TAMING PARETO REGRET WITH MARGINAL REGRET IN STOCHASTIC MULTI-OBJECTIVE BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic multi-objective bandits generalize bandits to d-dimensional reward vectors. Performance is commonly measured by Pareto regret, which sums the Pareto gaps of the selected arms, where each gap is the smallest uniform lift that makes the arm nondominated. As this measure is defined through the Pareto front, existing methods typically minimize it by estimating the front and thus inherit the difficulties of doing so. However, we find a direct link between Pareto regret and marginal regret in bandits that bypasses Pareto-front estimation. Beyond this, under regularity conditions, it even reaches the smaller scale of marginal regret. We formalize this idea through a width-guided method that maintains upper and lower confidence bounds for each arm and objective, races the top two arms within each objective, and commits once an objective has a strictly separated unique winner. For bounded i.i.d. instances with largest objective-wise suboptimality gap $g^\dagger > 0$, the method achieves Pareto regret $O(K \log T/g^\dagger)$ over horizon T with K arms, with no explicit multiplicative dependence on d in the leading gap-dependent term. We further construct an explicit subclass on which every uniformly good policy incurs $\Omega(K \log T/g^\dagger)$ regret asymptotically, establishing optimality. Experiments on synthetic and real-world datasets support the theory. Across the data families, our method reduces Pareto regret by orders of magnitude relative to the baselines. Moreover, these results empirically expose what Pareto regret omits: Pareto front recovery and fairness remain distinct goals and call for further study.

1 INTRODUCTION

Standard multi-armed bandits model sequential decisions under uncertainty with scalar rewards (Auer et al., 2002a;b), with applications ranging from healthcare (Zhou et al., 2023; Mate et al., 2020) to recommendation systems (Barraza-Urbina & Glowacka, 2020; Zhu & Van Roy, 2023). Many decision problems, however, involve several outcomes simultaneously. Multi-objective multi-armed bandits (MO-MABs) extend the bandit framework to such settings (Drugan & Nowe, 2013; Drugan et al., 2014; Xu & Klabjan, 2023; Hüyük & Tekin, 2021; Busa-Fekete et al., 2017): at each round, the learner chooses an arm and observes a d-dimensional reward vector. In the stochastic setting, each arm's reward vector follows a fixed distribution with mean μ_a. Since the objectives may conflict, no arm need be best on every objective. Pareto regret widely measures performance relative to the Pareto front of the true mean reward vectors. For arm a, its Pareto gap Δ_a^{P} is the infimum over uniform lifts $\epsilon \geq 0$ such that $\mu_a + \epsilon \mathbf{1}$ is not dominated by any other arm (Drugan & Nowe, 2013; Drugan et al., 2014; Lu et al., 2019; Tekin & Turğay, 2018; Turgay et al., 2018); Pareto regret is $R_T^{\mathrm{P}} = \sum_{t=1}^{T} \Delta_{A_t}^{\mathrm{P}}$.

A natural intuition in stochastic MO-MABs is that the problem should become harder as the number of objectives grows (Drugan & Nowe, 2013; Drugan et al., 2014; Lu et al., 2019). As d grows, Pareto dominance must be resolved across more objectives, and the Pareto front can encode increasingly many tradeoffs and require accurate estimation of all mean vectors simultaneously. Existing bounds on Pareto regret (Drugan & Nowe, 2013) also reflect this complexity. From this analysis, one might expect Pareto regret to carry a dependence on the ambient multi-objective complexity and to increase with dimensionality. *But must Pareto regret inherit the full geometric complexity of the Pareto front?*

Recent work by Xu & Klabjan (2023) shows that, in adversarial MO-MABs where reward vectors are arbitrarily chosen, Pareto regret can scale at the same order as single-objective bandit regret and can be linked to minimal (optimal) marginal regret. This suggests that *Pareto regret need not inherit*

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.

the full apparent complexity of the ambient multi-objective structure. However, this leaves a subtle question rather than a settled conclusion for the stochastic setting studied in Drugan & Nowe (2013); Drugan et al. (2014); Lu et al. (2019). First, the adversarial result concerns minimax worst-case rates over the horizon and adopts a regret notion different from that in stochastic multi-objective bandits. Second, although the connection between Pareto regret and marginal regret is conceptually suggestive, existing methods may still scale in theory with the largest, rather than the optimal, scale of marginal regret across dimensions. What is missing in the stochastic setting is an analogue that reveals which, as well as how, objective-wise scale fundamentally governs Pareto regret and how to attain the optimal scale. This distinction is crucial: it determines whether Pareto regret is driven by the hardest objective-wise gap or can instead attain the easiest marginal scale. Section A.2 presents a thorough review of related work. The main research problem is therefore

How does Pareto regret relate to marginal regret, and can this relationship be exploited to achieve the easiest marginal-regret scale, avoiding the full cost in Drugan & Nowe (2013)?

Our finding is that Pareto regret can be tamed through marginal regret in single-objective bandits. Marginal regret in objective j is $R_T^{(j)} = \sum_{t=1}^{T}(\max_a \mu_a^{(j)} - \mu_{A_t}^{(j)})$. Let g_j be the gap between the largest and second-largest means in objective j, and let $g^\dagger = \max_j g_j$. Thus, $g^\dagger > 0$ means that at least one objective has a strictly separated unique winner. On instances where at least one objective has a strictly separated unique winner, our method achieves Pareto regret $O(K \log T/g^\dagger)$, matching the scale of the easiest marginal problem. This extends the adversarial result (Xu & Klabjan, 2023) to stochastic MO-MAB, despite the different regret notion, and further goes beyond choosing an arbitrary coordinate (Xu & Klabjan, 2023) by identifying the easiest one.

We propose a novel algorithm that exploits this structural relationship. The key observation is that minimizing Pareto regret does not require learning the entire Pareto set; instead, it suffices to certify the unique winner of some objective and commit to that arm. The algorithm performs an adaptive search over objective-wise two-arm comparisons. For each objective, the algorithm maintains the two arms with the largest upper confidence bounds, encouraging additional exploration under objective-wise uncertainty. It then uses the gap between lower and upper confidence estimates to infer the objective-wise suboptimality gap and checks whether the leading arm is separated from the runner-up by their confidence intervals. Before such a separation occurs, the next pull is chosen from the two arms along the objective with the largest confidence-radius sum. This serves as a uniform upper-bound proxy for the objective-wise gap; within that comparison, the arm with the larger confidence radius is sampled. This creates a two-level exploration–exploitation mechanism: across objectives, the algorithm explores the comparison with the largest uncertainty and potential gap, while exploiting the proxy for the objective-wise gap; within that objective, it samples the more uncertain arm. The algorithm commits at the first strict separation in any objective.

Additionally, on instances with $g^\dagger > 0$, we prove a finite-time Pareto regret bound of order $O(K \log T/g^\dagger)$, where $g^\dagger$ denotes the largest objective-wise suboptimality gap. The leading term has no explicit multiplicative dependence on the number of objectives and depends on this objective-wise suboptimality gap rather than on a sum over dominated-arm Pareto gaps. We also prove an asymptotic lower bound of order $\Omega(K \log T/g^\dagger)$ on a natural duplicated-coordinate Bernoulli subclass of the full model. Consequently, the order $K \log T/g^\dagger$ cannot be improved in general by any uniformly good policy, and our upper bound is tight up to constant factors. As a byproduct, it yields a high-probability guarantee that the committed arm is also Pareto-optimal.

To examine the practical effectiveness of the proposed method, we evaluate the width-guided method on real-world Spotify and KuaiRec benchmarks and on a controlled synthetic family. The results confirm the theory across both types of datasets and further demonstrate the method's practical value against widely used baselines. On the real-data subsets, Width-guided attains the lowest Pareto regret in the reported groups, with order-of-magnitude ($\geq$ **111** times) improvements on Spotify datasets, and always selects a Pareto-optimal arm. We also report an empirical fairness measure. Although Width-guided is designed only to reduce Pareto regret and identify a Pareto-optimal arm, its fairness behavior is not uniformly worse: it loses fairness in some groups, but is surprisingly *best-of-all-worlds* in Pareto regret, Pareto optimal arm, and empirical fairness in others. This mixed behavior suggests that fairness deserves theory-oriented study, beyond as an empirical measure.

2 PROBLEM FORMULATION

We formulate the stochastic multi-objective bandit problem and introduce the notation used throughout the paper. We consider a stochastic multi-objective bandit problem with K arms and d objectives. For each arm $a \in [K] := \{1, \dots, K\}$, let $Y_{a,1}, Y_{a,2}, \dots \in [0,1]^d$ be i.i.d. reward vectors with mean $\mu_a = (\mu_a^{(1)}, \dots, \mu_a^{(d)}) \in \mathbb{R}^d$. The reward sequences are independent across arms. Dependence among the d coordinates of one reward vector is arbitrary. At each round $1 \le t \le T$ where T is time horizon, a policy (decision maker) selects an arm $A_t \in [K]$ based on the history observed so far and receives the next unused sample from that arm, $X_t := Y_{A_t, N_{A_t}(t-1)+1}$, where $N_a(t) := \sum_{s=1}^{t} \mathbf{1}\{A_s = a\}$ denotes the number of pulls of arm a up to time t. A key feature of the feedback model is that pulling an arm reveals the entire d-dimensional reward vector of that arm: all objectives of the selected arm are observed simultaneously, while no information is received about unchosen arms.

2.1 PARETO ORDER, PARETO OPTIMALITY, AND PARETO REGRET

Pareto order. We compare vector-valued rewards using the Pareto order (Drugan & Nowe, 2013). For $x, y \in \mathbb{R}^d$, write $x \succeq y$ if $x^{(j)} \ge y^{(j)}$ for all $j \in [d]$, and write $x \succ y$ if $x \succeq y$ and $x \ne y$. Thus, $x \succ y$ means that x is weakly larger than y in every coordinate and strictly larger in at least one coordinate. If neither $x \succeq y$ nor $y \succeq x$ holds, then x and y are incomparable.

Pareto optimality and regret. An arm a dominates arm b if $\mu_a \succ \mu_b$. The Pareto set is $A^\star := \{a \in [K] : \nexists b \in [K] \text{ such that } \mu_b \succ \mu_a\}$, and its elements are called Pareto-optimal arms. Following (Drugan & Nowe, 2013; Drugan et al., 2014; Lu et al., 2019; Tekin & Turğay, 2018; Turgay et al., 2018), the Pareto suboptimality gap of arm a is $\Delta_a^{\mathrm{P}} := \inf\{\epsilon \ge 0 : \mu_a + \epsilon \mathbf{1} \text{ is not dominated by any } \mu_b,\ b \in [K]\}$. Every Pareto-optimal arm has zero Pareto suboptimality gap, but the converse need not hold. Remark 2 gives a boundary example. We then measure performance by the commonly used metric of Pareto regret (Drugan & Nowe, 2013; Drugan et al., 2014; Lu et al., 2019): $R_T^{\mathrm{P}} := \sum_{t=1}^{T} \Delta_{A_t}^{\mathrm{P}} = \sum_{a:\Delta_a^{\mathrm{P}}>0} \Delta_a^{\mathrm{P}} N_a(T)$. Our goal is to minimize $\mathbb{E}[R_T^{\mathrm{P}}]$. *Our analysis relates Pareto regret to the objective-wise gaps below.*

2.2 PARETO REGRET AND OBJECTIVE-WISE GAPS

We first define the objective-wise gaps used in the regret bound. For objective j, define $\mu_{(1),j} := \max_{b\in[K]} \mu_b^{(j)}$ and $\Delta_a^{(j)} := \mu_{(1),j} - \mu_a^{(j)}$. Thus, $\Delta_a^{(j)} = 0$ means that arm a is a marginal maximizer of objective j. Any such arm has zero Pareto gap: for every $\epsilon > 0$, $\mu_a^{(j)} + \epsilon > \max_{b\in[K]} \mu_b^{(j)}$, so $\mu_a + \epsilon\mathbf{1}$ cannot be dominated by any arm.

For objective j, let $M_j := \arg\max_{a\in[K]} \mu_a^{(j)} = \{a \in [K] : \Delta_a^{(j)} = 0\}$ be the set of marginal maximizers. If M_j is a singleton, write $M_j = \{a_j^\star\}$ and define $g_j := \mu_{a_j^\star}^{(j)} - \max_{a\ne a_j^\star} \mu_a^{(j)}$. If M_j contains multiple arms, set $g_j := 0$. Thus, $g_j > 0$ exactly when objective j has a unique marginal maximizer whose objective-j mean is strictly larger than every other arm's objective-j mean. Finally, define $j^\dagger \in \arg\max_{j\in[d]} g_j$ and $g^\dagger := g_{j^\dagger}$, which is the largest objective-wise suboptimality gap. For objective j, define the marginal regret $R_T^{(j)} := \sum_{t=1}^{T} (\mu_{(1),j} - \mu_{A_t}^{(j)}) = \sum_{a=1}^{K} \Delta_a^{(j)} N_a(T)$.

For every arm a and objective j, $\Delta_a^{\mathrm{P}} \le \Delta_a^{(j)}$, because any uniform lift exceeding $\Delta_a^{(j)}$ exceeds every arm in coordinate j. Consequently, $\mathbf{R_T^{\mathrm{P}} \le R_T^{(j)}}$ **for every j**, $\mathbf{R_T^{\mathrm{P}} \le \min_j R_T^{(j)}}$. This inequality can be strict. Consider three arms with means $(0.9, 0.1)$, $(0.5, 0.5)$, and $(0.1, 0.9)$. All arms are Pareto optimal, so every policy incurs zero Pareto regret. However, $R_T^{(1)} + R_T^{(2)} = 0.8T$, so $\max\{R_T^{(1)}, R_T^{(2)}\} \ge 0.4T$. Thus, Pareto regret can be more permissive than marginal regret.

