# THE ROLE OF COORDINATES IN PARETO REGRET FOR ADVERSARIAL MULTI-OBJECTIVE BANDITS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Adversarial multi-objective bandits help us optimize choices (arms) whose reward is a multidimensional vector chosen by an adversary and whose performance is measured by Pareto regret. We define loss as one minus reward and measure the easiness of a coordinate by the smallest cumulative loss of the arms on it, and call the coordinate easier when this quantity is smaller. Existing work suggests that in theory an easier coordinate may reduce Pareto regret. However, in practice, one may not know which coordinate is easier. On the negative side, we show that this lack of information eliminates the possibility: a smaller cumulative loss does not improve the worst-case order of Pareto regret. Precisely, let $L_d$ be the smallest cumulative loss along coordinate $d$ over $T$ rounds. For $K \geq 4$ arms, $T \geq 6$ rounds, and at least 2 coordinates, we prove that the minimax expected Pareto regret is $\Omega(\min\{T - L_0, \sqrt{K(T - L_0)}\})$ where $L_0 = \min_d L_d$. It is monotonically decreasing in $L_0$, even when $L_0$ is known. On the positive side, it motivates that other coordinates may suffice to attain the optimal rate. When $L_0$ is known, we apply Poly-INF to a fixed coordinate and obtain an upper bound on Pareto regret that exhibits the same order and thus matches the lower bound. Without such knowledge, we create reward-doubling Poly-INF that adapts to this unknown quantity while still attaining the matching minimax rate. Another implication is that it has no extra $\log T$ factor and is independent of the number or choice of coordinates, but only with a deliberate policy along the chosen coordinate.

## 1 INTRODUCTION

Multi-armed bandits (Auer et al., 2002a;b) have elegantly enabled sequential decision making in a broad spectrum of applications, dating back to classical recommender systems (McInerney et al., 2018; Mahadik et al., 2020; Yang et al., 2020; Zhu & Van Roy, 2023; Yi et al., 2023) and extending to emerging large language models (Bouneffouf & Feraud, 2026; Chen et al., 2026; Poon et al., 2026). Here, a decision maker sequentially selects an option from multiple options (often called arms) and receives information about its performance in the form of a scalar value (often called a reward). The goal is to maximize the cumulative performance of the chosen options, usually in the language of regret, which measures the cumulative difference between the reward obtained by always selecting the best arm in hindsight and the reward received by the decision maker. Bandit models broadly fall into two categories, stochastic (Auer et al., 2002a) and adversarial (Auer et al., 2002b), depending on if rewards are independently drawn from a fixed distribution or chosen by an adversary. Both have witnessed measurable successes in theory and practice via the community's joint efforts (Bubeck & Cesa-Bianchi, 2012; Bubeck et al., 2013; Liau et al., 2018; Komiyama et al., 2015; Bouneffouf et al., 2020; Allenberg et al., 2006; Neu, 2015b; Lykouris et al., 2022; Wei & Luo, 2018).

The evolution of these applications has imposed the challenge of optimally selecting options whose performance/reward includes multiple objectives (coordinates) (Lacerda, 2017; Mehrotra et al., 2020; Wanigasekara et al., 2019; Li et al., 2026; Xue et al., 2026). Multi-objective bandits (MO-MAB) have emerged as a promising remedy (Drugan & Nowe, 2013; Busa-Fekete et al., 2017a; Garivier et al., 2024; Xu & Klabjan, 2023; Roijers et al., 2017; Yahyaa et al., 2014; Auer et al., 2016; Ararat & Tekin, 2023; Hüyük & Tekin, 2021; Lu et al., 2019), where the reward of each arm is a multidimensional vector and the optimality of arms, namely Pareto optimality, is defined via Pareto dominance. Pareto dominance (Drugan & Nowe, 2013; Lu et al., 2019; Yahyaa et al., 2014; Garivier et al., 2024) compares two arms across coordinates. An arm is Pareto optimal if it is at least as good on every



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.

coordinate and strictly better on at least one; otherwise, it is suboptimal (dominated). This intricate geometry motivates algebraic notions of Pareto regret, w.r.t. the minimum value to be added uniformly to *all coordinates* when comparing with Pareto optimal arms in hindsight, although their precise definitions differ between stochastic (Drugan & Nowe, 2013) and adversarial models (Xu & Klabjan, 2023). This minimum reflects that the improvement needed to overcome Pareto domination can differ across coordinates. By Pareto dominance, performing well on even *one coordinate* can be sufficient to prevent an arm from being dominated. A smaller required improvement means that domination is easier to overcome along that coordinate. This heterogeneity suggests that coordinates, such as easier ones, may play a role in Pareto regret and thus in surmounting the challenges of MO-MAB.

To date, existing studies on MO-MAB have jointly solved the puzzle for the stochastic problem, where the reward vectors are independent and identically distributed (i.i.d.) across the time horizon. The seminal studies measure Pareto regret by cumulative Pareto gaps (Drugan & Nowe, 2013; Turgay et al., 2018; Lu et al., 2019), specify how objectives are combined or ordered (Busa-Fekete et al., 2017b; Hüyük & Tekin, 2021), or identify Pareto optimal arms (Auer et al., 2016; Ararat & Tekin, 2023). Here, a coordinate is easier when the gap between its top two expected rewards is larger, making the best arm easier to distinguish. Under assumptions, Guan & Xu (2026) recently show that Pareto regret can follow the easiest coordinate. The algorithm searches for such a coordinate across arms and coordinates, then certifies an arm as Pareto optimal. Notably, this search is feasible with this nice i.i.d. structure, as repeated samples provide persistent statistical information and thus offer hope for learning. Yet, such an assumption can oftentimes fail measurably in the adversarial counterpart. Consequently, these arguments, and the insights behind them, do not extend to adversarial MO-MAB.

In fact, to the best of our knowledge, much less is known about this puzzle in adversarial MO-MAB. Encouragingly, several results from the limited literature on adversarial MO-MAB and classical adversarial bandits suggest that an easy coordinate may still help. Among these, Xu & Klabjan (2023) introduce a definition of Pareto regret and show that it can be upper bounded by classical regret on any coordinate fixed before play. Accordingly, they select an arbitrary coordinate, as *their guarantees depend only on the time horizon.* The roles of the coordinates thus seem to be entirely irrelevant. However, as suggested by the line of work on adversarial bandits, the horizon dependence can be overly pessimistic. Indeed, the minimum cumulative loss among the arms can be smaller than the horizon where loss is one minus reward. Seminal work has exploited this structure and achieved regret upper bounds that increase with this minimum cumulative loss (Allenberg et al., 2006; Neu, 2015b; Lykouris et al., 2022; Wei & Luo, 2018). Hence, *if the coordinate with the smallest cumulative loss were known, namely, the easiest coordinate, one might hope to incur the least Pareto regret by optimizing along it.* It remains unexplored in adversarial MO-MAB if, following these results, coordinate structure can similarly be exploited to reduce Pareto regret as in stochastic MO-MAB.

