, we represent its observations by an independent infinite sequence of potential observations, and let $\widehat{\mu}_i[n]$ denote the mean of the first n of them. For any deterministic count vector $n = (n_1, \dots, n_K)$ with $n_i \geq 1$,

$$Q(n) = \frac{1}{2} \sum_i n_i \|\widehat{\mu}_i[n_i] - \mu_i\|_2^2 \sim \frac{1}{2}\chi^2_D, \qquad D = KM.$$

For $0 < \lambda < 1$,

$$\log \mathbb{E} \exp\{\lambda(Q(n) - D/2)\} = \frac{D}{2}\big[-\log(1-\lambda) - \lambda\big] \leq \frac{D\lambda^2}{4(1-\lambda)}.$$

Chernoff's inequality with $\lambda = 2\sqrt{x}/(\sqrt{D} + 2\sqrt{x})$ then gives

$$\mathbb{P}\left(Q(n) > \frac{D}{2} + \sqrt{Dx} + x\right) \leq e^{-x}, \qquad x > 0. \tag{25}$$

At time t, there are at most $(t+1)^K$ possible count vectors. Applying Eq. (25) with $x = \log(2(t+1)^{K+2}/\delta)$ to each of them, before the adaptive count vector is selected, and taking a union bound over count vectors and times gives

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \geq K : \frac{1}{2}\sum_i N_i(t)\|\widehat{\mu}_i(t) - \mu_i\|_2^2 > \beta(t,\delta)\right) \\ \leq \frac{\delta}{2}\sum_{t \geq K}(t+1)^{-2} \leq \delta.\end{aligned} \tag{26}$$

This argument does not require the empirical means to remain Gaussian after conditioning on the adaptively chosen counts.

If $\widehat{A}_t \neq A^*$, the true instance belongs to $\Theta \setminus \Theta_{\widehat{A}_t}$, so

$$Z_t \leq \frac{1}{2}\sum_i N_i(t)\|\widehat{\mu}_i(t) - \mu_i\|_2^2.$$

Therefore, an incorrect recommendation at a threshold crossing can occur only outside the confidence event in Eq. (26), so the probability of an incorrect stop is at most δ. Almost-sure stopping is established in Step 6.

Step 3: bounded tracking discrepancy and forced exploration. Let

$$S_i(t) = 1 + \sum_{s=K}^{t-1} q_{s,i}, \qquad b_i(t) = S_i(t) - N_i(t).$$

We have $b_i(K) = 0$ and $\sum_i b_i(t) = 0$. Before the pull at time $t+1$, the deficits are $b_i(t) + q_{t,i}$ and sum to one, so the selected deficit is at least $1/K$ before it is reduced by one, while every unselected deficit can only increase. By induction, $b_i(t) > -1$ for all i and t; since the deficits sum to zero, also $b_i(t) < K - 1$. Thus,

$$|N_i(t) - S_i(t)| \leq K - 1, \qquad \left\|\frac{N(t)}{t} - \frac{S(t)}{t}\right\|_1 \leq \frac{K(K-1)}{t}. \tag{27}$$

The exploration mixture satisfies $q_{s,i} \geq s^{-1/2}/K$, whence

$$N_i(t) \geq 1 + \frac{1}{K}\sum_{s=K}^{t-1} s^{-1/2} - (K-1). \tag{28}$$

Consequently, there exist deterministic constants $c_0 > 0$ and $t_0 \geq K$, depending only on K, such that $N_i(t) \geq c_0\sqrt{t}$ for all i and all $t \geq t_0$.

Step 4: an integrable time after which the estimates remain accurate. Fix $r > 0$. If $\|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_F > r$, then at least one scalar coordinate has error greater than $r/\sqrt{KM}$. For a deterministic sample size n, the Gaussian tail bound gives

$$\mathbb{P}_{\boldsymbol{\mu}}\left(|\widehat{\mu}_{i,m}[n] - \mu_{i,m}| > \frac{r}{\sqrt{KM}}\right) \le 2\exp\left(-\frac{nr^2}{2KM}\right).$$

Using Eq. (28) and a union bound over all possible counts yields constants $C_r, c_r > 0$ such that

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\mu}}(\|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_F > r) &\le 2KM \sum_{n \ge \lceil c_0\sqrt{t}\rceil} \exp\left(-\frac{nr^2}{2KM}\right) \\ &\le C_r e^{-c_r\sqrt{t}}, \qquad t \ge t_0.\end{aligned}$$

Let L_r be one plus the last integer $t \ge t_0$ at which this error exceeds r, with $L_r = t_0$ if no such time exists. Then

$$\mathbb{P}_{\boldsymbol{\mu}}(L_r > u) \le \sum_{t \ge u} C_r e^{-c_r\sqrt{t}} \quad (u \ge t_0),$$

and

$$\mathbb{E}_{\boldsymbol{\mu}}[L_r] \le t_0 + \sum_{t \ge t_0} (t+1) C_r e^{-c_r\sqrt{t}} < \infty.$$

For $r < \rho/3$, an estimate within distance r of $\boldsymbol{\mu}$ is at distance at most r from Θ_{A^*} and at distance at least $\rho - r > r$ from every other answer region. Hence, the exact answer-region fit in Eq. (12) selects $\widehat{A}_t = A^*$ for all $t \ge L_r$. This argument also covers estimates outside Θ.