Debate and Objective. Classical gap-dependent bounds for marginal regret scale as $1/g_j$ (Lai & Robbins, 1985). The Pareto-UCB1 bound depends on the collection of dominated-arm Pareto gaps and contains d inside a logarithm (Drugan & Nowe, 2013). *Our question is whether cumulative Pareto regret can instead attain the easiest objective-wise scale* $1/\max_j g_j = 1/g^\dagger$ *when* $g^\dagger > 0$. For example, suppose $d = 2$, $g_1 = 10^{-3}$, and $g_2 = 1$. Then $g^\dagger = 1$. The classical marginal regret

scale is $1/g_1 = 10^3$ for objective 1, but only $1/g_2 = 1$ for objective 2. A bound governed by the hardest objective would scale with $1/\min_j g_j = 10^3$, possibly with an additional factor depending on d, whereas the bound in the latter case scales with $1/g^\dagger = 1$. Thus, Pareto regret follows the easier objective-wise scale rather than the hardest one. Our upper bound achieves the easier scale when $g^\dagger > 0$, and our lower bound matches it on an explicit subclass for uniformly good policies.

Motivated by the definitions of Pareto regret and marginal regret, we identify a subtle distinction between zero Pareto gap and Pareto optimality: the two notions are not equivalent. Prop. 1 shows that, although the implication $a \in A^\star \Rightarrow \Delta_a^{\mathrm{P}} = 0$ always holds, the reverse implication can fail. A tied marginal maximizer can be dominated. Remark 2 records this distinction.

Proposition 1 (Zero Pareto gap $\neq$ Pareto optimality)**.** *There exists an arm $a \notin A^\star$ with $\Delta_a^{\mathrm{P}} = 0$.*

Remark 2 (Ties and marginal-Pareto consistency). For example, if $\mu_1 = (1, 1)$ and $\mu_2 = (1, 0)$, then arm 2 is dominated by arm 1, while $\Delta_2^{\mathrm{P}} = 0$, since $(1 + \epsilon, \epsilon)$ is not dominated by $(1, 1)$ for any $\epsilon > 0$. Thus, zero Pareto gap alone does not necessarily imply Pareto optimality. The marginal-Pareto consistency property states that every marginal maximizer is Pareto-optimal: $\Delta_a^{(j)} = 0 \implies a \in A^\star, \forall a \in [K],\ j \in [d]$. Equivalently, $\bigcup_{j=1}^{d} \arg\max_{a \in [K]} \mu_a^{(j)} \subseteq A^\star$.

Remark 2 describes the problem class in which objective-wise optimality certifies a Pareto-optimal arm, thereby making marginal regret meaningful for Pareto optimality. It is weaker than requiring every zero-Pareto-gap arm to be Pareto-optimal, since it only restricts zero-gap arms that arise as marginal maximizers. It is also weaker than requiring every objective to have a unique maximizer: uniqueness implies the condition, but the condition allows ties as long as all tied marginal maximizers are Pareto-optimal. If $g^\dagger > 0$, then objective $j^\dagger$ has a unique maximizer $a^\star_{j^\dagger}$, and this arm belongs to $A^\star$ without requiring marginal-Pareto consistency: every other arm is strictly worse in that coordinate.

3 Methodology: Width-Guided Objective-Wise UCB

We propose a method that uses confidence bounds to verify that an arm has a strictly larger mean than every other arm on one objective. On the simultaneous-confidence event, such an arm belongs to $A^\star$ because every other arm is strictly worse in that coordinate, and therefore has zero Pareto suboptimality gap. Throughout, a *certificate* means the confidence-bound test that verifies this statement. With this, the policy *commits* to the certified arm and pulls it for all remaining rounds.

Before a certificate is accepted, the policy must decide which arm to sample. Algorithm 1 presents the complete pseudo code. Algorithm 1 compares arms separately within each objective using upper and lower confidence bounds. For each objective, it keeps the arms with the largest and second-largest upper confidence bounds. It then chooses the objective whose two retained arms have the largest sum of confidence radii, and pulls the arm among those two with the larger confidence radius. Since one pull reveals the full reward vector of the selected arm, this pull updates the estimates of all objectives for that arm. Thus the algorithm does not need to estimate the whole Pareto set. It stops after one objective-wise comparison verifies a unique maximizer, at which point all future pulls have zero Pareto regret on the simultaneous-confidence event. We now define the empirical quantities used to implement these comparisons. After the warm start, let $\widehat{\mu}_a^{(j)}(t-1) := {}^1\!/_{N_a(t-1)} \sum_{s=1}^{t-1} X_s^{(j)} \mathbf{1}\{A_s = a\}$ denote the empirical mean of objective j for arm a based on its first $N_a(t-1)$ samples.

At round $t > K$, define the fixed-horizon confidence radius $\beta_a(t) := \sqrt{{}^{2\log T}\!/_{N_a(t-1)}}$. For every arm a and objective j, $U_a^{(j)}(t) := \widehat{\mu}_a^{(j)}(t-1) + \beta_a(t)$, $L_a^{(j)}(t) := \widehat{\mu}_a^{(j)}(t-1) - \beta_a(t)$. For each objective j, let $b_j(t) := \arg\max_{a \in [K]} U_a^{(j)}(t)$, $c_j(t) := \arg\max_{a \in [K] \setminus \{b_j(t)\}} U_a^{(j)}(t)$, and define the pair width $W_j(t) := \beta_{b_j(t)}(t) + \beta_{c_j(t)}(t)$. If some objective j already satisfies $L_{b_j(t)}^{(j)}(t) > U_{c_j(t)}^{(j)}(t)$, then $b_j(t)$ is certified as a Pareto-optimal arm on the simultaneous-confidence event, and the algorithm stores $b_j(t)$. If no objective satisfies this inequality, the algorithm chooses an objective with largest $W_j(t)$ and samples the arm in $\{b_j(t), c_j(t)\}$ with the larger confidence radius $\beta(t)$.

In the mean update, after the played arm's count reaches n, its previous mean m becomes $m + (X_t - m)/n = ((n-1)m + X_t)/n$. We call a round $t > K$ *non-certifying* if, at the decision point of round t, no certified arm has been stored and no objective $j \in [d]$ satisfies $L_{b_j(t)}^{(j)}(t) > U_{c_j(t)}^{(j)}(t)$. At a

Algorithm 1 Width-Guided First-Certification UCB

Require: Number of arms K, number of objectives d, horizon $T \geq K$

Initialize $N_a \leftarrow 0$, $\widehat{\mu}_a \leftarrow 0 \in \mathbb{R}^d$, and stored certified arm $\widehat{a} \leftarrow \emptyset$

Warm start: pull each arm once

for $a = 1, 2, \ldots, K$ **do**

  Pull arm a and observe X_a^{init}

  $N_a \leftarrow 1$, $\widehat{\mu}_a \leftarrow X_a^{\text{init}}$

Adaptive exploration and commitment

for $t = K+1, K+2, \ldots, T$ **do**

  $\beta_a(t) \leftarrow \sqrt{2 \log T / N_a}$ for all $a \in [K]$

  $U_a^{(j)}(t) \leftarrow \widehat{\mu}_a^{(j)} + \beta_a(t)$ and $L_a^{(j)}(t) \leftarrow \widehat{\mu}_a^{(j)} - \beta_a(t)$ for all a, j

  for $j = 1, 2, \ldots, d$ **do**

    $b_j(t) \leftarrow \arg\max_{a \in [K]} (\widehat{\mu}_a^{(j)} + \beta_a(t))$, $c_j(t) \leftarrow \arg\max_{a \in [K] \setminus \{b_j(t)\}} (\widehat{\mu}_a^{(j)} + \beta_a(t))$

    $W_j(t) \leftarrow \beta_{b_j(t)}(t) + \beta_{c_j(t)}(t)$

  if $\widehat{a} \neq \emptyset$ **then**

    $A_t \leftarrow \widehat{a}$

  else if some objective j satisfies $L_{b_j(t)}^{(j)}(t) > U_{c_j(t)}^{(j)}(t)$ **then**

    choose one such objective $\widehat{j}$, store $\widehat{a} \leftarrow b_{\widehat{j}}(t)$, and set $A_t \leftarrow \widehat{a}$

  else

    $j_t \leftarrow \arg\max_{j \in [d]} W_j(t)$, $A_t \leftarrow \arg\max_{a \in \{b_{j_t}(t), c_{j_t}(t)\}} \beta_a(t)$

  Pull arm A_t and observe X_t

  $N_{A_t} \leftarrow N_{A_t} + 1$, $\widehat{\mu}_{A_t} \leftarrow \widehat{\mu}_{A_t} + {}^{(X_t - \widehat{\mu}_{A_t})}/_{N_{A_t}}$

non-certifying round, the algorithm selects $j_t \in \arg\max_{j \in [d]} W_j(t)$, $b_t := b_{j_t}(t)$, $c_t := c_{j_t}(t)$, and then pulls $A_t \in \arg\max_{a \in \{b_t, c_t\}} \beta_a(t)$.

Remark 3 (Comparison). Existing methods use confidence bounds to establish different types of guarantees. Pareto-UCB-style algorithms build an optimistic Pareto set and bound how often arms with positive Pareto suboptimality gap remain optimistically nondominated. Pareto-front identification methods decide the Pareto status of many arms. Scalar best-arm identification methods, including LUCB-style algorithms, reduce to methods with scalarized objective (Kalyanakrishnan et al., 2012; Jamieson & Nowak, 2014). Algorithm 1 aims to verify that one arm is the unique maximizer of a target objective via a two-layer exploration-exploitation trade-off across arms and objectives, where each layer involves a new design. It compares two arms within one objective, but the objective is chosen adaptively via the current width $W_j(t)$. It does not maintain an estimated Pareto set or decide the Pareto status of all arms. It stops once one objective-wise comparison verifies a unique maximizer in any objective. This is enough for Pareto regret as such an arm has zero Pareto suboptimality gap. It also ensures that Pareto regret scales inversely with the largest suboptimality gap, since the exploration sample complexity leading to the dominant regret term depends on this gap.

4 Theoretical Analyses

The finite-time analysis below shows that the leading gap-dependent Pareto regret bound is controlled by $g^\dagger$, the largest objective-wise suboptimality gap, rather than by a sum over all dominated-arm Pareto suboptimality gaps. The leading term has no explicit multiplicative factor in the number of objectives. The reason is that the algorithm stops exploration once the confidence bounds verify a unique maximizer of one objective, which is Pareto-optimal because every other arm is strictly worse in that objective. Together with the lower bound below, this shows that the standard Pareto regret can have the same leading order as a single-objective bandit. We refer to Appendix A.6 for the proofs.

4.1 Lower Bound Analysis

Before stating the finite-time upper bound, we record a lower bound through a simple duplicated-coordinate subclass. This subclass is still d-dimensional and is part of the full stochastic MO-MAB model. The lower bound has a uniform full-class role: any uniformly good policy with a smaller leading order uniformly over the full stochastic MO-MAB model would also have to improve on this subclass. This is the lower-bound notion matched to this Pareto-regret comparison. It does not

claim that every stochastic MO-MAB instance has instance complexity $K \log T/g^\dagger$, nor that $g^\dagger$ is the only relevant geometric quantity for all possible Pareto-optimal-arm certificates. Rather, the relevant obstruction is already present inside the model. In particular, the statement is not that every geometry is governed by $g^\dagger$, but that the full model already contains instances with a matching lower bound. Consequently, The $O(K \log T/g^\dagger)$ order cannot be improved uniformly over the class.

Proposition 4. *Assume $K \geq 2$ and fix $\Delta_{\mathrm{sc}} \in (0, 1/4]$. Consider the duplicated-coordinate Bernoulli instance $Y_{a,n} = (Z_{a,n}, \ldots, Z_{a,n}) \in [0,1]^d$, where $Z_{1,n} \sim \mathrm{Bernoulli}(\frac{1}{2} + \Delta_{\mathrm{sc}})$, $Z_{a,n} \sim \mathrm{Bernoulli}(\frac{1}{2})$ for $a = 2, \ldots, K$, and all latent samples are independent across arms and pulls. Suppose moreover that, on this duplicated-coordinate subclass, the induced scalar policy obtained by identifying each observed vector $(Z_{a,n}, \ldots, Z_{a,n})$ with its underlying Bernoulli sample $Z_{a,n}$ is uniformly good in the Lai–Robbins sense: for every Bernoulli mean vector $\theta \in (0,1)^K$ and every $\alpha > 0$, its expected single-objective regret under θ is $o(T^\alpha)$ as $T \to \infty$. Then, because $g^\dagger = \Delta_{\mathrm{sc}}$ on this instance, $\mathbb{E}[R_T^{\mathrm{P}}] = \Omega(\frac{K \log T}{g^\dagger})$ as $T \to \infty$. More precisely, $\liminf_{T\to\infty} \frac{\mathbb{E}[R_T^{\mathrm{P}}]}{\log T} \geq \frac{3}{8}\frac{K-1}{\Delta_{\mathrm{sc}}}$.*

Proposition 4 shows that multi-objective bandits contain instances with lower bound $\Omega(K^{\log T}/_{g^\dagger})$. Therefore, no uniformly good policy can improve the order $K^{\log T}/_{g^\dagger}$ over the full model class. The upper bound in Theorem 5 holds for every instance with $g^\dagger > 0$, while the lower bound establishes matching leading order on the explicit subclass for the stated uniformly good policies. This is not an arbitrary-instance characterization or a minimax statement. Duplicating coordinates reveals only one Bernoulli sample per pull, so the scalar information lower bound transfers to this subclass.