Yet, the decision maker generally does not know which coordinate attains this minimum. This leaves us with a dilemma. Directly applying existing bandit algorithms to an arbitrary coordinate seems to yield a Pareto regret guarantee that scales with the loss on that coordinate, which can exceed the minimum coordinate loss and thus be less desirable. Identifying the coordinate attaining this minimum, on the other hand, may itself incur substantial regret. Importantly, we do not yet know whether either difficulty is fundamental in adversarial MO-MAB. The gap in Pareto regret may be an artifact of the presumed scaling with the minimum coordinate loss, analogous to classical bandits, and the apparent need to identify that coordinate may simply arise from this dependence. This raises:

*Does the Pareto regret scale with minimum coordinate loss in some sense, and should we try to identify coordinates en route to the optimality?*

Counterintuitively, the optimal rate is governed by the minimum coordinate loss but is monotonically decreasing in it, and attaining this rate does not require identifying the coordinate that realizes the minimum. Precisely, on coordinate $d$, let $L_d$ denote the minimum cumulative loss among the arms, and let $L^\star = \min_d L_d$. Thus, a small $L^\star$ means that some arm has small cumulative loss (thus large cumulative reward) on at least one coordinate, but the coordinate attaining $L^\star$ is unknown. *Our first result is surprisingly negative.* For every $L_0 \in [0, T]$, we derive a lower bound of order $\min\{T - L_0, \sqrt{K(T - L_0)}\}$ over the class of instances satisfying $L^\star = L_0$. This lower bound is monotonically decreasing in $L_0$; making the easiest coordinate easier, i.e., decreasing its minimum cumulative loss instead increases the lower bound. *On the flip side, this negative result has an optimistic implication: perhaps we do not always need to identify coordinates.* We confirm this by

showing that the optimal rate in $L_0$ can be attained *using any fixed coordinate but with deliberately designed algorithms along the chosen coordinate*, perhaps as a metaphor for a philosophy of life that *"many roads lead to Rome"* (Kruglanski et al., 2011; Pažout et al., 2026). We show this with existing and new methods that match the lower bound and relatively cleanly remove several extra factors.

The main contributions are as follows. We first extend Pareto regret in Xu & Klabjan (2023) so that it is well defined. We determine the optimal worst-case order of Pareto regret as a function of $L_0$, the minimum coordinate loss. Surprisingly, this lower bound is monotone decreasing in $L_0$. Hence in contrast to the positive scaling in bandits, it yields a new bound authentic to MO-MAB. Even when this loss is 0, the lower bound has the same order as in the unrestricted case.

We then develop routes to optimize Pareto regret without finding this coordinate attaining $L_0$; in fact any coordinate suffices for *minimax optimality* but the algorithm along it has to be targeted. When $L_0$ is known, we apply Poly-INF (Audibert & Bubeck, 2010) and set the reward scale using the horizon minus this loss. When it is unknown, we create a reward-doubling version that restarts the algorithm and doubles the scale whenever the reward collected in the current run reaches that scale.

We prove that the upper bound of these methods matches the lower bound, particularly with respect to $L_0$. This also implies the minimax optimality of the methods and the tightness of the lower bound. As a separate contribution, they also sharpen the dependence on the problem parameters. They remove the common $\sqrt{\log T}$ factor and are independent of the number of coordinates, with the order governed by the intrinsic complexity $L_0$. We also adapt multiple popular algorithms from adversarial bandits and derive corresponding bounds to assess if they attain the optimal rate in MO-MAB; several yield different rates or extra factors. Lastly, numerical experiments further validate these findings.

## 2 Problem Formulation

We study the adversarial multi-objective multi-armed bandit (MO-MAB) problem. There are $K$ arms, $T$ rounds, and $D$ reward coordinates. Before the first round, an oblivious adversary specifies the reward vectors $r = r^{i,t} = (r_d^{i,t})_{1\le d\le D} \in [0,1]^D$ of every arm $i \in [K]$ in every round $t \in [T]$. We call this collection $r$ the reward sequence. In each round $1 \le t \le T$, the learner chooses an arm $a_t$, possibly using randomization, and observes only the chosen arm's reward $r^{a_t,t}$.

Pareto dominance is a standard tool for comparing vectors. For vectors $a, b \in \mathbb{R}^D$, $a$ weakly dominates $b$ if $a_d \ge b_d$ for every coordinate $d$; we write $a \succeq b$, or equivalently $b \preceq a$. If in addition $a \ne b$, then $a$ dominates $b$, denoted by $a \succ b$. Vectors $a$ and $b$ are incomparable, denoted by $a \parallel b$, if $a_d > b_d$ and $a_{d'} < b_{d'}$ for some coordinates $d, d'$. Finally, $\mathbf{1} \in \mathbb{R}^D$ denotes the all-ones vector.

Let $G_i := \sum_{t=1}^T r^{i,t}$ denote the cumulative reward vector of arm $i$, which is unknown during the process. We define the set of distinct nondominated cumulative reward vectors as $O' := \{G_i : i \in [K],\ \nexists j \in [K] \text{ with } G_j \succ G_i\}$ (1). As $O'$ is defined over reward vectors rather than arms, tied arms contribute only one vector, and $O'$ is always nonempty. We call $O'$ the Pareto front of the cumulative reward vectors. Ultimately, we care about which arms are optimal. An arm whose cumulative reward vector belongs to $O'$ is Pareto optimal, and the collection of such arms, denoted by $O$ is called the Pareto-optimal set, or Pareto set for short. Naturally, both $O'$ and $O$ are unknown to the learner.

### 2.1 Regret

Regret has long been the go-to language in bandits for measuring how far our arm choices are from choosing the optimal arms as if with hindsight knowledge of $G_i$, which, of course, is unavailable throughout. We first recap scalar regret in classical bandits (also called coordinate regret) on an individual coordinate, which helps understand the roles of coordinates. We then turn to our main objective, Pareto regret, which evaluates the entire reward vector through Pareto dominance.