Step 5: information accumulation after an integrable time. Fix $\varepsilon \in (0, \Gamma)$. By joint continuity (Step 1) and compactness of the simplex, we can choose $r \in (0, \rho/3)$ such that

$$\sup_{w \in \Delta_K} |g(\boldsymbol{x}, w) - g(\boldsymbol{\mu}, w)| \le \frac{\varepsilon}{8} \quad \text{whenever } \|\boldsymbol{x} - \boldsymbol{\mu}\|_F \le r.$$

Choose a deterministic $t_1 \ge t_0$ such that $t^{-1/2}\Gamma \le \varepsilon/8$ for all $t \ge t_1$, and let $L = \max\{L_r, t_1\}$. Then $\mathbb{E}_{\boldsymbol{\mu}}[L] < \infty$. For $t \ge L$, the candidate answer is A^*, and the exact allocation rule gives

$$\begin{aligned}g(\boldsymbol{\mu}, w_t) &\ge g(\widehat{\boldsymbol{\mu}}_t, w_t) - \frac{\varepsilon}{8} \\ &= \max_{w \in \Delta_K} g(\widehat{\boldsymbol{\mu}}_t, w) - \frac{\varepsilon}{8} \\ &\ge \Gamma - \frac{\varepsilon}{4}.\end{aligned}$$

Let $u_K = \mathbf{1}_K/K$. By concavity and nonnegativity,

$$g(\boldsymbol{\mu}, q_t) \ge (1 - t^{-1/2}) g(\boldsymbol{\mu}, w_t) + t^{-1/2} g(\boldsymbol{\mu}, u_K) \ge \Gamma - \frac{3\varepsilon}{8}.$$

Since $S(t)/t = (K/t)u_K + t^{-1}\sum_{s=K}^{t-1} q_s$, another application of concavity gives

$$g\left(\boldsymbol{\mu}, \frac{S(t)}{t}\right) \ge \frac{1}{t}\sum_{s=L}^{t-1} g(\boldsymbol{\mu}, q_s) \ge \left(1 - \frac{L}{t}\right)\left(\Gamma - \frac{3\varepsilon}{8}\right).$$

For $t \ge 8\Gamma L/\varepsilon$, this is at least $\Gamma - \varepsilon/2$. By Eq. (27) and the uniform continuity of $g(\boldsymbol{\mu}, \cdot)$ on the simplex, there exists a deterministic $t_2 \ge t_0$ such that

$$\left|g\left(\boldsymbol{\mu}, \frac{N(t)}{t}\right) - g\left(\boldsymbol{\mu}, \frac{S(t)}{t}\right)\right| \le \frac{\varepsilon}{8} \qquad (t \ge t_2).$$

Define

$$H_\varepsilon = \max\left\{t_2, \left\lceil \frac{8\Gamma L}{\varepsilon} \right\rceil\right\}.$$

Then $H_\varepsilon \ge L$, $\mathbb{E}_{\boldsymbol{\mu}}[H_\varepsilon] < \infty$, and for all $t \ge H_\varepsilon$,

$$\frac{Z_t}{t} = g\left(\widehat{\boldsymbol{\mu}}_t, \frac{N(t)}{t}\right) \ge \Gamma - \frac{\varepsilon}{2} - \frac{\varepsilon}{8} - \frac{\varepsilon}{8} \ge \Gamma - \varepsilon. \tag{29}$$

The random time H_ε does not depend on δ. This argument relies on the average information value of the allocations rather than on the convergence of individual maximizers, and therefore does not require a unique optimal allocation.

Step 6: finite expectation and the asymptotic upper bound. For fixed δ, the threshold satisfies $\beta(t,\delta)/t \to 0$, so there exists a deterministic $b_\delta \geq K$ such that $\beta(t,\delta) < (\Gamma - \varepsilon)t$ for every integer $t \geq b_\delta$. Equation (29) gives

$$\tau_\delta \leq \max\{H_\varepsilon, b_\delta\}.$$

This proves that the algorithm stops almost surely and has a finite expected stopping time. Together with Step 2, this establishes δ-correctness.

To obtain the leading constant, set $\ell_\delta = \log(1/\delta)$, fix $\zeta > 0$, and define

$$n_\delta = \left\lceil \frac{(1+\zeta)\ell_\delta}{\Gamma - \varepsilon} \right\rceil.$$

For fixed K, M,

$$\log \frac{2(n_\delta+1)^{K+2}}{\delta} = \ell_\delta + O(\log \ell_\delta), \qquad \beta(n_\delta, \delta) = \ell_\delta + o(\ell_\delta).$$

Hence, $\beta(n_\delta,\delta) < (\Gamma - \varepsilon)n_\delta$ for sufficiently small δ, and this inequality continues to hold for every $t \geq n_\delta$. Indeed, extending the threshold to real $t \geq K$,

$$t\,\partial_t \beta(t,\delta) \leq (K+2)\left(1 + \frac{\sqrt{D}}{2\sqrt{\log \frac{2(t+1)^{K+2}}{\delta}}}\right),$$

$$\beta(t,\delta) \geq \log \frac{2(t+1)^{K+2}}{\delta} \geq \ell_\delta.$$

For sufficiently small δ, the second quantity exceeds the first uniformly in t, so $\beta(t,\delta)/t$ is decreasing in t. Combining this with Eq. (29) gives

$$\tau_\delta \leq \max\{H_\varepsilon, n_\delta\}, \qquad \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \leq n_\delta + \mathbb{E}_{\boldsymbol{\mu}}[H_\varepsilon].$$

It follows that

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq \frac{1+\zeta}{\Gamma - \varepsilon}.$$

Letting $\zeta \downarrow 0$ and then $\varepsilon \downarrow 0$ yields the upper bound $1/\Gamma = T_\preceq(\boldsymbol{\mu})$. Theorem 1 provides the matching lower bound, which establishes Eq. (16).

□