4.2 Upper Bound Analysis

Theorem 5 (Finite-time bound for objective-wise Pareto-optimal certification)**.** *Under the stochastic model above, suppose $K \geq 2$, $T \geq K$, rewards are bounded in $[0,1]$, and the instance has positive first-certificate scale $g^\dagger > 0$. Let $\Delta_{\max}^{\mathrm{P}} := \max_{a\in[K]} \Delta_a^{\mathrm{P}}$. Then Algorithm 1 satisfies $\mathbb{E}[R_T^{\mathrm{P}}] \leq K\Delta_{\max}^{\mathrm{P}} + 64\,\frac{K \log T}{g^\dagger} + 2Kd\,\Delta_{\max}^{\mathrm{P}} T^{-2}$. In particular, $\mathbb{E}[R_T^{\mathrm{P}}] = O(\frac{K \log T}{g^\dagger})$.*

The theorem is a guarantee for the objective-wise Pareto-optimal certificate family, not a characterization of all Pareto geometries. When $g^\dagger > 0$, this route supplies a statistically separated certificate and Algorithm 1 reaches the single-objective-like scale above. When $g^\dagger = 0$, Pareto regret remains well defined, but this particular objective-wise route has no positive separation.

The key feature of Theorem 5 lies in its leading term: apart from the displayed bad-event tail $2Kd\Delta_{\max}^{\mathrm{P}} T^{-2}$, there is no explicit multiplicative dependence on the number of objectives, and the first-certificate scale $g^\dagger$ controls the finite-time upper bound for the objective-wise route used by Algorithm 1. The dimension dependence in the theorem is therefore confined to a vanishing confidence-failure term, not to the leading exploration cost. Operationally, $g^\dagger$ is not a relaxation of the Pareto benchmark; it is the gap controlling the exploration cost before the first objective-wise certificate after which the same benchmark charges zero regret. Thus, for gap-dependent Pareto regret under selected-arm vector feedback, the multi-objective problem remains single-objective-like at leading order when the first certificate has positive separation.

Remark 6 (Comparison with Pareto UCB1). Theorem 1 of (Drugan & Nowe, 2013) gives $\mathbb{E}[R_T^{\mathrm{P}}] \leq \sum_{i\notin A^\star} \frac{8\log(T(d|A^\star|)^{1/4})}{\Delta_i^{\mathrm{P}}} + (1 + \frac{\pi^2}{3}) \sum_{i\notin A^\star} \Delta_i^{\mathrm{P}}$. Consider the generic non-boundary case in which every dominated arm has positive Pareto gap; zero-gap dominated arms are regret-neutral, and if $A^\star = [K]$, the comparison is trivial. Let $\Delta_{\min}^{\mathrm{P}} := \min_{i\notin A^\star} \Delta_i^{\mathrm{P}}$. Then the leading term is at most $8(K - |A^\star|)\log(T(d|A^\star|)^{1/4})/\Delta_{\min}^{\mathrm{P}}$, which can be written on the scale of Theorem 5 as $(8 \cdot \frac{K-|A^\star|}{K} \cdot \frac{g^\dagger}{\Delta_{\min}^{\mathrm{P}}} \cdot \frac{\log(T(d|A^\star|)^{1/4})}{\log T}) \frac{K \log T}{g^\dagger}$. Thus, the Pareto-UCB1 leading term is organized through all dominated-arm Pareto gaps and retains d and $|A^\star|$ inside the logarithm, whereas our bound reaches the first-certificate scale $g^\dagger$. The full derivation is given in Appendix A.6.8.

For an exact leading-term comparison, write the Pareto-UCB1 leading term as $C_{\mathrm{PUCB}}(T,\mu) K \log T/g^\dagger$, where $C_{\mathrm{PUCB}}(T,\mu) := 8 \cdot \frac{g^\dagger}{K}(\sum_{i\notin A^\star} \frac{1}{\Delta_i^{\mathrm{P}}}) \frac{\log(T(d|A^\star|)^{1/4})}{\log T}$. Our coefficient 64 is smaller exactly when $\frac{g^\dagger}{K}\sum_{i\notin A^\star} \frac{1}{\Delta_i^{\mathrm{P}}} > 8 \cdot \frac{\log T}{\log(T(d|A^\star|)^{1/4})}$. Since $\frac{\log T}{\log(T(d|A^\star|)^{1/4})} = (1 + \frac{\log(d|A^\star|)}{4\log T})^{-1}$, the right-hand side is asymptotically close to 8. Hence, the

comparison is driven by how the dominated-arm Pareto gaps relate to $g^\dagger$. If $\Delta^{\mathrm{P}}_{\min} \ll g^\dagger$, or more generally if $\sum_{i \notin A^\star} 1/\Delta^{\mathrm{P}}_i$ is large, the Pareto-UCB1 coefficient can be much larger than 64. This is precisely the regime our method targets: one objective can be certified quickly even when many dominated arms lie close to the Pareto frontier. If the dominated-arm Pareto gaps are all of the same order as $g^\dagger$ and only a few dominated arms remain, then the Pareto-UCB1 coefficient remains constant-order and may be smaller than 64. This comparison concerns the leading $K \log T/g^\dagger$ term.

Proof Intuition. On the confidence event, each post-warm-start non-certifying pull costs at most four times its confidence radius. The best-separated objective keeps the sampled radius at least $g^\dagger/4$, limiting each arm to $O(\log T/(g^\dagger)^2)$ such pulls. Summing the radius charges and adding the warm-start cost $K\Delta^{\mathrm{P}}_{\max}$ gives the stated bound. The difficulty is that the marginal gaps are unknown and the learner chooses arms, not coordinates: exploring an objective can delay exploitation of a good arm, while each selected-arm observation tightens all its coordinates.

We derive the above upper bound using a new analytical framework. The proof differs from Pareto-UCB analyses in how regret is accounted for: Pareto UCB controls regret by bounding the number of times each dominated arm remains an optimistic Pareto candidate, whereas we focus only on pulls made before the first Pareto-optimal-arm certificate in any objective. The bound rests on a single global confidence event and three deterministic implications on that event:

- ▶ certification stops regret because the certified arm has zero Pareto gap;
- ▶ each non-certifying pull incurs only uncertainty-scale Pareto regret;
- ▶ non-certifying play cannot continue indefinitely as the champion objective enforces a width floor.

More specifically, Lemma 10 shows how a single-coordinate statistical guarantee controls Pareto regret: a certified unique maximizer of a single objective already belongs to $A^\star$. Lemma 11 then attributes each non-certifying pull to the confidence radius of the sampled endpoint in an objective-level top-two race, rather than tracking the arm's role within an estimated Pareto set. Finally, Lemma 12 shows that the easiest separated objective keeps these races active only until the relevant radii reach the $g^\dagger$ scale. Thus, the proof focuses entirely on the certificate witness rather than resolving the Pareto status of individual arms: we never need to show that every positive-gap arm eventually leaves an optimistic Pareto set. Once the first certificate is found, the remaining arms no longer affect Pareto regret. Together, these ingredients attribute all pre-certification regret to confidence radii, while the width floor converts these terms into a total $O(K \log T/g^\dagger)$ bound.

Beyond its Pareto regret guarantee, the method also identifies a Pareto-optimal arm through certification-style committed arm selection. This property is not guaranteed by Pareto UCB, which randomly selects an arm from the estimated Pareto front. The formal statement is as follows:

Corollary 7 (High-probability validity of accepted certificates)**.** *Run Algorithm 1 with the theorem-matched confidence radius $\beta_a(t) = \sqrt{2\log T/N_a(t-1)}$. Let $\widehat{a}_{\mathrm{cert}}(T) \in [K] \cup \{\emptyset\}$ denote the stored certified leader at the end of horizon T, with $\widehat{a}_{\mathrm{cert}}(T) = \emptyset$ if no certificate has been stored. Then $\Pr(\widehat{a}_{\mathrm{cert}}(T) \neq \emptyset$ and $\widehat{a}_{\mathrm{cert}}(T) \notin A^\star) \leq 2KdT^{-3}$. Equivalently, with probability at least $1 - 2KdT^{-3}$, every arm accepted by the algorithm is Pareto-optimal.*

5 Computational Study

Our empirical goal is to evaluate finite-horizon performance from two complementary angles. The real-data benchmark tests outcome metrics under realistic multi-objective reward distributions and compares the method against widely used baselines, while the synthetic study provides a stylized validation of the bound and comparison with Pareto-UCB1; synthetic results are in Appendix A.3.

Under the stated regularity condition $g^\dagger > 0$, our Pareto-regret upper bound matches the best coordinate's single-objective scale. In addition, the guarantee of identifying an optimal arm also yields the identification of a Pareto-optimal arm as a byproduct. To the best of our knowledge, fairness in general multi-objective bandits is primarily an empirical measure without a formal characterization. Motivated by this, we also examine this measure numerically. Thus, we report three measures: Pareto regret, detection of a Pareto-optimal arm, and empirical fairness. The results not only confirm and align with the theoretical message of the paper, but also reveal surprising insights and identify a potential hardness associated with distinct learning goals; details are in Appendix A.4.

Real-world Datasets. We evaluate our approach using two real-world datasets. In the Spotify instance, the arms represent music genres, and each pull samples a track from the selected genre and reveals all six attributes jointly. The six objectives correspond to popularity, danceability, energy, loudness, acousticness, and valence. In the KuaiRec instance, the arms are videos, and each objective represents a specific user cohort. The reward is defined as a normalized watch-ratio sample for that particular cohort. One simulated pull draws one stored response from each cohort, so this is an offline batched panel evaluation rather than a single user impression. The prepared instances contain 114 Spotify arms with $d = 6$ objectives and 300 KuaiRec arms with $d = 26$ objectives. For both datasets, each evaluated benchmark subset consists of $K = 10$ arms. The Spotify instance is built from the Spotify Tracks dataset on Hugging Face (Spotify Tracks Dataset); the dataset card for the mirror used in our scripts lists license `bsd`. The KuaiRec instance uses KuaiRec 2.0 (Gao et al., 2022); the release used here includes a Creative Commons Attribution-ShareAlike 4.0 International license. Rewards are bounded after normalization, and the certificate scale is strictly positive.

We select these subsets prior to policy simulation and hold them fixed throughout the evaluation. We categorize the subsets into three empirical-geometry groups. First, *misleading near-front* subsets contain attractive arms that appear close to the Pareto front but actually possess a positive Pareto gap. Second, *friendly near-front* subsets feature near-front arms that have zero or nearly zero gaps, making exploratory mistakes relatively cheap. Finally, *easy separated* subsets offer clear separation with few near-front distractors. We apply the geometry selector directly to the Spotify data, using an independently frozen geometry-selected easy-separated suite for the Spotify easy row. For KuaiRec, we use the same categorizations but employ a contamination-balanced selector to emphasize empirical-front contamination within the geometry groups.

Baselines. We compare our method against three widely recognized baselines that are commonly used in practice. Pareto UCB1 (Drugan & Nowe, 2013) forms coordinate-wise UCB vectors and samples uniformly from the nondominated set of these optimistic vectors. Annealing-Pareto (Yahyaa et al., 2014) uses an ϵ-Pareto candidate rule with the reported decay $\epsilon_s = 0.4^s/(Kd)$. Scalarized UCB applies scalar UCB to a fixed finite set of normalized nonnegative linear weights.

Performance Metrics and Implementation. Pareto regret is defined as before. The terminal recommendation, denoted as $\widehat{a}_T$, is defined as the arm most frequently played during the final 20% of rounds. The Pareto set $A^\star$ is derived from the empirical mean vectors of the held-out subset. Additionally, fairness is measured as the cumulative max-min regret: $\sum_{t=1}^{T}(\max_a \min_j \mu_a^{(j)} - \min_j \mu_{A_t}^{(j)})$. Reported results are held-out averages over a horizon of $T = 10^6$. The only hyperparameter we tune is this empirical confidence coefficient, because the theoretical radius is intentionally conservative for an *upper bound*, rather than for the *actual regret* observed in numerical experiments. For the baselines, we use the best parameter settings recommended in their original papers for a fair comparison. Further implementation and experiment details are available in Appendix A.4. The anonymous code and computational artifact are available at `https://anonymous.4open.science/r/momab-width-guided-certification-ACF7`.