Let $A_T := \sum_{t=1}^T r^{a_t,t}$ denote the cumulative reward vector of the learner with arm sequence $\mathcal{A} = \{a_t\}_{t=1}^T$; scalar $A_{T,d}$ is its element in coordinate $d$. The comparison with the best arm (Auer et al., 2002b; Lattimore & Szepesvári, 2020) defines the coordinate regret as $R_T^d := \max_i G_{i,d} - A_{T,d}$ (2). We denote Pareto regret by $R_T^{\text{fs}}(r)$, or simply $R_T^{\text{fs}}$ with a slight abuse of notation:

$$R_T^{\text{fs}} := \inf\{\epsilon \ge 0 : \nexists \sigma \in O' \text{ with } \sigma \succ A_T + \epsilon\mathbf{1}\} = [\max_{i\in[K]} \min_{d\in[D]} (G_{i,d} - A_{T,d})]_+. \tag{3}$$

Eq. (3) has a particular elegant interpretation. For every $\epsilon \geq 0$, we have the equivalence: $R_T^{\text{fs}} > \epsilon \iff \exists i \in [K]$ s.t. $G_{i,d} > A_{T,d} + \epsilon$ for every $d \in [D]$. Intuitively, it measures the largest uniform improvement that one arm offers over the learner's cumulative reward across all coordinates. Positive regret captures a uniform advantage of one fixed arm over the learner across all objectives, while zero regret means that every fixed arm fails to improve on the learner in at least one objective. To see why the second equality holds, we prove the following result regarding the minimax expression.

**Proposition 2.1** (Pathwise representation)**.** *For every reward sequence and every realized play path, the two expressions in (3) are equal. The maximum over arms can equivalently be taken over the distinct nondominated vectors in $O'$. The identity holds even when cumulative arm vectors coincide or no shift makes the learner incomparable with every vector on the front.*

Notably, our definition builds on and extends that of Xu & Klabjan (2023). They define Pareto regret as $\min\{\epsilon \geq 0 : A_T + \epsilon \mathbf{1} \parallel \sigma \text{ for every } \sigma \in O'\}$, on the domain where $A_T$ is weakly Pareto dominated by every vector in $O'$. However, it does not cover three edge cases. First, the feasible set in the minimum can be empty. For $A_T = (0, 0)$ and $O' = \{(1, 1)\}$, no shift makes $A_T + \epsilon\mathbf{1}$ incomparable with $(1, 1)$, so $\text{Dist}(A_T, O)$ is undefined. Second, feasible shifts may exist but the minimum may not be attained. For $A_T = (0, 0)$ and $O' = \{(1, 2)\}$, the feasible shifts are $\epsilon \in (1, 2)$, which has an infimum but no minimum. Third, ties can make the Pareto front $O'$ empty, as the Pareto set $O$ is defined before $O'$ therein. This order matters. If two arms have the same cumulative reward vector $G_i = (1, 1)$, neither satisfies the non-domination condition, so both can be excluded from $O$. This notion is also closely related to Jiang et al. (2023) in multi-objective online convex optimization, but here the comparators are fixed arms and no convexity is assumed.

We address these issues by first deduplicating vectors in $O'$ and then defining the Pareto set of arms, while using the boundary of the same uniform-shift condition. This preserves the Pareto comparison and ensures a well defined regret for *every reward sequence*. To formalize the relation, let us define $a_- = \max_{\sigma \in O} \min_d(\sigma_d - A_{T,d})$ and $a_+ = \min_{\sigma \in O} \max_d(\sigma_d - A_{T,d})$ (4). Our definition agrees with that of Xu & Klabjan (2023) whenever theirs is well defined and remains valid otherwise.

**Proposition 2.2.** *Suppose $O' \neq \varnothing$ and $A_T \preceq \sigma$ for every $\sigma \in O'$. The feasible shifts $\epsilon$ in Xu & Klabjan (2023) form $(a_-, a_+)$. When this interval is nonempty, its infimum $a_-$ is at most $R_T^{\text{fs}}$, but its minimum is not attained. If vectors $G_i$ are distinct, their front equals $O'$, and the infimum equals $R_T^{\text{fs}}$.*

### 2.2 Mathematical form of the main question

We are now ready to present the precise form of our main question of interest. For coordinate $d$, let $U_d(r) := \max_i G_{i,d}$ be the cumulative reward of its best arm, and let $L_d(r) := T - U_d(r)$ be that arm's cumulative loss, obtained by subtracting each step's reward coordinate from 1. Notably, the reward and loss notation give identical regret. In MO-MAB, we have $R_T^{\text{fs}} \leq [\max_i(G_{i,d} - A_{T,d})]_+ = [R_T^d]_+$. In classical bandits, the regret bound becomes smaller as $L_d$ decreases (Lykouris et al., 2022). Thus, a coordinate with smaller $L_d$ can give a tighter bound on Pareto regret. The smallest loss and the corresponding reward are $L^\star(r) := \min_d L_d(r) = \min_{i,d} \sum_{t=1}^T \big(1 - r_d^{i,t}\big) = T - U, U^\star(r) := \max_d U_d = \max_{i,d} G_{i,d}$ (5). We use $d^\star(r)$ to denote a coordinate attaining this minimum. In other words, $L^\star(r) = L_1$ when $D = 1$; the regret scales with $L_1$.

The first part of the main question thus concerns whether Pareto regret admits an analogue through $L^\star(r)$ when $D \geq 2$. If such a dependence exists, then all reward sequences with the same value $L^\star(r) = L_0$ must share the same bound. The sharpest such guarantee at $L_0$ is therefore obtained by taking the worst case over these sequences and then optimizing over policies. Precisely, for a fixed $L_0 \in [0, T]$, the minimax Pareto regret among reward sequences with $L^\star = L_0$ is

$$\mathcal{V}_{K,D,T}(L_0) := \inf_{\mathcal{A}} \sup_{r:\, L^\star(r) = L_0} \mathbb{E}_{\mathcal{A}}[R_T^{\text{fs}}] \tag{6}$$

The infimum is over all policies $\mathcal{A}$, so that it represents the optimal policy; the expectation accounts for the policy's randomness. The supremum is over all instances with the same $L_0$, so that it represents the worst case. Thus, $\mathcal{V}_{K,D,T}(L_0)$ gives the optimal worst-case Pareto regret for $L_0$. Fixing $L^\star = L_0$ means that the losses on other coordinates can still vary. Notably, classical adversarial bandits are already minimax in that it suffices to exhibit a worst-case reward sequence. Our formulation refines this notion by additionally parameterizing the minimax value by $L_0$, a quantity arising

from the multiple coordinates in MO-MAB, alongside $K$ and $T$. We also consider the case where all coordinate losses are fixed, but the learner does not know which value corresponds to which coordinate. Specifically, define $V^{\text{perm}}_{K,D,T}(r) := \inf_{\mathcal{A}} \sup_{r \in \mathcal{C}(r)} \mathbb{E}_{\mathcal{A}}[R^{\text{fs}}_T]$, where $\mathcal{C}(r)$ contains all reward sequences obtained by permuting the coordinates of the same reward sequence $r$.