Experiment Results. Table 1 reports the full results. Across these held-out real-world subsets, Width-guided achieves the lowest Pareto regret in every Spotify and KuaiRec group under the stated hyperparameter protocol. The performance gap is most pronounced in the misleading near-front and easy separated groups. In these settings, repeatedly playing positive-gap arms proves costly for methods that attempt to maintain or exploit a broader empirical front. While the friendly groups also show favorable results for our method, the interpretation differs slightly. Because near-front mistakes in these subsets are cheap, the absolute scale of regret remains small across all algorithms. Measured against the best non-Width-guided Pareto-regret mean in each block, Width-guided is lower by about two or more orders of magnitude in all three Spotify groups ($\mathbf{111}\times$, $\mathbf{223}\times$, and $\mathbf{591}\times$), by more than one order of magnitude in KuaiRec friendly $\mathbf{16}\times$, and by smaller factors in KuaiRec misleading and KuaiRec easy ($\mathbf{1.7}\times$ and $\mathbf{1.2}\times$). Thus the main pattern is order-of-magnitude regret reduction in four of six groups, with two closer KuaiRec comparisons. These scale differences are intrinsic to the dataset structures themselves; the performance gap is especially large in the Spotify groups and in KuaiRec friendly, while KuaiRec misleading and KuaiRec easy show closer comparisons. Our algorithm applies regardless and can actually identify such structures based on the results. Furthermore, Width-guided has terminal detection frequency 1.000 in every reported group, so the low-regret rows also select $\widehat{a}_T \in A^\star$ at the terminal recommendation stage.

The fairness column highlights the complementary nature of these learning objectives. In some instances (e.g., Spotify misleading, Spotify easy, and KuaiRec misleading), although Scalarized UCB suffers from substantially larger Pareto regret, it achieves smaller max-min fairness. This behavior is consistent with the role of first certification: our approach is designed to minimize Pareto regret, rather than to ensure full-front coverage or fairness-style balancing. In other instances (e.g., Spotify friendly, KuaiRec friendly, and KuaiRec easy), however, our method is *best-of-all-worlds on all three reported metrics*: Pareto regret, terminal Pareto-optimal-arm detection, and fairness.

Front coverage and recovery. The count-based coverage entropy in Appendix A.4 is 0.000–0.003 for Width-guided, showing that it does not diversify over the front when the true front has several arms. KuaiRec easy has a one-arm true front, so entropy is zero for every method. After certification, the runner accounts for Width-guided's committed tail using the exact benchmark means, while the baselines remain fully sampled. This leaves the actions, Pareto regret, detection, and count-based entropy unchanged, but makes Width-guided's front precision and recall hybrid plug-in diagnostics that should not be compared across methods. Entropy measures diversity of play, not identification correctness or sample complexity; a fully sampled front-recovery comparison remains open. Full-front identification is a complementary pure-exploration goal whose complexity depends on arm-wise gaps and front geometry rather than on $g^\dagger$ alone (Auer et al., 2016; crepon et al., 2024).

Taken together, these empirical results highlight the practical utility of our method, besides complementing the theoretical guarantees. They also suggest that fairness deserves formal study and may reflect a new source of hardness, despite of being treated largely as an empirical measure.

Table 1: Real-world benchmark on held-out Spotify and KuaiRec subsets at $T = 10^6$. Subsets are grouped before policy simulation. The main target is Pareto regret and detection of at least one Pareto-optimal arm; fairness is reported as an empirical measure.

Subset type	Method	Pareto regret ↓	$\Pr(\widehat{a}_T \in A^\star)$ ↑	Fairness regret ↓
Spotify d6				
Misleading near-front	Width-guided	$\mathbf{3.05 \pm 1.99}$	$\mathbf{1.000}$	$1.27 \times 10^5 \pm 8.86 \times 10^4$
Misleading near-front	Pareto UCB1	2584.77 ± 60.31	$\mathbf{1.000}$	$1.67 \times 10^5 \pm 1.10 \times 10^4$
Misleading near-front	Annealing-Pareto	337.59 ± 1504.82	0.983	$1.37 \times 10^5 \pm 2.53 \times 10^4$
Misleading near-front	Scalarized UCB	1435.03 ± 258.58	$\mathbf{1.000}$	$\mathbf{1.07} \times 10^5 \pm \mathbf{2.14} \times 10^4$
Friendly near-front	Width-guided	$\mathbf{0.12 \pm 0.07}$	$\mathbf{1.000}$	$\mathbf{1.40} \times 10^5 \pm \mathbf{9.26} \times 10^4$
Friendly near-front	Pareto UCB1	212.44 ± 61.38	0.967	$2.16 \times 10^5 \pm 2.44 \times 10^4$
Friendly near-front	Annealing-Pareto	26.81 ± 71.75	0.983	$2.05 \times 10^5 \pm 3.62 \times 10^4$
Friendly near-front	Scalarized UCB	65.10 ± 39.01	$\mathbf{1.000}$	$1.55 \times 10^5 \pm 2.62 \times 10^4$
Easy separated	Width-guided	$\mathbf{0.69 \pm 0.43}$	$\mathbf{1.000}$	$2.23 \times 10^5 \pm 5.89 \times 10^4$
Easy separated	Pareto UCB1	1916.61 ± 57.93	$\mathbf{1.000}$	$1.83 \times 10^5 \pm 4.15 \times 10^4$
Easy separated	Annealing-Pareto	408.06 ± 1935.09	0.983	$1.79 \times 10^5 \pm 4.64 \times 10^4$
Easy separated	Scalarized UCB	1198.16 ± 88.62	$\mathbf{1.000}$	$\mathbf{1.25} \times 10^5 \pm \mathbf{2.97} \times 10^4$
KuaiRec d26				
Misleading near-front	Width-guided	$\mathbf{8.00 \pm 4.41}$	$\mathbf{1.000}$	$2.85 \times 10^4 \pm 3.76 \times 10^4$
Misleading near-front	Pareto UCB1	4772.64 ± 1766.07	$\mathbf{1.000}$	$5.03 \times 10^4 \pm 6.82 \times 10^3$
Misleading near-front	Annealing-Pareto	13.36 ± 8.80	$\mathbf{1.000}$	$2.80 \times 10^4 \pm 2.04 \times 10^4$
Misleading near-front	Scalarized UCB	2158.68 ± 585.78	$\mathbf{1.000}$	$\mathbf{1.95} \times 10^4 \pm \mathbf{4.95} \times 10^3$
Friendly near-front	Width-guided	$\mathbf{0.19 \pm 0.14}$	$\mathbf{1.000}$	$\mathbf{1.67} \times 10^4 \pm \mathbf{2.53} \times 10^4$
Friendly near-front	Pareto UCB1	200.54 ± 44.32	0.950	$4.67 \times 10^4 \pm 1.76 \times 10^4$
Friendly near-front	Annealing-Pareto	3.03 ± 6.95	$\mathbf{1.000}$	$4.22 \times 10^4 \pm 1.66 \times 10^4$
Friendly near-front	Scalarized UCB	47.52 ± 7.27	$\mathbf{1.000}$	$2.27 \times 10^4 \pm 4.91 \times 10^3$
Easy separated	Width-guided	$\mathbf{7.53 \pm 3.08}$	$\mathbf{1.000}$	$\mathbf{0.012} \times 10^3 \pm \mathbf{0.005} \times 10^3$
Easy separated	Pareto UCB1	2855.97 ± 455.90	$\mathbf{1.000}$	$6.33 \times 10^3 \pm 2.02 \times 10^3$
Easy separated	Annealing-Pareto	8.66 ± 5.15	$\mathbf{1.000}$	$0.015 \times 10^3 \pm 0.009 \times 10^3$
Easy separated	Scalarized UCB	1789.05 ± 175.84	$\mathbf{1.000}$	$3.31 \times 10^3 \pm 0.538 \times 10^3$

AI USE STATEMENT

Generative AI tools assisted with coding, manuscript, literature and reference checks, and review of mathematical claims. The authors are responsible for the final content of this work.

REFERENCES

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina Drugan. Pareto front identification from stochastic bandit feedback. In Arthur Gretton and Christian C. Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 939–947, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/auer16.html.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013. doi: 10.1109/FOCS.2013.30.

Andrea Barraza-Urbina and Dorota Glowacka. Introduction to bandits in recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pp. 748–750, 2020.

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

Ji Cheng, Bo Xue, Chengyu Lu, Ziqiang Cui, and Qingfu Zhang. Multi-objective neural bandits with random scalarization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4914–4922, 2025.

Ji Cheng, Song Lai, Shunyu Yao, and Bo Xue. Offline multi-objective bandits: From logged data to pareto-optimal policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36636–36644, 2026.

élise crepon, Aurélien Garivier, and Wouter M Koolen. Sequential learning of the Pareto front for multi-objective bandits. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3583–3591. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/crepon24a.html.

Madalina M Drugan and Ann Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2013.

Mădălina M Drugan, Ann Nowé, and Bernard Manderick. Pareto upper confidence bounds algorithms: an empirical study. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pp. 1–8. IEEE, 2014.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. Kuairec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 540–550, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557220. URL https://doi.org/10.1145/3511808.3557220.

Patrick Groetzner and Ralf Werner. Multiobjective optimization under uncertainty: A multiobjective robust (relative) regret approach. *European Journal of Operational Research*, 296(1):101–115, 2022.

Wen-Jing Hong, Peng Yang, and Ke Tang. Evolutionary computation for large-scale multi-objective optimization: A decade of progresses. *International Journal of Automation and Computing*, 18(2): 155–169, 2021.

Tian Huang, Shengbo Wang, and Ke Li. Direct preference-based evolutionary multi-objective optimization with dueling bandits. *Advances in Neural Information Processing Systems*, 37: 122206–122258, 2024.

Alihan Hüyük and Cem Tekin. Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives. *Machine Learning*, 110(6):1233–1266, 2021.

Kevin G. Jamieson and Robert D. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2014. URL https://api.semanticscholar.org/CorpusID:1631520.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 227–234, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

İlter Onat Korkmaz, Yaşar Cahit Yıldırım, Çağın Ararat, and Cem Tekin. Vector optimization with Gaussian process bandits. *Machine Learning*, 115(4), 2026. doi: 10.1007/s10994-026-07024-y. URL https://link.springer.com/article/10.1007/s10994-026-07024-y. Article 75.

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

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. IJCAI'19, pp. 3080–3086. AAAI Press, 2019. ISBN 9780999241141.

Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Rishabh Mehrotra, Niannan Xue, and Mounia Lalmas. Bandit based optimization of multiple objectives on a music streaming platform. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3224–3233, 2020.

Spotify Tracks Dataset. Spotify tracks dataset. https://huggingface.co/datasets/sfiore/spotify-tracks-dataset. Hugging Face dataset card; license listed as `bsd`.

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

Nirandika Wanigasekara, Yuxuan Liang, Siong Thye Goh, Ye Liu, Joseph Jay Williams, and David S Rosenblum. Learning multi-objective rewards and user utility function in contextual bandits for personalized ranking. In *IJCAI*, volume 19, pp. 3835–3841, 2019.

Mengfan Xu and Diego Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023.

Saba Q Yahyaa, Madalina M Drugan, and Bernard Manderick. Annealing-pareto multi-objective multi-armed bandit algorithm. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pp. 1–8. IEEE, 2014.

Tongxin Zhou, Yingfei Wang, Lu Yan, and Yong Tan. Spoiled for choice? personalized recommendation for healthcare decisions: A multiarmed bandit approach. *Information Systems Research*, 34 (4):1493–1512, 2023.

Zheqing Zhu and Benjamin Van Roy. Scalable neural contextual bandit for recommender systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 3636–3646, 2023.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

The appendix is organized as follows.

1. **Appendix A.1: Discussions.** Summarizes the paper and exhibits further discussions.
2. **Appendix A.2: Related Work.** Supports Section 1.
3. **Appendix A.3: Synthetic Experiments.** Supports Section 5.
4. **Appendix A.4: Reproducibility Details.** Provides implementation details for Section 5.
5. **Appendix A.5: Additional Theory.** Provides additional theoretical statements for Section 4.
6. **Appendix A.6: Proofs.** Contains the proofs of the results in Section 4.

A.1 DISCUSSION AND LIMITATIONS

We study stochastic multi-objective bandits under cumulative Pareto regret. Under regularity conditions, we show that Pareto regret can be controlled through the smallest marginal regret among the objectives. Our width-guided method achieves $O(K \log T/g^\dagger)$ Pareto regret for bounded i.i.d. instances with full-vector feedback on the selected arm, provided that some objective has a strictly separated unique winner. We also prove a matching asymptotic lower bound of $\Omega(K \log T/g^\dagger)$ for uniformly good policies on an explicit subclass. The analysis has several limitations that may motivate future work. First, the rate requires $g^\dagger > 0$, and the lower bound is class-level rather than instance-dependent. It also does not formally address other aspects of multi-objective learning, such as fairness and Pareto front recovery. An instance-dependent information lower bound over alternatives to an arbitrary instance is a natural extension in the style of Lai and Robbins (Lai & Robbins, 1985). The current analysis permits arbitrary within-vector dependence but does not exploit covariance. The confidence argument applies coordinatewise across i.i.d. pulls. A shared model for latent means can transfer information across arms while observations remain conditionally independent across arms. Structured models relating latent arm means, including multi-output Gaussian processes (GP), provide a separate future direction. Their confidence construction and rates may involve both a certificate separation and kernel or posterior information complexity, rather than $g^\dagger$ alone. Related work includes Korkmaz et al. (2026), which studies Pareto-set identification under a multi-output GP.

A.2 RELATED WORK

Stochastic Multi-objective Bandits. Multi-objective multi-armed bandits (MO-MAB) generalize the classical MAB framework by replacing the scalar reward with a vector-valued reward, which substantially complicates both algorithm design and performance analysis. A natural starting point is to extend confidence-bound methods from MAB to the multi-objective setting. In this direction, Pareto UCB (Drugan & Nowe, 2013) maintains coordinate-wise upper confidence bounds for each arm and identifies the estimated Pareto front rather than a single best arm; an arm is then selected uniformly at random from this set. Using a multi-dimensional Chernoff–Hoeffding bound, Pareto UCB achieves logarithmic Pareto regret, and subsequent refinements further improve its performance (Drugan et al., 2014). The Pareto-UCB1 bound contains the dimension inside a logarithm and depends on the collection of dominated-arm Pareto gaps. Our result instead gives a leading $O(K \log T/g^\dagger)$ Pareto-regret bound under $g^\dagger > 0$, without an explicit multiplicative d-factor. Other related variants have also been studied for stochastic MO-MAB, including PF-LEX, which incorporates lexicographic order and satisficing objectives (Hüyük & Tekin, 2021), whereas we consider the standard Pareto order relation.

A different stream of related work is rooted in the multi-objective optimization (MOO) literature. A number of algorithms developed in that area have demonstrated strong empirical effectiveness in practical settings, including evolutionary methods for large-scale problems that explicitly navigate the exploration–exploitation trade-off (Hong et al., 2021), as well as Annealing Pareto Knowledge Gradient, which resembles a Thompson-sampling-type strategy (Yahyaa et al., 2014). Despite these promising empirical results, such methods generally do not come with regret guarantees, which limits their applicability in online learning settings where theoretical reliability is important. Another common strategy inspired by MOO is to apply scalarization, for example through linear or Chebyshev

functions, so that the original vector-valued problem is converted into a standard scalar bandit problem. Although this simplification is often computationally convenient, it may obscure important structure in the multi-objective reward and tends to depend heavily on the specific choice of scalarization. Closely related is the line of work that optimizes the General Gini Index (GGI) via online convex optimization (Busa-Fekete et al., 2017); however, the associated regret is measured relative to GGI, rather than with respect to Pareto performance.

Contextual Multi-objective Bandits. For *contextual* multi-objective bandits, the literature has mostly focused on the linear contextual bandit case, where the reward vector is a linear function of contexts or feature vectors. One representative example is MO-LinUCB (Mehrotra et al., 2020), which extends linear contextual bandits to vector-valued rewards. This line of work highlights that, once context is introduced, the challenge is no longer only to learn the Pareto structure across arms, but also to adapt that structure to context-dependent reward models. Recent exceptions include the contextual case in (Turgay et al., 2018), where the linear function is replaced by a similarity distance, meaning that the reward mean is Lipschitz continuous in the context, and (Wanigasekara et al., 2019), where this assumption is removed but the theoretical guarantee is also unknown, as well as neural bandits (Cheng et al., 2025). We focus on the classical stochastic bandit case, while establishing a class-level optimal leading Pareto-regret rate under the stated regularity condition.

Adversarial Multi-objective Bandits. Recently, adversarial multi-objective bandits have attracted increasing attention because they generalize naturally to non-stochastic reward vectors. (Xu & Klabjan, 2023) introduces the first formulation of Pareto regret that directly measures the one-shot distance between the cumulative received reward vector and the Pareto front induced by the cumulative reward vectors of all arms. They show the surprising result that this notion of Pareto regret is smaller than any marginal regret, and then propose a best-of-both-worlds algorithm that works in both stochastic and adversarial settings. Notably, our notion of regret differs from theirs: we consider cumulative distance rather than one-shot distance, so their results do not directly apply here. In addition, our algorithm attains a gap-dependent cumulative-Pareto-regret bound at the easiest marginal scale when $g^{\dagger} > 0$, whereas their best-of-both-worlds guarantee concerns a different cumulative-vector criterion. We emphasize that the subtle connection they uncover between their notion of Pareto regret and marginal regret is what inspires our approach here, although the settings and regret notions differ substantially, as described above.

Multi-objective Optimization. In a broader class of multi-objective decision problems under uncertainty, one common approach is to encode preferences through utility functions, whether linear or nonlinear, as is often done in economics. Conceptually, these methods serve a similar purpose to scalarization by converting vector-valued outcomes into a single evaluative criterion. A more recent direction is the relative minimax regret framework of (Groetzner & Werner, 2022), which comes with theoretical guarantees. Their method evaluates each dimension against the worst-case realization of uncertainty and then optimizes decisions according to Pareto dominance. Despite its theoretical appeal, the framework requires all optimal values under the possible realizations of uncertainty to be computed in advance, making it difficult to adapt to online bandit problems where decisions must be taken sequentially without such offline precomputation. Indeed, there is classical work on using online bandits to solve MOO (Huang et al., 2024). Another relevant line of work is offline multi-objective bandits (Cheng et al., 2026), which is designed for the offline setting, whereas our focus is online, highlighting the main difference.

Bandits with knapsacks. Both settings learn unknown vector-valued means through confidence-based allocation. In bandits with knapsacks (Badanidiyuru et al., 2013), a scalar payoff is accompanied by resource consumptions subject to cumulative budgets, so pulls change future feasibility and optimal play may mix arms. Here all coordinates are reward criteria and no budget couples rounds. A zero-gap arm incurs no further Pareto regret; on the confidence event, our method sustains this after one strict-winner certificate. The sources of difficulty are therefore cumulative feasibility in the knapsack setting and statistical certification under noise in ours.

A.3 Numerical Results on Synthetic Datasets

We also include a controlled synthetic family to isolate the mechanism behind the comparison with Pareto UCB1. This experiment uses the same horizon and Monte Carlo budget as the real-data benchmark: $T = 10^6$ and 20 repetitions for every setting. Unlike the real-data Width-guided

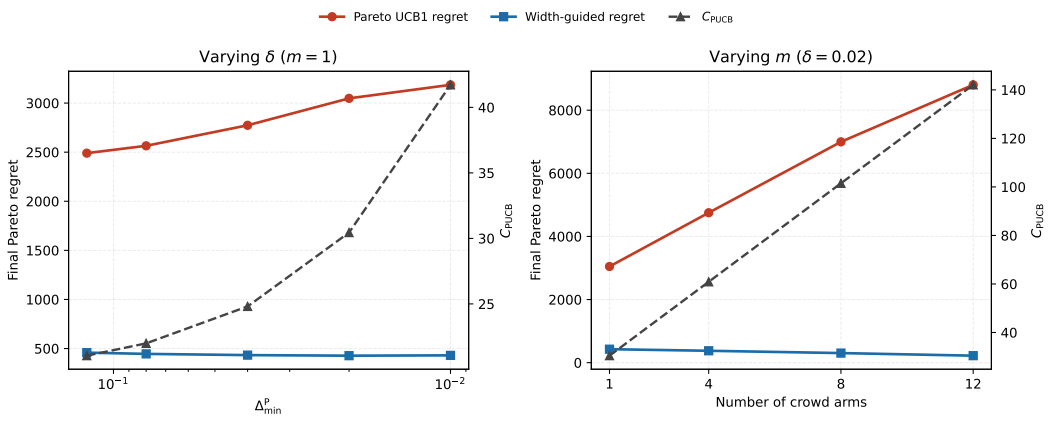


Figure 1: Synthetic controlled family at $T = 10^6$ with 20 repetitions. The first panel decreases the Pareto gap δ with one near-front dominated arm; the second increases the number m of near-front dominated arms at fixed $\delta = 0.02$. The first-certification scale $g^\dagger = 0.55$ is fixed throughout.

Table 2: Synthetic family with fixed certification gap and small Pareto gaps at $T = 10^6$; $g^\dagger = 0.55$ throughout and $\delta = \Delta^{\mathrm{P}}_{\min}$. Regret entries report mean $\pm$ standard deviation over 20 runs.

δ	m	C_{PUCB}	Pareto UCB1	Width-guided	Cert. rate
Panel A: vary δ, $m = 1$					
0.120	1	21.05	2490.0 ± 46.8	458.5 ± 27.2	100%
0.080	1	21.99	2564.8 ± 70.3	445.6 ± 28.7	100%
0.040	1	24.81	2773.7 ± 97.3	433.2 ± 26.3	100%
0.020	1	30.45	3048.1 ± 104.4	427.4 ± 23.1	100%
0.010	1	41.72	3185.8 ± 146.7	430.6 ± 30.9	100%
Panel B: vary m, $\delta = 0.020$					
0.020	4	60.89	4751.0 ± 236.1	377.0 ± 19.8	100%
0.020	8	101.48	6997.2 ± 395.6	304.0 ± 12.9	100%
0.020	12	142.08	8803.7 ± 343.8	221.7 ± 10.7	100%

implementation in Table 1, the synthetic experiment uses the theorem-matched confidence radius $\sqrt{2 \log T / N_a(t-1)}$, with no empirical coefficient tuning.

Each synthetic instance has $K = 20$ arms and $d = 2$ objectives. Two arms form the Pareto front, with means $(p + g, p)$ and $(p, p + g)$, where $p = 0.25$ and $g = 0.55$. A crowd of m dominated arms has mean $(p - \eta, p + g - \delta)$, with $\eta = 0.20$, and the remaining arms are low-reward fillers. Thus the first-certification scale remains fixed at $g^\dagger = 0.55$, while the Pareto-UCB burden grows either by decreasing the dominated-arm Pareto gap $\delta = \Delta^{\mathrm{P}}_{\min}$ or by increasing the number m of near-front dominated arms.

Remark 8 (Synthetic view of the C_{PUCB} comparison). The coefficient C_{PUCB} is the normalized leading constant appearing in the comparison with the Pareto-UCB1 gap-dependent bound. The sufficient comparison condition $64 < C_{\mathrm{PUCB}}(T, \mu)$ from Section 4 is intentionally conservative, not a necessary condition for finite-horizon improvement. In Table 2, Width-guided already has smaller regret in the smaller-C_{PUCB} settings, and the advantage becomes more pronounced as C_{PUCB} increases through smaller δ or larger m. This is the synthetic counterpart of the real-data pattern: when many positive-gap arms remain attractive near the empirical front, first certification can avoid repeated positive-gap pulls.

A.4 Additional Computational Details

This appendix records reproducibility details for the computational study in Section 5. It specifies the computational resources used, the real-data benchmark construction, validation/test split, policy definitions, selected Width-guided coefficient, and front-identification diagnostics. It also records the matched-budget synthetic-family setup. The main real-data table reports outcome metrics only;

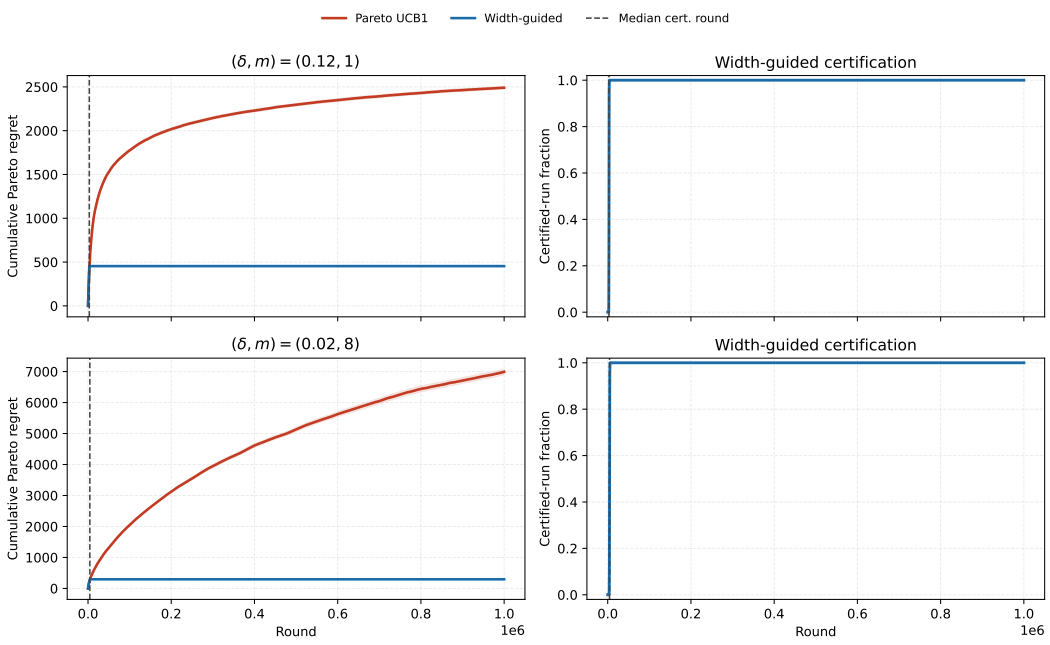


Figure 2: Synthetic regret and certification trajectories for two representative settings. Width-guided certifies in every run and then stops accumulating Pareto regret, while Pareto UCB1 continues to sample from a broader optimistic front.

the diagnostics here are included to clarify what the first-certification objective does and does not optimize.

Computer resources. The final evaluation scripts in the open-source artifact were rerun as CPU-only Slurm array jobs; no GPUs were requested or used. For every final real-data and synthetic evaluation, we used 20 array shards, and each shard requested 20 CPU cores, 40 GB of memory, and a conservative 12-hour wall-clock limit. The rerun outputs are included in the open-source artifact. In this rerun, all completed final-evaluation shards finished within 6 minutes of wall-clock time, excluding scheduler queueing. The software environment used Python 3.9.21, NumPy 2.0.2, and Matplotlib 3.9.4.