The question is whether these quantities depend on $L_0$ and, further, scale with $L_0$, as suggested by classical adversarial bandits. This also motivates us the second part. Let $\mathcal{A}$ denote a policy that commits to one coordinate $d$ before the first round, without knowing $d^\star(r)$, and operates only on that coordinate. Xu & Klabjan (2023) show that such a policy can attain the optimal order w.r.t. the horizon $T$. With a more carefully designed policy $\mathcal{A}$, can the upper bound $\sup_{r:\, L^\star(r)=L_0} \mathbb{E}_{\mathcal{A}}[R^{\text{fs}}_T]$ attain the same order as $\mathcal{V}_{K,D,T}(L_0)$, particularly w.r.t. the coordinate-related quantity $L_0$?

## 3 Lower Bound Analyses

To answer the first part, this section is dedicated to the analyses of the minimax Pareto regret. Before analyzing the general dependence on $L_0$, we begin with the special case $L_0 = 0$ to gather insights with more interpretable intuition and motivate the general-case analysis. Here, $L_0 = 0$ means a problem class where there exists a coordinate $d^\star$ on which the best arm incurs zero cumulative loss; equivalently, that arm receives reward one on coordinate $d^\star$ in every round. The main challenge is that neither a lower bound on coordinate regret nor a condition on coordinate loss directly implies a lower bound on Pareto regret, which depends on the joint rewards across coordinates. Our construction of instances for this class links the coordinates to reflect coordinate-level properties.

Let us first introduce several notations for ease of presentations. We denote $q := \lfloor K/2 \rfloor$, $W_q := \min\{T, \sqrt{qT}\}$, $h := \lfloor W_q/256 \rfloor$, $n := T - h$, and $\alpha := \frac{1}{8}\min\{1, \sqrt{q/n}\}$. We first consider instances with $W_q \geq 512$. We split $2q$ arms into groups $A_1, \ldots, A_q$ and $B_1, \ldots, B_q$; each group has at least two arms when we assume the number of arms $K \geq 4$. During the first $h$ rounds, every $A_i$ receives $(1, 0)$, and every $B_i$ receives $(0, 1)$. Before play, one chooses a case $J \in \{A, B\}$ uniformly and a hidden index $I \in [q]$ uniformly as in Figure 1. The rewards $Y_{i,t}$ are independent Bernoulli with mean $1/2 + \alpha$ for $i = I$ and mean $1/2$ otherwise. If $K$ is odd, the remaining arm receives $(0, 0)$ during the first block and the reward $(0, 1/2)$ in case $A$ or $(1/2, 0)$ in case $B$ during the remaining rounds. Each reward sequence is therefore oblivious. In case $A$, every $A_i$ receives reward one in coordinate 1 on all $T$ rounds. In case $B$, every $B_i$ receives reward one in coordinate 2 on all $T$ rounds. Hence every realization has $L^\star = 0$. Probabilities below include these draws and the algorithm's randomization.

**Case A, coordinate 1 has zero loss**

| Group | First $h$ rounds | Remaining $n$ rounds |
|---|---|---|
| $A_i$ | $(1, 0)$ | $(1, Y_{i,t})$ |
| $B_i$ | $(0, 1)$ | $(0, 1/2)$ |

The arm with larger mean reward is in group $A$.

**Case B, coordinate 2 has zero loss**

| Group | First $h$ rounds | Remaining $n$ rounds |
|---|---|---|
| $A_i$ | $(1, 0)$ | $(1/2, 0)$ |
| $B_i$ | $(0, 1)$ | $(Y_{i,t}, 1)$ |

The arm with larger mean reward is in group $B$.

Figure 1: The first block is identical; the remaining rewards determine which group has the zero-loss coordinate and better arm in the other coordinate.

The first block ensures that the learner falls behind in at least one of the two possible coordinates. Let $X_A$ and $X_B$ denote the numbers of pulls from groups $A$ and $B$ during this block. If $J = A$, the shortfall is $d_A := h - X_A$; if $J = B$, it is $d_B := h - X_B$. Since $X_A + X_B \leq h$, we have $d_A + d_B \geq h$. Moreover, the observations from the first block do not reveal $J$. Therefore, $\Pr\{d_J \geq h/2 \mid \text{history from the first block}\} \geq 0.5$ (7). After this block, every arm in group $J$, including the hidden arm, receives reward one on this coordinate; the shortfall $d_J$ cannot decrease.

On the other coordinate, the remaining $n$ rounds form a $q$-armed bandit problem with a hidden arm whose mean reward is larger by $\alpha$. Index $J$ may be known afterwards. Let $C_I$ denote the reward of the hidden arm and $S$ the learner's reward on this coordinate over the remaining $n$ rounds. With $V = \min\{n, \sqrt{qn}\}$, Lemma A.3 gives $\Pr\{C_I - S \geq V/64 \mid \text{history from the first block}, J\} \geq 1/12288$ (8). Our choice of $h$ and $n$ gives $h \geq W_q/512$, $V \geq (255/256)W_q$, and $V/64 \geq 2h$. When both events in (7) and (8) occur, the same hidden arm is ahead on both coordinates. On the zero-loss coordinate, it is ahead by $\geq h/2$. On the other coordinate, the learner can make up $\leq h$ during the

first block, so the hidden arm is ahead overall by at least $V/64 - h \geq h$. Thus, $R_T^{\text{fs}} \geq h/2$. The two events occur together with probability at least $1/24576$, and hence $\mathbb{E}[R_T^{\text{fs}}] \geq h/49152 = \Omega(W_q)$.

It remains to consider $W_q < 512$. One chooses one arm uniformly at random, gives it reward $(1, 1)$ in the first round, and gives every other arm $(0, 0)$. From $t = 2$ onward, every arm receives $(1, 1)$. The expected regret is then $1 - 1/K \geq 3/4$. Since $q = \lfloor K/2 \rfloor \geq K/4$ for $K \geq 4$, we have $W_q \geq \frac{1}{2} \min\{T, \sqrt{KT}\}$. Since $W_q < 512$, this implies $\min\{T, \sqrt{KT}\} < 1024$, so the constant lower bound is of the required order. Together with $W_q \geq 512$, this proves the result for all $T \geq 6$.

Finally, for any randomized algorithm, averaging over the construction gives a reward sequence with at least the same expected regret. For $D > 2$, we copy coordinate 1 to the remaining coordinates.