Dataset construction. The held-out benchmark uses two prepared real-data instances. In Spotify, the 114 arms are genres, $d = 6$, and rewards are empirical track feature vectors sampled with replacement from the selected genre. In KuaiRec, the 300 arms are videos, $d = 26$, and rewards are cohort-specific normalized watch-ratio samples. All reward coordinates are scaled to $[0, 1]$, and each benchmark subset contains $K = 10$ arms. The Spotify objectives are popularity, danceability, energy, loudness, acousticness, and valence. The KuaiRec objectives are the 26 largest retained user cohorts formed from user activity, follow-count range, and registration-age range. Spotify subsets use the geometry selector, while KuaiRec uses the contamination-balanced geometry selector, which applies the same misleading/friendly/easy labels but prioritizes candidates where bootstrapped empirical fronts contain positive-gap near-front arms.

On the frozen Spotify subsets, $g^\dagger$ is $0.375856, 0.107449, 0.112970$ (misleading), $0.173673, 0.101168, 0.138873$ (friendly), and $0.257544, 0.310974, 0.328786$ (easy). The corresponding KuaiRec values are $0.154020, 0.118469, 0.128796$ (misleading), $0.060137, 0.125906, 0.148372$ (friendly), and $0.367829, 0.352207, 0.348669$ (easy).

Validation, coefficient selection, and held-out evaluation. For Width-guided, the only hyperparameter selected on our validation subsets is the confidence coefficient c in

$$\beta_a(t) = \sqrt{\frac{c \log T}{N_a(t-1)}}.$$

The analyzed theorem corresponds to the conservative choice $c = 2$. For the real-data benchmark, c is selected on validation subsets at the same horizon $T = 10^6$, and the selected value is frozen before held-out testing. This empirical implementation should not be read as carrying the high-probability guarantee of Theorem 5. Baseline hyperparameters are taken from the recommended settings in their original papers rather than reselected on our validation subsets.

Table 3: Validation/test split and selected Width-guided coefficient. "Subsets $\times$ seeds" is per dataset and subset type. All held-out rows use the same test seed; Spotify easy uses a separately frozen geometry-selected easy suite.

Dataset	Validation seed	Test seed	Validation	Test	Selected c
Spotify d6, misleading/friendly	20260604	20260704	3×3	3×20	0.02
Spotify d6, easy	20260604	20260704	3×3	3×20	0.00894
KuaiRec d26, all types	20260604	20260704	3×3	3×20	0.02

The held-out table therefore averages 3 subsets and 20 reward seeds per subset type, for 60 runs per method and type. Candidate subset generation uses frozen subset files selected before the policy simulations; the Spotify easy row uses a separately frozen geometry-selected easy suite. The validation and test seeds are different, and the same held-out test seed is used across rows, so no held-out reward stream is used when choosing c.

Policy implementations. Within a Monte Carlo run, all policies use the same reward stream for each arm. Width-guided is Algorithm 1 with the selected coefficient above; after an objective-wise certificate appears, it commits to the certified arm. Pareto UCB1 pulls each arm once, forms coordinate-wise upper confidence vectors, and samples uniformly from the optimistic empirical Pareto set. Annealing-Pareto uses the reported empirical decay setting from its original experiments and otherwise samples uniformly from the empirical Pareto set. Scalarized UCB applies ordinary UCB over a fixed multi-weight scalarization set.

Front-identification and coverage diagnostics. Table 4 reports diagnostics on the same held-out runs. For the fully sampled baselines, precision and recall compare the Pareto front of final empirical means to the benchmark front. Width-guided instead accounts for its committed tail using exact benchmark means, so its precision and recall are hybrid plug-in diagnostics and should not be compared across methods. This shortcut leaves the actions, Pareto regret, detection, and count-based coverage entropy unchanged. Coverage entropy is the normalized entropy of total pull counts over the true Pareto arms; it measures diversity of play, not identification correctness or sample complexity.

The diagnostics show the separation between objectives. Pareto UCB1 and Annealing-Pareto often have greater count-based coverage entropy because they continue sampling across the front. Width-guided concentrates on one certified Pareto-optimal arm, which is why its coverage entropy is close to zero. The KuaiRec easy subsets have singleton true fronts, so all methods have zero entropy there. These comparisons concern diversity of play and do not establish comparative full-front-recovery performance.

Synthetic-family reproducibility. The synthetic family in Figure 1, Table 2, and Figure 2 uses the same horizon and repetition count as the real-data benchmark: $T = 10^6$, 20 final-regret repetitions per setting, and 20 trajectory repetitions per plotted setting. Within each setting, all policies are evaluated on the same synthetic instance family and matched Monte Carlo repetitions. Unlike the real-data benchmark, the synthetic Width-guided runs use the theorem-matched radius $\sqrt{2 \log T / N_a(t-1)}$, with no validation-tuned coefficient.

Table 4: Front-identification and coverage diagnostics on the held-out real-data subsets. Width-guided rows marked † use hybrid plug-in precision and recall, which are not comparable with the fully sampled baselines. Coverage entropy remains comparable across methods.

Dataset	Subset type	Method	Front precision ↑	Front recall ↑	Coverage entropy ↑
Spotify d6	Misleading	Width-guided†	0.760	0.947	0.001
Spotify d6	Misleading	Pareto UCB1	1.000	1.000	1.000
Spotify d6	Misleading	Annealing-Pareto	0.782	0.840	0.848
Spotify d6	Misleading	Scalarized UCB	1.000	1.000	0.812
Spotify d6	Friendly	Width-guided†	0.850	0.921	0.002
Spotify d6	Friendly	Pareto UCB1	0.971	1.000	1.000
Spotify d6	Friendly	Annealing-Pareto	0.879	0.835	0.820
Spotify d6	Friendly	Scalarized UCB	0.962	1.000	0.775
Spotify d6	Easy	Width-guided†	0.713	0.919	0.000
Spotify d6	Easy	Pareto UCB1	1.000	0.972	1.000
Spotify d6	Easy	Annealing-Pareto	0.852	0.881	0.838
Spotify d6	Easy	Scalarized UCB	1.000	0.964	0.733
KuaiRec d26	Misleading	Width-guided†	0.416	0.993	0.001
KuaiRec d26	Misleading	Pareto UCB1	1.000	0.990	0.667
KuaiRec d26	Misleading	Annealing-Pareto	0.925	0.940	0.620
KuaiRec d26	Misleading	Scalarized UCB	1.000	0.987	0.447
KuaiRec d26	Friendly	Width-guided†	0.815	0.950	0.003
KuaiRec d26	Friendly	Pareto UCB1	0.963	1.000	1.000
KuaiRec d26	Friendly	Annealing-Pareto	0.910	0.912	0.934
KuaiRec d26	Friendly	Scalarized UCB	0.961	1.000	0.737
KuaiRec d26	Easy	Width-guided†	0.502	1.000	0.000
KuaiRec d26	Easy	Pareto UCB1	1.000	1.000	0.000
KuaiRec d26	Easy	Annealing-Pareto	1.000	1.000	0.000
KuaiRec d26	Easy	Scalarized UCB	1.000	1.000	0.000

A.5 Additional Theoretical Results

Because the confidence radius depends only on the realized sample count, the natural concentration event is a single horizon-T event that covers every arm, every objective, and every realized sample count up to time T.

Proposition 9 (Global confidence event)**.** *Let*

$$\Omega_T := \left\{ \left| \widehat{\mu}_a^{(j)}(t-1) - \mu_a^{(j)} \right| \le \beta_a(t) \quad \forall K < t \le T,\ a \in [K],\ j \in [d] \right\}.$$

Then

$$\Pr(\Omega_T^c) \le 2KdT^{-3}.$$

Proposition 9 resolves the concentration step. Its role is to reduce the adaptive sampling process to a single event Ω_T on which every empirical mean lies inside its current confidence interval, simultaneously across all arms, objectives, and rounds. Once we condition on Ω_T, the remaining work is deterministic and algorithmic: we must show that certification returns a Pareto-optimal arm, that each pull at a non-certifying round incurs only uncertainty-scale Pareto regret, and that at every non-certifying round the champion objective still has width of order $g^\dagger$.

Lemma 10 (Certification identifies a Pareto-optimal arm)**.** *Fix a round $K < t \le T$ and an objective j. On the event Ω_T, if*

$$L_{b_j(t)}^{(j)}(t) > U_{c_j(t)}^{(j)}(t),$$

then $b_j(t)$ is the unique maximizer of objective j, which implies that $b_j(t) \in A^\star$ and hence $\Delta^{\mathrm{P}}_{b_j(t)} = 0$.

Lemma 10 is the point where a one-coordinate statistical statement becomes a Pareto regret statement. The algorithm does not certify the full Pareto front; it certifies that one arm is the winner of one objective, and therefore lies in $A^\star$. Once certification occurs, the stored leader can be exploited forever because every Pareto-optimal arm has zero Pareto gap.

The remaining main task is to control the regret accumulated before certification. On the confidence event, all positive Pareto regret after the warm start comes from non-certifying rounds, that is, rounds at which no certified leader has yet been stored and no objective certifies. The key bridge is that the Pareto gap is upper-bounded by the deficit to the best arm on any single objective. Thus, once the selected objective is fixed, controlling an endpoint's objective-wise deficit controls its Pareto regret, even though the benchmark itself is not scalarized. At a non-certifying round, the algorithm selects the objective with the largest pair width and samples the more uncertain endpoint of its top-two UCB pair. The next lemma shows that every such pull has Pareto regret of only confidence-radius scale.

Lemma 11 (Non-certifying pulls have uncertainty-scale regret)**.** *Fix a round $K < t \le T$ on Ω_T. Suppose round t is non-certifying. Let*

$$j_t \in \arg\max_{j\in[d]} W_j(t), \qquad A_t \in \arg\max_{a\in\{b_{j_t}(t),\, c_{j_t}(t)\}} \beta_a(t).$$

Then

$$\Delta^{\mathrm{P}}_{A_t} \le 4\beta_{A_t}(t).$$

Lemma 11 is the local regret-to-uncertainty bridge. At a non-certifying round, the sampled arm is one endpoint of an unresolved top-two UCB pair. If that arm had a large Pareto gap, then on the selected objective it would lie too far below the true winner; on Ω_T, that would either force the true winner to outrank it in the UCB ordering or make the pair certify already. Hence the chosen arm can remain active only because uncertainty still masks its deficit, and its instantaneous Pareto regret is controlled by its current confidence radius.

By itself, Lemma 11 does not yet imply a cumulative regret bound. A complementary ingredient is needed: before certification occurs, the champion objective must still have unresolved width of order $g^\dagger$, and because the algorithm selects the widest pair, every sampled arm at a non-certifying round must still have radius at least of order $g^\dagger$. This is where the scheduling rule matters. The learner does not know in advance which objective will provide the first certificate, but as long as no certificate exists, the champion objective remains unresolved at scale $g^\dagger$. Choosing the widest race transfers this lower bound to the race actually sampled.

Lemma 12 (Champion width floor at non-certifying rounds)**.** *Fix a round $K < t \le T$ on Ω_T and suppose $g^\dagger > 0$. Suppose round t is non-certifying. Then the champion objective satisfies*

$$\beta_{b_{j^\dagger(t)}}(t) + \beta_{c_{j^\dagger(t)}}(t) \ge \frac{g^\dagger}{2}.$$

Consequently, if

$$j_t \in \arg\max_{j\in[d]} W_j(t), \qquad A_t \in \arg\max_{a\in\{b_{j_t}(t),\, c_{j_t}(t)\}} \beta_a(t),$$

with

$$b_t := b_{j_t}(t), \qquad c_t := c_{j_t}(t),$$

then

$$\beta_{b_t}(t) + \beta_{c_t}(t) \ge \frac{g^\dagger}{2}, \qquad \beta_{A_t}(t) \ge \frac{g^\dagger}{4}.$$

Lemma 12 supplies the global complement to Lemma 11. Lemma 11 says that at a non-certifying round, regret can be charged to the current confidence radius of the sampled arm. Lemma 12 says that as long as certification has not yet happened, those radii cannot collapse to an arbitrarily small scale: the champion objective still forces unresolved width of order $g^\dagger$, so every sampled arm must still have radius at least of order $g^\dagger$.

The main theorem then follows by combining these two facts: each pre-certification pull costs only radius-scale regret, each arm can incur only finitely many such pulls before its radius drops below the level allowed by Lemma 12, and after certification the regret becomes identically zero.

A.6 PROOFS

Let

$$\mathcal{F}_t := \sigma(A_1, X_1, \ldots, A_t, X_t)$$

be the natural filtration. The policy is non-anticipatory: A_{t+1} is $\mathcal{F}_t$-measurable. Under the latent-sequence construction, if $N_a(t-1) \geq 1$, then the empirical mean used by the algorithm is exactly

$$\widehat{\mu}_a^{(j)}(t-1) := \frac{1}{N_a(t-1)} \sum_{s=1}^{t-1} X_s^{(j)} \mathbf{1}\{A_s = a\} = \frac{1}{N_a(t-1)} \sum_{n=1}^{N_a(t-1)} Y_{a,n}^{(j)}.$$

A.6.1 PROOF OF PROPOSITION 1

Proof. Let $\mu_1 = (1,1)$ and $\mu_2 = (1,0)$. Arm 1 dominates arm 2, so $2 \notin A^\star$. However, for every $\epsilon > 0$,

$$\mu_2 + \epsilon \mathbf{1} = (1+\epsilon, \epsilon)$$

is not dominated by μ_1, because its first coordinate is larger than that of μ_1. Hence $\Delta_2^{\mathrm{P}} = 0$. □

A.6.2 PROOF OF COROLLARY 7

Proof. On Ω_T, Lemma 10 shows that whenever the certification condition

$$L_{b_j(t)}^{(j)}(t) > U_{c_j(t)}^{(j)}(t)$$

holds, the certified leader $b_j(t)$ is the unique maximizer of objective j. Hence every other arm is strictly worse than $b_j(t)$ on coordinate j, so no other arm can Pareto dominate it. Therefore $b_j(t) \in A^\star$. Since Algorithm 1 stores only arms accepted through this certificate condition, a stored non-Pareto-optimal certified leader can occur only on Ω_T^c. Proposition 9 gives $\Pr(\Omega_T^c) \leq 2KdT^{-3}$, proving the claim. □

A.6.3 PROOF OF PROPOSITION 9

Proof. Fix an arm a, an objective j, and a count $n \in \{1, \ldots, T\}$. Let

$$\bar{Y}_{a,n}^{(j)} := \frac{1}{n} \sum_{m=1}^{n} Y_{a,m}^{(j)}.$$

Because the latent samples

$$Y_{a,1}^{(j)}, Y_{a,2}^{(j)}, \ldots$$

are i.i.d. with mean $\mu_a^{(j)}$, the average $\bar{Y}_{a,n}^{(j)}$ is the mean of n i.i.d. $[0,1]$-valued random variables. By Hoeffding's inequality,

$$\Pr\Big(\Big|\bar{Y}_{a,n}^{(j)} - \mu_a^{(j)}\Big| > \sqrt{\frac{2\log T}{n}}\Big) \leq 2\exp(-4\log T) = 2T^{-4}.$$

Define

$$E_{a,j,n} := \left\{ \Big|\bar{Y}_{a,n}^{(j)} - \mu_a^{(j)}\Big| > \sqrt{\frac{2\log T}{n}} \right\}.$$

Then $\Pr(E_{a,j,n}) \leq 2T^{-4}$.

Now fix any round $K < t \leq T$. By the latent-sequence construction, the empirical mean of arm a and objective j after $t-1$ rounds is the average of the first $N_a(t-1)$ latent samples from that arm, namely

$$\widehat{\mu}_a^{(j)}(t-1) = \bar{Y}_{a,N_a(t-1)}^{(j)}.$$

Hence if Ω_T fails, then for some (t,a,j) with $K < t \leq T$,

$$\Big|\widehat{\mu}_a^{(j)}(t-1) - \mu_a^{(j)}\Big| > \sqrt{\frac{2\log T}{N_a(t-1)}}.$$

Setting $n := N_a(t-1)$, the event $E_{a,j,n}$ occurs. Therefore

$$\Omega_T^c \subseteq \bigcup_{a\in[K]} \bigcup_{j\in[d]} \bigcup_{n=1}^{T} E_{a,j,n}.$$

Applying the union bound yields

$$\Pr(\Omega_T^c) \le 2KdT \cdot T^{-4} = 2Kd\,T^{-3}.$$

□

A.6.4 Proof of Lemma 10

Proof. Since $b_j(t)$ and $c_j(t)$ are the two largest-UCB arms on objective j, every arm $a \neq b_j(t)$ satisfies

$$U_a^{(j)}(t) \le U_{c_j(t)}^{(j)}(t) < L_{b_j(t)}^{(j)}(t) \le \mu_{b_j(t)}^{(j)}.$$

Also, on Ω_T,

$$\mu_a^{(j)} \le U_a^{(j)}(t).$$

Hence

$$\mu_a^{(j)} < \mu_{b_j(t)}^{(j)} \qquad \forall a \neq b_j(t),$$

so $b_j(t)$ is the unique objective maximizer on objective j. No other arm can dominate it, because every other arm is strictly worse on coordinate j. Therefore $b_j(t) \in A^\star$, and hence $\Delta^{\mathrm{P}}_{b_j(t)} = 0$. □

A.6.5 Proof of Lemma 11

Proof. Let j_t be the selected objective and write

$$b_t := b_{j_t}(t), \qquad c_t := c_{j_t}(t).$$

For any arm a and objective j, define

$$m_j := \max_{u\in[K]} \mu_u^{(j)}, \qquad \delta_j(a) := m_j - \mu_a^{(j)} \ge 0.$$

We first claim that for every arm a and objective j,

$$\Delta_a^{\mathrm{P}} \le \delta_j(a).$$

Fix any $\eta > \delta_j(a)$. Then

$$\mu_a^{(j)} + \eta > m_j \ge \mu_b^{(j)} \qquad \forall b \in [K],$$

so the lifted vector $\mu_a + \eta\mathbf{1}$ is strictly largest in coordinate j and therefore cannot be dominated by any arm. By the definition of the Pareto suboptimality gap, $\Delta_a^{\mathrm{P}} \le \eta$ for every $\eta > \delta_j(a)$, and thus $\Delta_a^{\mathrm{P}} \le \delta_j(a)$.

Applying this with $j = j_t$, it suffices to upper-bound

$$\delta_{j_t}(A_t) = m_{j_t} - \mu_{A_t}^{(j_t)}.$$

Because round t is non-certifying, $A_t \in \{b_t, c_t\}$.

Case 1: $A_t = b_t$. If $\mu_{b_t}^{(j_t)} = m_{j_t}$, then A_t maximizes objective j_t and $\Delta_{A_t}^{\mathrm{P}} = 0$. Otherwise, choose any $a^+ \in \arg\max_{a\in[K]} \mu_a^{(j_t)}$. Since b_t has the largest UCB on objective j_t,

$$U_{b_t}^{(j_t)}(t) \ge U_{a^+}^{(j_t)}(t) \ge m_{j_t},$$

while on Ω_T,

$$U_{b_t}^{(j_t)}(t) \le \mu_{b_t}^{(j_t)} + 2\beta_{b_t}(t).$$

Hence

$$m_{j_t} - \mu_{b_t}^{(j_t)} \le 2\beta_{b_t}(t),$$

so $\delta_{j_t}(A_t) \le 2\beta_{A_t}(t)$ and therefore

$$\Delta_{A_t}^{\mathrm{P}} \le 2\beta_{A_t}(t).$$

Case 2: $A_t = c_t$**.** If $\mu_{b_t}^{(j_t)} = m_{j_t}$, then because round t is non-certifying,

$$L_{b_t}^{(j_t)}(t) \le U_{c_t}^{(j_t)}(t).$$

On Ω_T,

$$L_{b_t}^{(j_t)}(t) \ge m_{j_t} - 2\beta_{b_t}(t), \qquad U_{c_t}^{(j_t)}(t) \le \mu_{c_t}^{(j_t)} + 2\beta_{c_t}(t),$$

which gives

$$m_{j_t} - \mu_{c_t}^{(j_t)} \le 2\beta_{b_t}(t) + 2\beta_{c_t}(t).$$

Since $A_t = c_t$ is the more uncertain endpoint,

$$\beta_{b_t}(t) \le \beta_{c_t}(t) = \beta_{A_t}(t),$$

so $\delta_{j_t}(A_t) \le 4\beta_{A_t}(t)$ and hence

$$\Delta_{A_t}^{\mathrm{P}} \le 4\beta_{A_t}(t).$$

If instead $\mu_{b_t}^{(j_t)} < m_{j_t}$, then some objective maximizer $a^+ \in \arg\max_{a\in[K]} \mu_a^{(j_t)}$ belongs to the candidate set for c_t, so

$$U_{c_t}^{(j_t)}(t) \ge U_{a^+}^{(j_t)}(t) \ge m_{j_t}.$$

On Ω_T,

$$U_{c_t}^{(j_t)}(t) \le \mu_{c_t}^{(j_t)} + 2\beta_{c_t}(t),$$

which yields

$$m_{j_t} - \mu_{c_t}^{(j_t)} \le 2\beta_{c_t}(t) = 2\beta_{A_t}(t).$$

Thus $\Delta_{A_t}^{\mathrm{P}} \le 2\beta_{A_t}(t) \le 4\beta_{A_t}(t)$. □

A.6.6 Proof of Lemma 12

Proof. Since $g^\dagger > 0$, objective $j^\dagger$ has a separated winner, denoted by $a_{j^\dagger}^\star$. Let

$$b^\dagger := b_{j^\dagger}(t), \qquad c^\dagger := c_{j^\dagger}(t).$$

Case 1: $b^\dagger = a_{j^\dagger}^\star$**.** Because round t is non-certifying, objective $j^\dagger$ does not certify, so

$$L_{b^\dagger}^{(j^\dagger)}(t) \le U_{c^\dagger}^{(j^\dagger)}(t).$$

On Ω_T,

$$L_{b^\dagger}^{(j^\dagger)}(t) \ge \mu_{a_{j^\dagger}^\star}^{(j^\dagger)} - 2\beta_{b^\dagger}(t).$$

Since $c^\dagger \ne a_{j^\dagger}^\star$ and $g^\dagger = g_{j^\dagger}$, the definition of $g_{j^\dagger}$ gives

$$\mu_{c^\dagger}^{(j^\dagger)} \le \mu_{a_{j^\dagger}^\star}^{(j^\dagger)} - g^\dagger.$$

Therefore, on Ω_T,

$$U_{c^\dagger}^{(j^\dagger)}(t) \le \mu_{a_{j^\dagger}^\star}^{(j^\dagger)} - g^\dagger + 2\beta_{c^\dagger}(t).$$

Combining the displays gives

$$2\beta_{b^\dagger}(t) + 2\beta_{c^\dagger}(t) \ge g^\dagger,$$

and hence

$$\beta_{b^\dagger}(t) + \beta_{c^\dagger}(t) \ge \frac{g^\dagger}{2}.$$

Case 2: $b^\dagger \neq a^\star_{j^\dagger}$. By definition of $b^\dagger$,

$$U^{(j^\dagger)}_{b^\dagger}(t) \geq U^{(j^\dagger)}_{a^\star_{j^\dagger}}(t) \geq \mu^{(j^\dagger)}_{a^\star_{j^\dagger}}.$$

On Ω_T,

$$U^{(j^\dagger)}_{b^\dagger}(t) \leq \mu^{(j^\dagger)}_{b^\dagger} + 2\beta_{b^\dagger}(t) \leq \mu^{(j^\dagger)}_{a^\star_{j^\dagger}} - g^\dagger + 2\beta_{b^\dagger}(t),$$

which implies

$$\beta_{b^\dagger}(t) \geq \frac{g^\dagger}{2}.$$

Thus

$$\beta_{b^\dagger}(t) + \beta_{c^\dagger}(t) \geq \frac{g^\dagger}{2}.$$

In both cases,

$$\beta_{b^\dagger}(t) + \beta_{c^\dagger}(t) \geq \frac{g^\dagger}{2}.$$

Since round t is non-certifying, the algorithm picks

$$j_t \in \arg\max_{j\in[d]} W_j(t),$$

so

$$W_{j_t}(t) \geq W_{j^\dagger}(t).$$

Recalling that $W_j(t) = \beta_{b_j(t)}(t) + \beta_{c_j(t)}(t)$, we obtain

$$\beta_{b_t}(t) + \beta_{c_t}(t) \geq \frac{g^\dagger}{2}.$$

Finally,

$$\beta_{A_t}(t) = \max\{\beta_{b_t}(t), \beta_{c_t}(t)\} \geq \frac{\beta_{b_t}(t) + \beta_{c_t}(t)}{2} \geq \frac{g^\dagger}{4}.$$

□

A.6.7 Proof of Theorem 5

Proof. Let τ_{cert} be the first round after the warm start at which some objective certifies, with the convention $\tau_{\text{cert}} := T + 1$ if no certification occurs by time T.

Good-event decomposition. The warm start contributes at most

$$K\Delta^{\text{P}}_{\max}.$$

On Ω_T, Lemma 10 shows that the first certified leader has zero Pareto gap. Since the algorithm stores that leader and exploits it forever,

$$\Delta^{\text{P}}_{A_t} = 0 \qquad \text{for every } t \geq \tau_{\text{cert}} \quad \text{on } \Omega_T.$$

Thus only rounds $K < t < \tau_{\text{cert}}$ contribute further regret on Ω_T.