**Theorem 3.1** (Zero-loss)**.** *For $K \geq 4, T \geq 6, D \geq 2$ and any randomized algorithm, there is a deterministic reward sequence fixed before play with $L^\star = 0$ satisfying $\mathcal{V}_{K,D,T}(0) = \Omega(\min\{T, \sqrt{KT}\})$.*

For fixed $K$, expected regret can grow as $\sqrt{T}$ even though an arm has zero loss in one coordinate. Knowing the value $L^\star = 0$ does not identify that coordinate. In contrast, if such coordinate is known apriori, an elimination rule guarantees $R_T^{\text{fs}} \leq \min\{T, K - 1\}$ (Lemma A.4). When $D = 1$, it is known automatically. The gap therefore comes from not knowing which coordinate has zero loss.

For a general class $L^\star = L_0$ and $U_0 = T - L_0$. If $U_0 = 0$, the lower bound is immediate. If $0 < U_0 < 1$, choose one arm uniformly at random and give it reward $U_0\mathbf{1}$ in the first round, with all other rewards equal to zero. Then $L^\star = L_0$, and the expected regret is $(1 - 1/K)U_0$. Now suppose $U_0 \geq 1$, and let $m := \lfloor U_0 \rfloor$. We embed the zero-loss construction into the first $m$ rounds, while the learner may still use the original $T$ and $L_0$. If $m < T$, every arm receives the same reward $(U_0 - m)\mathbf{1}$ in round $m + 1$, followed by zero rewards. This makes the largest cumulative reward over all arms and coordinates exactly $U_0$, and hence $L^\star = L_0$, while the common reward does not change $R_T^{\text{fs}}$. Since $m \geq U_0/2$, $\min\{m, \sqrt{Km}\} \geq \frac{1}{2} \min\{U_0, \sqrt{KU_0}\}$. It therefore yields the general bound:

**Theorem 3.2.** *Assuming the same, for any $L_0$, we have $\mathcal{V}_{K,D,T}(L_0) = \Omega\big(\min\{U_0, \sqrt{KU_0}\}\big)$.*

At $L_0 = 0$, we have $U_0 = T$ consistent with Theorem 3.1. It answers the main question. For $U_0 \leq K$, the lower bound is $\Omega(U_0)$. For $U_0 \geq K$, it is $\Omega(\sqrt{KU_0})$, which comes from the cost of finding the hidden best arm. As $L_0$ increases, $U_0 = T - L_0$ decreases, and so does the regret lower bound.

**Corollary 3.3** (Permutation)**.** *Assuming the same, we have $\mathcal{V}^{\text{perm}}_{K,D,T}(r) \geq 2^{-28} \min\{U_0, \sqrt{KU_0}\}$.*

The permutation case reveals a useful distinction. It incurs an extra factor of $2^{-28}$ w.r.t. the class with fixed $L_0$, yet retains the same dependence on $L_0$. In this smaller class, the reward sequence is fixed and only the coordinates are permuted. By contrast, fixing $L_0$ alone allows the remaining coordinate losses, and hence the reward sequence, to vary. Thus, uncertainty about which coordinate attains $L_0$ is sufficient for the rate, while allowing other losses to vary strengthens the constant in our lower bound. Hypothetically, a similar phenomenon can arise from permuting arms in adversarial bandits.

*Comparison.* Xu & Klabjan (2023) obtain lower bounds of order $\sqrt{T}$ by reducing to bandits with identical coordinate means or rewards. Even when $L_0 = 0$, identical rewards across coordinates would allow the learner to choose any coordinate and obtain regret bounds of order $O(1)$ by Lemma A.4, which cannot yield the desired rate discovered herein. In contrast, Theorem 3.2 proves the rate over reward sequences with $L^\star = L_0$, and Corollary 3.3 further relaxes this to permutation. Lai & Robbins (1985) establish an asymptotic lower bound for stochastic instances, assuming subpolynomial regret over the policy class. Our lower bound holds at a finite horizon and for every randomized policy.

## 4 Algorithms and Upper Bound Analyses

The lower bound analysis shows that any guarantee on Pareto regret that holds uniformly over the class $L^\star = L_0$ must incur the rate $\min\{T - L_0, \sqrt{K(T - L_0)}\}$. We now turn to the second part of our question: whether this rate can be attained without identifying $d^\star$, the coordinate attaining $L_0$. The encouraging answer is yes. It is enough to choose any coordinate before the first round and carefully optimize along it. We introduce two algorithms, both of which attain the same rate. When $L_0$ is known, the condition $L^\star = L_0$ provides the common reward bound $U_0 = T - L_0$ across coordinates, and we run Poly-INF on an arbitrary coordinate using this bound. When $L_0$ is unknown, we develop a reward-doubling version of Poly-INF that adapts the reward scale from observations.

The intuition behind these results follows from an interplay between the rate identified by the lower bound and Pareto geometry. The rate from the lower bound is monotonically decreasing in $L_0$, while $L_0$ also lower bounds the cumulative loss on every coordinate. This contrasts with the intuition from classical adversarial bandits that smaller loss is better. However, the rate increases with the complementary reward quantity $U_0 = T - L_0$. For every coordinate $d$, we have $U_d \leq U_0$. Importantly, some algorithms for classical adversarial bandits admit regret bounds that increase with $U_d$. Thus, any coordinate regret can also scale with $U_0$. The following lemma crucially builds the route from coordinate regret to Pareto regret: the positive part of the regret on any fixed coordinate controls Pareto regret. The key subsequent task is therefore to control this positive part, rather than only the expected regret commonly analyzed before, thereby transferring the same rate to Pareto regret.

**Lemma 4.1** (Reduction). *For every reward sequence and play path, and every coordinate $d'$, $0 \leq R_T^{\mathrm{fs}} \leq [R_T^{d'}]_+$ and $R_T^{\mathrm{fs}} \leq U$ (9). More generally, let $\Delta_D$ be the set of nonnegative coordinate weights that sum to one. For any $w \in \Delta_D$, put $g_{i,t} = \sum_d w_d r_d^{i,t}$ and $R_T^w = \max_i \sum_t g_{i,t} - \sum_t g_{a_t,t}$. Then $R_T^{\mathrm{fs}} \leq [R_T^w]_+$ and $\max_i \sum_t g_{i,t} \leq U$ (10).*

We use the coordinate case in what follows, while, more generally, Lemma 4.1 shows that the same comparison extends to any fixed weighted average of the coordinates. One subtlety is that the reduction involves $[R_T^{d'}]_+$: a bound on signed expected regret alone is insufficient, as positive and negative regret can offset. The weighted version may have implications for scenarios where the decision maker has known a priori preferences, in the form of weights, over the coordinates.