Length of the pre-certification phase. Fix any round $K < t < \tau_{\text{cert}}$ with $A_t = a$. Since round t is non-certifying, Lemma 12 gives

$$\beta_{A_t}(t) \geq \frac{g^\dagger}{4}.$$

Hence

$$\sqrt{\frac{2\log T}{N_a(t-1)}} \geq \frac{g^\dagger}{4},$$

which implies

$$N_a(t-1) \leq \frac{32\log T}{(g^\dagger)^2}.$$

Now let m_a be the number of times arm a is pulled after the warm start and before certification. The successive pre-pull counts of those pulls are exactly

$$1, 2, \dots, m_a.$$

Since each must be at most $32 \log T/(g^\dagger)^2$, we obtain the direct bound

$$m_a \le \frac{32 \log T}{(g^\dagger)^2}.$$

Regret accumulated before certification. For each arm a, let

$$t_{a,1} < t_{a,2} < \cdots < t_{a,m_a}$$

be the rounds at which arm a is pulled after the warm start and before certification. Then

$$\sum_{t=K+1}^{\tau_{\text{cert}}-1} \Delta^{\text{P}}_{A_t} = \sum_{a=1}^{K} \sum_{s=1}^{m_a} \Delta^{\text{P}}_{A_{t_{a,s}}}.$$

Since each $t_{a,s}$ is non-certifying, Lemma 11 gives

$$\Delta^{\text{P}}_{A_{t_{a,s}}} \le 4\beta_{A_{t_{a,s}}}(t_{a,s}) = 4\beta_a(t_{a,s}).$$

Because $t_{a,s}$ is the sth post-warm-start pull of arm a,

$$N_a(t_{a,s} - 1) = s, \qquad \beta_a(t_{a,s}) = \sqrt{\frac{2 \log T}{s}}.$$

Therefore

$$\sum_{t=K+1}^{\tau_{\text{cert}}-1} \Delta^{\text{P}}_{A_t} \le 4 \sum_{a=1}^{K} \sum_{s=1}^{m_a} \sqrt{\frac{2 \log T}{s}}.$$

Using $\sum_{s=1}^{m_a} s^{-1/2} \le 2\sqrt{m_a}$ and the bound on m_a gives

$$\begin{aligned}
\sum_{a=1}^{K} \sum_{s=1}^{m_a} \sqrt{\frac{2 \log T}{s}} &\le \sum_{a=1}^{K} \sqrt{2 \log T} \sum_{s=1}^{m_a} s^{-1/2} \\
&\le 2\sqrt{2 \log T} \sum_{a=1}^{K} \sqrt{m_a} \\
&\le 2K \sqrt{2 \cdot \frac{32 \log T}{(g^\dagger)^2} \cdot \log T} \le 16 \frac{K \log T}{g^\dagger}.
\end{aligned}$$

Hence

$$\sum_{t=K+1}^{\tau_{\text{cert}}-1} \Delta^{\text{P}}_{A_t} \le 64 \frac{K \log T}{g^\dagger}.$$

Combining this estimate with zero regret after certification yields

$$R^{\text{P}}_T \mathbf{1}\{\Omega_T\} \le K \Delta^{\text{P}}_{\max} + 64 \frac{K \log T}{g^\dagger}.$$

Bad-event contribution. Trivially,

$$R^{\text{P}}_T \le T \Delta^{\text{P}}_{\max}.$$

Therefore Proposition 9 yields

$$\mathbb{E}\big[R^{\text{P}}_T \mathbf{1}\{\Omega^c_T\}\big] \le T \Delta^{\text{P}}_{\max} \Pr(\Omega^c_T) \le 2Kd\, \Delta^{\text{P}}_{\max} T^{-2}.$$

Adding the bad-event term proves the theorem. □

A.6.8 Comparison with Pareto UCB1

Theorem 1 of (Drugan & Nowe, 2013) states that the Pareto UCB1 policy satisfies

$$\mathbb{E}\big[R_T^{\mathrm{P}}\big] \le \sum_{i \notin A^\star} \frac{8\,\log(T(d|A^\star|)^{1/4})}{\Delta_i^{\mathrm{P}}} \;+\; (1+\frac{\pi^2}{3}) \sum_{i \notin A^\star} \Delta_i^{\mathrm{P}}.$$

This comparison is most transparent in the generic non-boundary case where every arm outside the Pareto set has positive Pareto gap. If boundary dominated arms with zero Pareto gap exist, they are regret-neutral under the criterion and the same comparison should be read over the positive-gap arms. If $A^\star = [K]$, then the leading term is the empty sum and there is nothing further to compare. We therefore focus on the nontrivial generic case $A^\star \neq [K]$ and define

$$\Delta_{\min}^{\mathrm{P}} := \min_{i \notin A^\star} \Delta_i^{\mathrm{P}}.$$

To compare leading terms, factor out the common logarithmic term:

$$\sum_{i \notin A^\star} \frac{8\,\log(T(d|A^\star|)^{1/4})}{\Delta_i^{\mathrm{P}}} = 8\,\log(T(d|A^\star|)^{1/4}) \sum_{i \notin A^\star} \frac{1}{\Delta_i^{\mathrm{P}}}.$$

Since every dominated arm satisfies $\Delta_i^{\mathrm{P}} \ge \Delta_{\min}^{\mathrm{P}}$, we have

$$\sum_{i \notin A^\star} \frac{1}{\Delta_i^{\mathrm{P}}} \le \frac{K - |A^\star|}{\Delta_{\min}^{\mathrm{P}}}.$$

Therefore the leading term is at most

$$\frac{8(K - |A^\star|)\,\log(T(d|A^\star|)^{1/4})}{\Delta_{\min}^{\mathrm{P}}}.$$

Writing this on the same $K \log T / g^\dagger$ scale as Theorem 5 gives

$$\frac{8(K - |A^\star|)\,\log(T(d|A^\star|)^{1/4})}{\Delta_{\min}^{\mathrm{P}}} = (8 \cdot \frac{K - |A^\star|}{K} \cdot \frac{g^\dagger}{\Delta_{\min}^{\mathrm{P}}} \cdot \frac{\log(T(d|A^\star|)^{1/4})}{\log T}) \frac{K \log T}{g^\dagger}.$$

From the original leading term, the exact normalized coefficient induced by (Drugan & Nowe, 2013) is

$$C_{\mathrm{PUCB}} := 8 \cdot \frac{g^\dagger}{K} (\sum_{i \notin A^\star} \frac{1}{\Delta_i^{\mathrm{P}}}) \frac{\log(T(d|A^\star|)^{1/4})}{\log T}.$$

Our leading coefficient 64 is smaller exactly when

$$64 < C_{\mathrm{PUCB}},$$

that is, when

$$\frac{g^\dagger}{K} \sum_{i \notin A^\star} \frac{1}{\Delta_i^{\mathrm{P}}} > 8 \cdot \frac{\log T}{\log(T(d|A^\star|)^{1/4})}.$$

This is the exact pointwise comparison between the two leading terms.

The $\Delta_{\min}^{\mathrm{P}}$ rewriting used in Remark 6 is a coarser upper envelope. Indeed,

$$C_{\mathrm{PUCB}} \le \overline{C}_{\mathrm{PUCB}} := 8 \cdot \frac{K - |A^\star|}{K} \cdot \frac{g^\dagger}{\Delta_{\min}^{\mathrm{P}}} \cdot \frac{\log(T(d|A^\star|)^{1/4})}{\log T}.$$

Using

$$\frac{\log T}{\log(T(d|A^\star|)^{1/4})} = (1 + \frac{\log(d|A^\star|)}{4 \log T})^{-1},$$

the factor multiplying $g^\dagger / \Delta_{\min}^{\mathrm{P}}$ is asymptotically close to $8(K - |A^\star|)/K$. This upper envelope highlights the setting in which the two bounds separate most clearly: when $\Delta_{\min}^{\mathrm{P}} \ll g^\dagger$, or more generally when $\sum_{i \notin A^\star} 1/\Delta_i^{\mathrm{P}}$ is large, the earlier leading coefficient can be much larger than 64. If instead the dominated-arm Pareto gaps are all comparable to $g^\dagger$ and only a few dominated arms remain, the earlier coefficient is also constant-order and can be numerically smaller.

A.6.9 PROOF OF PROPOSITION 4

Proof. We first compute the Pareto geometry of the subclass and verify that the theorem parameter satisfies $g^\dagger = \Delta_{\mathrm{sc}}$. We then show that Pareto regret reduces exactly to the ordinary single-objective regret of the reduced Bernoulli bandit instance. Finally, because each reward vector is one Bernoulli sample duplicated across coordinates, the problem is equivalent to that reduced single-objective Bernoulli bandit, so Theorem 2 of (Lai & Robbins, 1985) applies.

Set

$$\theta_1 = \frac{1}{2} + \Delta_{\mathrm{sc}}, \qquad \theta_a = \frac{1}{2} \quad \text{for } a = 2, \dots, K.$$

Then every arm has mean vector

$$\mu_a = \theta_a \mathbf{1}.$$

Hence

$$\mu_1^{(j)} = \frac{1}{2} + \Delta_{\mathrm{sc}} > \frac{1}{2} = \mu_a^{(j)} \qquad \forall a \geq 2,\ \forall j \in [d],$$

so arm 1 dominates every other arm and therefore

$$A^\star = \{1\}.$$

Because arm 1 is the unique maximizer on every objective,

$$g_j = \mu_1^{(j)} - \max_{a \neq 1} \mu_a^{(j)} = \Delta_{\mathrm{sc}} \qquad \forall j \in [d].$$

Thus $g^\dagger = \Delta_{\mathrm{sc}}$.

For arm 1, $\Delta_1^{\mathrm{P}} = 0$. Fix any arm $a \geq 2$. Since

$$\mu_a + \epsilon \mathbf{1} = (\tfrac{1}{2} + \epsilon)\mathbf{1},$$

and every arm $b \geq 2$ has mean vector $(1/2)\mathbf{1}$, it is enough to compare only with μ_1. The lifted vector is dominated by $\mu_1 = (\frac{1}{2} + \Delta_{\mathrm{sc}})\mathbf{1}$ whenever $\epsilon < \Delta_{\mathrm{sc}}$, and it is no longer dominated once $\epsilon \geq \Delta_{\mathrm{sc}}$. Therefore

$$\Delta_a^{\mathrm{P}} = \Delta_{\mathrm{sc}} \qquad \text{for every } a = 2, \dots, K.$$

It follows that

$$R_T^{\mathrm{P}} = \sum_{a=2}^{K} \Delta_a^{\mathrm{P}} N_a(T) = \Delta_{\mathrm{sc}} \sum_{a=2}^{K} N_a(T).$$

Because each reward vector is obtained by duplicating one scalar Bernoulli sample across all coordinates, observing

$$Y_{a,n} = (Z_{a,n}, \dots, Z_{a,n})$$

is exactly the same as observing the scalar sample $Z_{a,n}$. Thus one pull in the duplicated-coordinate instance provides exactly the same information as one pull in the reduced single-objective K-armed Bernoulli bandit with means

$$\theta_1 = \frac{1}{2} + \Delta_{\mathrm{sc}}, \qquad \theta_a = \frac{1}{2} \quad \text{for } a = 2, \dots, K.$$

Let $N_a(T)$ denote the pull count of arm a up to time T. By assumption, the policy induced by the duplicated-coordinate reduction is uniformly good for the reduced Bernoulli bandit. We may therefore apply Theorem 2 of (Lai & Robbins, 1985), which lower-bounds the expected number of samples drawn from an inferior population under such a rule. Fix a suboptimal arm $a \in \{2, \dots, K\}$ and match it to population $j = a$ in Lai–Robbins. In our reduced Bernoulli bandit,

$$p(\theta_a) = \frac{1}{2}, \qquad p(\theta^\star) = \frac{1}{2} + \Delta_{\mathrm{sc}}.$$

The quantity $I(\theta_a, \theta^\star)$ is the least information cost of changing the inferior population into one whose mean exceeds $p(\theta^\star)$. For Bernoulli distributions this is the Bernoulli Kullback–Leibler divergence

$$\mathrm{kl}(p, q) := p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

Since the inferior arm has Bernoulli mean $1/2$, the alternative means making it better than the current best arm are exactly those with

$$q > \frac{1}{2} + \Delta_{\mathrm{sc}}.$$

Therefore,

$$I(\theta_a, \theta^\star) = \inf_{q > 1/2 + \Delta_{\mathrm{sc}}} \mathrm{kl}(\frac{1}{2}, q).$$

For fixed $p = 1/2$, the map $q \mapsto \mathrm{kl}(1/2, q)$ is increasing on $(1/2, 1)$, so

$$I(\theta_a, \theta^\star) = \mathrm{kl}(\frac{1}{2}, \frac{1}{2} + \Delta_{\mathrm{sc}}).$$

Hence Theorem 2 of (Lai & Robbins, 1985) gives, for each suboptimal arm $a = 2, \ldots, K$,

$$\liminf_{T \to \infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{\mathrm{kl}(1/2,\ 1/2 + \Delta_{\mathrm{sc}})}.$$

Summing over $a = 2, \ldots, K$ and using

$$R_T^{\mathrm{P}} = \Delta_{\mathrm{sc}} \sum_{a=2}^{K} N_a(T)$$

yields

$$\liminf_{T \to \infty} \frac{\mathbb{E}[R_T^{\mathrm{P}}]}{\log T} \geq (K-1) \frac{\Delta_{\mathrm{sc}}}{\mathrm{kl}(1/2,\ 1/2 + \Delta_{\mathrm{sc}})}.$$

Next,

$$\mathrm{kl}(\frac{1}{2}, \frac{1}{2} + \Delta_{\mathrm{sc}}) = -\frac{1}{2} \log(1 - 4\Delta_{\mathrm{sc}}^2).$$

Using

$$-\log(1-x) \leq \frac{x}{1-x} \qquad \text{for } 0 \leq x < 1,$$

with $x = 4\Delta_{\mathrm{sc}}^2$, we obtain

$$\mathrm{kl}(\frac{1}{2}, \frac{1}{2} + \Delta_{\mathrm{sc}}) \leq \frac{2\Delta_{\mathrm{sc}}^2}{1 - 4\Delta_{\mathrm{sc}}^2}.$$

Because $\Delta_{\mathrm{sc}} \leq 1/4$, we have $1 - 4\Delta_{\mathrm{sc}}^2 \geq 3/4$, and therefore

$$\mathrm{kl}(\frac{1}{2}, \frac{1}{2} + \Delta_{\mathrm{sc}}) \leq \frac{8}{3} \Delta_{\mathrm{sc}}^2.$$

Substituting this bound yields

$$\liminf_{T \to \infty} \frac{\mathbb{E}[R_T^{\mathrm{P}}]}{\log T} \geq \frac{3}{8} \frac{K-1}{\Delta_{\mathrm{sc}}}.$$

Since $g^\dagger = \Delta_{\mathrm{sc}}$, the claimed $\Omega(K \log T / g^\dagger)$ lower bound follows.

□