*Remark* 4.2. The coordinate regret $R_T^d$ is the standard regret against the best fixed arm in hindsight and can be negative, since the learner may switch among arms and outperform every fixed arm on coordinate $d$. The sign becomes relevant here because Pareto regret is nonnegative; consequently, our comparison involves the positive part $[R_T^d]_+$, rather than $R_T^d$ itself. In particular, a bound on $\mathbb{E}[R_T^d]$ does not by itself imply the same bound on $\mathbb{E}[[R_T^d]_+]$. Moreover, the weighted version generalizes the result in Xu & Klabjan (2023), which considers only the coordinate case without weights.

These results reduce the remaining task to developing scalar methods whose positive part of the regret on an arbitrary fixed coordinate can be controlled at the target rate in $U_0/L_0$. We first consider the case where $L_0$ is known (Section 4.1) and then the case where it is unknown (Section 4.2).

### 4.1 When $L_0$ is known

The lower bound holds even if $L_0$ is known, which motivates us to first consider this setting. We use the Poly-INF policy of Audibert & Bubeck (2010, Theorem 18), whose parameters can be set using an upper bound $B \geq 81K$ on cumulative reward. We denote the resulting policy by Poly-INF$(B)$. Since $U_0$ is known and the cumulative reward of every arm on every coordinate is at most $U_0$, we can set $B = U_0$ for any coordinate fixed before the first round.

Algorithm 2 therefore runs Poly-INF$(U_0)$ when $U_0 \geq 81K$. When $U_0 < 81K$, it instead plays a fixed arm throughout, as the deterministic bound $R_T^{\mathrm{fs}} \leq U_0$ already gives the desired rate. To see this, for every $0 \leq U_0 < 81K$, $R_T^{\mathrm{fs}} \leq U_0 \leq 9\min\{U_0, \sqrt{KU_0}\}$. If $U_0 \leq K$, the minimum equals $U_0$. If $K < U_0 < 81K$, then $U_0/\sqrt{KU_0} = \sqrt{U_0/K} < 9$. Naturally, it leads to the following:

**Theorem 4.3.** *For $K \geq 2, T \geq 2, D \geq 1$ and every $L_0 \in [0, T]$, with $U_0 = T - L_0$, Algorithm 2 with $B = U_0$ satisfies $\mathbb{E}[R_T^{\mathrm{fs}}] \leq 10\min\{U_0, \sqrt{KU_0}\}$ on every reward sequence with $L^\star = L_0$.*

For $K \geq 4, T \geq 6, D \geq 2$, this bound holds uniformly over all reward sequences $r$ satisfying $L^\star(r) = L_0$, for each fixed known $L_0$. Taking the supremum over this class therefore yields the same order, which matches the lower bound in Theorem 3.2 for $\mathcal{V}_{K,D,T}(L_0)$. Hence, a policy using any coordinate fixed in advance is minimax optimal without identifying the coordinate attaining $L_0$.

A new aspect of our analysis is controlling the positive part of scalar regret. For the Poly-INF policy of Audibert & Bubeck (2010, Theorem 18), the parameters are the same for every confidence level. Integrating its tail bound therefore gives, for $U_0 \geq 81K$, $\mathbb{E}[[R_T^{d'}]_+] \leq (8.5 + \sqrt{2})\sqrt{KU_0}$. Together with the reduction from Pareto regret to coordinate regret, this yields the upper bound.

Notably, our choice of Poly-INF is deliberate. To see why, we also examine several popular algorithms. In particular, modified EXP3 uses the logarithmic estimator of Audibert & Bubeck (2010, Theorem 21) with $B = U_0$. Gain-based EXP3 and EXP3.P retain their updates, with their exploration parameters

set using $U_0$. EXP3-IX uses a horizon-based learning rate. We find that they may incur additional factors in their regret bounds or retain a dependence on the time horizon $T$ as shown in Lemma A.14, which can be larger than $U_0$ since $U_0 \leq T$. Thus, *while the choice of coordinate does not matter for attaining the minimax rate, how the learner optimizes along that coordinate does.*

**Proposition 4.4** (EXP3 variants)**.** *Fix $K \geq 2, T \geq 2, D \geq 1$ and a coordinate $d'$ before play. The corresponding bounds on $\mathbb{E}[R_T^{\text{fs}}]$ are $O(\min\{U_0, \sqrt{KU_0 \log(3K)}\})$ for modified EXP3, $O(\min\{U_0, (K \log K)^{1/3} U_0^{2/3}\})$ for gain-based EXP3, $O(\min\{U_0, \sqrt{KT \log(KT)}\})$ for EXP3.P and adversarial MO-KS, and $O(\min\{U_0, \sqrt{KT \log K}\})$ for EXP3-IX.*

At the level of these guarantees, Poly-INF is the only algorithm among those considered that matches the lower bound uniformly over $U_0$. Modified EXP3 incurs an additional $\sqrt{\log K}$ factor, gain-based EXP3 does not attain the target dependence on $U_0$, and EXP3.P and EXP3-IX retain a dependence on $T$. This comparison shows that attaining the minimax rate does not require identifying a favorable coordinate, but does require a favorable algorithm with the appropriate dependence on $U_0$.

### 4.2 When $L_0$ is unknown

More generally, $L_0$ may be unknown to the learner, for example when the problem setting is completely black-box. In this case, a direct fallback is to run Poly-INF with $B = T$, the worst-case upper bound on $U_0$, though this can clearly be overly pessimistic. This gives $O(\min\{U_0, \sqrt{KT}\})$, where the second term fails to reflect the role of $U_0$. To recover this dependence without knowing $L_0$, we propose **reward-doubling Poly-INF, Algorithm 5.** It fixes one coordinate, starts with $B = 81K$, and runs Poly-INF$(B)$. Whenever the reward collected during the current run reaches $B$, it doubles $B$ and starts a fresh copy on the next round. In this way, the scale adapts to the observed rewards while the chosen coordinate remains fixed.

**Theorem 4.5.** *Let $K \geq 2$, $T \geq 2$, and $D \geq 1$, and let Alg. 5 run without knowing $L_0$. For every $L_0 \in [0, T]$, we have $\sup_{r: L_*(r) = L_0} \mathbb{E}[R_T^{\text{fs}}] \leq 100 \min\{T - L_0, \sqrt{K(T - L_0)}\}$* (11).

The doubling scheme allows the unknown reward scale to reveal itself through the observed rewards. Across epochs, we compare against the best arm on the chosen coordinate, so the epoch regrets sum to the full coordinate regret. Whenever this regret is positive, the largest scale reached by the algorithm is controlled by the cumulative reward $U_d$ of that arm. Since the epoch scales grow geometrically, the regret contributions across epochs remain of order $\sqrt{KU_d}$. A single supermartingale controls the estimation error across all epochs, avoiding a union bound over restarts and hence an additional logarithmic factor. Combined with $U_d \leq U_0$ and Lemma 4.1, this yields the regret bound.

An important implication is that the optimal rate dependence on $L_0$ can be attained even when neither its value nor the coordinate attaining it is known. The learner adapts through the rewards observed on an arbitrary fixed coordinate, while $U_d \leq U_0 = T - L_0$ connects this adaptation to the target rate. Knowing $L_0$, however, reduces the constant in our upper bound from 100 to 10.

Another natural approach is to apply existing bandit algorithms to one coordinate. We closely examine whether these can recover the optimal dependence on $L_0$ without knowing $L_0$. Their guarantees reveal two obstacles. For several EXP3 variants (Auer et al., 2002b; Audibert & Bubeck, 2010; Neu, 2015a), the resulting bounds retain a dependence on $T$, $L_d$, or additional factors, and therefore do not recover the rate. For Tsallis-INF (Zimmert & Seldin, 2021), EXP3++ (Seldin & Lugosi, 2017), and MO-US (Xu & Klabjan, 2023), the guarantees instead control signed expected scalar regret, whereas our reduction requires its positive part. Converting these guarantees introduces an extra sequence-dependent term that is not controlled by $L_0$. As an ablation of our doubling scheme, we also restart these algorithms independently over fixed blocks. This removes the term but still yields a $T^{3/4}$ dependence (Appendix A.7.7). We summarize all results in Table 1; full results are in Appendix. On the other hand, for a particular sequence, $L_d$ may yield a coordinate-dependent bound (Corollary A.21) through GREEN-IX (Lykouris et al., 2022). Identifying a favorable $d$, however, can itself be difficult with little knowledge of the problem instance, and a poor choice may yield a worse bound than one based on $L_0$. By contrast, $L_0$ reflects worst-case performance, a key departure herein.

Collectively, the lower and upper bounds yield the following minimax characterization.

**Theorem 4.6** (Minimax rate)**.** *For $K \geq 4, T \geq 6, D \geq 2$ and every $L_0 \in [0, T]$, Eq. (6) satisfies $\mathcal{V}_{K,D,T}(L_0) = \Theta(\min\{U_0, \sqrt{KU_0}\})$* (12).

| Method | Knows $L_0$? | $b$ in $\mathbb{E}[R_T^{\text{fs}}] \lesssim \min\{U_0, b\}$ |
|---|---|---|
| **Poly-INF** | **Yes** | $\sqrt{\mathbf{KU_0}}$ |
| Modified EXP3 | Yes | $\sqrt{KU_0 \log(3K)}$ |
| Gain-based EXP3 | Yes / no | $(K \log K)^{1/3} V^{2/3}$ with $V = U_0 \,/\, T$ |
| **Reward-doubling Poly-INF (Ours)** | **No** | $\sqrt{\mathbf{KU_0}}$ |
| Poly-INF with $B = T$ | No | $\sqrt{KT}$ |
| EXP3.P and adversarial MO-KS | Either | $\sqrt{KT \log(KT)}$ |
| EXP3-IX | Either | $\sqrt{KT \log K}$ |
| GREEN-IX | No | $\sqrt{KL_{d'} \log(KT)} + K \log(KT)$ |
| Original EXP3.1 | No | $\sqrt{KU_0 \log K} + (KU_0^3 / \log K)^{1/4} + K \log K$ |
| Tsallis-INF (IW) | No | $\sqrt{KT} + S_{d'}$ |
| Tsallis-INF (RV) | No | $\sqrt{KT} + K \log T + S_{d'}$ |
| EXP3++ | No | $\sqrt{KT \log K} + S_{d'}$ |
| MO-US | No | $\sqrt{KT \log K} + K + S_{d'}$ |
| Fixed-block Tsallis-INF (IW) | No | $K^{1/4} T^{3/4}$ |
| Fixed-block Tsallis-INF (RV) and MO-US | No | $K^{1/4} T^{3/4} \log(2T)$ |
| Fixed-block EXP3++ | No | $K^{1/4} T^{3/4} \sqrt{\log K}$ |

Table 1: Newly derived upper bounds with a coordinate $d'$ fixed before play.

At $L_0 = 0$, the minimax rate is $\Theta(W_{K,T})$ where $W_{K,T} := \min\{T, \sqrt{KT}\}$, recovering the existing bound and removing the extra $\sqrt{\log T}$ factor in Xu & Klabjan (2023). More generally, these bounds characterize the dependence of the minimax Pareto regret on $K$, $T$, and, for the first time, $L_0$, an intrinsic coordinate-level quantity that may or may not be known to the learner. The horizon enters only through $U_0 = T - L_0$. When $U_0 \leq K$, the rate is $\Theta(U_0)$, corresponding to the deterministic reward cap. When $U_0 \geq K$, the rate becomes $\Theta(\sqrt{KU_0})$, reflecting the cost of learning the best arm. For $K \in \{2, 3\}$, all upper bounds remain valid, while the matching lower bound remains open. Lastly, the role of the coordinate dimension is surprisingly limited. The rate is independent of $D$, and a coordinate is sufficient to attain the upper bound. Notably, the same dependence on $L_0$ persists under full information, where the learner observes every arm's reward (Appendix A.3.6). Broadly, we believe it may also apply to adaptive adversaries (Cesa-Bianchi et al., 2013) by replacing the bound.

*Computational study.* We also conduct numerical experiments to test the behavior suggested by theory. Across five settings, both proposed Poly-INF methods attain lower Pareto regret than all baselines, while the hardest instance differs across policies. Details are provided in Appendix B.

## 5 Conclusions and Future Work

Multi-objective bandits have been widely used or hold the potential to be widely used across various application domains. Understanding the role of coordinates in adversarial multi-objective bandits is particularly important, yet underexplored, for optimizing performance with the right handle. Pareto dominance seems to complicate the problem, but we have also been fortunately equipped with numerous methods from classical single-objective bandits. We take steps toward this herein by studying several aspects of coordinates in the context of Pareto regret optimization. Several findings are perhaps counterintuitive or surprising at first sight, but later turn out to be quite interpretable. The most crucial role of a coordinate-level quantity is reflected in the minimum coordinate loss $L_0$. It plays an important role in coordinate regret and, in theory, governs the best admissible coordinate regret. A surprising lower bound on Pareto regret is monotonically decreasing in $L_0$, yielding an instance-dependency through $L_0$. We also develop algorithms that achieve this rate uniformly over the class, proving both tightness and optimality for every $L_0$ and over instances in the minimax sense. In contrast, the roles of the coordinate dimension ($D$) and coordinate choice ($d$) are quite limited. The bound has no dependence on $D$ or $d$. Yet, the choice of algorithm along the chosen coordinate matters and must therefore be made carefully, perhaps echoing, in a way, a philosophy of life. At the same time, we feel obligated to highlight that the findings are limited to these specific intrinsic quantities and do not exclude the possibility that there may be other quantities that are also related to coordinates and affect Pareto regret. We hope the work proposed herein could inspire future work toward fully understanding the role of coordinates for performances in multi-objective problems.

## DISCLOSURE OF LLM USE

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


Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Algorithmic Learning Theory*, volume 4264 of *Lecture Notes in Computer Science*, pp. 229–243. Springer, 2006. doi: 10.1007/11894841_20.

Çağın Ararat and Cem Tekin. Vector optimization with stochastic bandit feedback. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2165–2190. PMLR, 2023. URL https://proceedings.mlr.press/v206/ararat23a.html.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(94):2785–2836, 2010. URL https://www.jmlr.org/papers/v11/audibert10a.html.

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

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b. doi: 10.1137/S0097539701398375.

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

Djallel Bouneffouf and Raphael Feraud. Multi-armed bandits meet large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39682–39690, 2026.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE congress on evolutionary computation (CEC)*, pp. 1–8. IEEE, 2020.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pp. 122–134. PMLR, 2013.

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. PMLR, 2017a.

Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. Multi-objective bandits: Optimizing the generalized Gini index. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 625–634. PMLR, 2017b. URL https://proceedings.mlr.press/v70/busa-fekete17a.html.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. *Advances in Neural Information Processing Systems*, 26, 2013.

Sanxing Chen, Xiaoyin Chen, Yukun Huang, Roy Xie, and Bhuwan Dhingra. When greedy wins: Emergent exploitation bias in meta-bandit llm training. In *International Conference on Learning Representations*, volume 2026, pp. 57389–57421, 2026.

Élise Crépon, Aurélien Garivier, and Wouter M. Koolen. Sequential learning of the Pareto front for multi-objective bandits. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3583–3591. PMLR, 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.

Aurélien Garivier, Wouter M Koolen, et al. Sequential learning of the pareto front for multi-objective bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3583–3591. PMLR, 2024.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems 29*, pp. 1198–1206, 2016.

Changkun Guan and Mengfan Xu. Are stochastic multi-objective bandits harder than single-objective bandits? *arXiv preprint arXiv:2604.07096*, 2026.

Alihan Hüyük and Cem Tekin. Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives. *Machine Learning*, 110(6):1233–1266, 2021. doi: 10.1007/s10994-021-05956-1.

Jiyan Jiang, Wenpeng Zhang, Shiji Zhou, Lihong Gu, Xiaodong Zeng, and Wenwu Zhu. Multi-objective online learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dKkMnCWfVmm.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pp. 1152–1161. PMLR, 2015.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit Pareto set identification: The fixed budget setting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2548–2556. PMLR, 2024. URL https://proceedings.mlr.press/v238/kone24a.html.

Arie W Kruglanski, Antonio Pierro, and Anna Sheveland. How many roads lead to rome? equifinality set-size and commitment to goals and means. *European Journal of Social Psychology*, 41(3):344–352, 2011.

Anisio Lacerda. Multi-objective ranked bandits for recommender systems. *Neurocomputing*, 246:12–24, 2017.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. doi: 10.1016/0196-8858(85)90002-8. URL https://doi.org/10.1016/0196-8858(85)90002-8.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Donghao Li, Chengshuai Shi, Weijuan Ou, Cong Shen, and Jing Yang. Efficient multi-objective prompt optimization via pure-exploration bandits. *ICLR*, 2026.

David Liau, Zhao Song, Eric Price, and Ger Yang. Stochastic multi-armed bandits in constant space. In *International Conference on Artificial Intelligence and Statistics*, pp. 386–394. PMLR, 2018.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 3080–3086. International Joint Conferences on Artificial Intelligence Organization, 2019. doi: 10.24963/ijcai.2019/427.

Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Small-loss bounds for online learning with partial information. *Mathematics of Operations Research*, 47(3):2186–2218, 2022. doi: 10.1287/moor.2021.1204. URL https://arxiv.org/abs/1711.03639.

Kanak Mahadik, Qingyun Wu, Shuai Li, and Amit Sabne. Fast distributed bandits for online recommendation systems. In *Proceedings of the 34th ACM international conference on supercomputing*, pp. 1–13, 2020.

James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 31–39, 2018.

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.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28, 2015a. URL https://proceedings.neurips.cc/paper/2015/hash/e5a4d6bf330f23a8707bb0d6001dfbe8-Abstract.html.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1360–1375. PMLR, 2015b. URL https://proceedings.mlr.press/v40/Neu15.html.

Adam Pažout, Tom Brughmans, Pau de Soto, Matteo Mazzamurro, María Coto-Sarmiento, Clara Filet, Eduardo Herrera Malatesta, Magnus L Nielsen, Gustav Emil Ølgaard, and Peter B Vahlstrup. Network science reveals the structure and lasting impact of the roman road system. *Nature communications*, 17(1):9551, 2026.

Manhin Poon, XiangXiang Dai, Xutong Liu, Fang Kong, John CS Lui, and Jinhang Zuo. Online multi-llm selection via contextual bandits under unstructured context evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24855–24863, 2026.

Diederik M Roijers, Luisa M Zintgraf, and Ann Nowé. Interactive thompson sampling for multi-objective multi-armed bandits. In *International conference on algorithmic decision theory*, pp. 18–34. Springer, 2017.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1743–1759. PMLR, 2017. URL https://proceedings.mlr.press/v65/seldin17a.html.

Eralp Turgay, Doruk Öner, and Cem Tekin. Multi-objective contextual bandit problem with similarity information. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1673–1681. PMLR, 2018. URL https://proceedings.mlr.press/v84/turgay18a.html.

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.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1263–1291. PMLR, 2018. URL https://proceedings.mlr.press/v75/wei18a.html.

Mengfan Xu and Diego Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38499–38517. PMLR, 2023. URL https://proceedings.mlr.press/v202/xu23i.html.

Bo Xue, Zhi Hong, Jiayi Li, Yuanyu Wan, Ji Cheng, and Shuang Qiu. Cost-aware multi-objective bandits: Theory and application to budgeted llm configuration evaluation. *arXiv preprint arXiv:2608.04333*, 2026.

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 (ADPRL)*, pp. 1–8. IEEE, 2014.

Liu Yang, Bo Liu, Leyu Lin, Feng Xia, Kai Chen, and Qiang Yang. Exploring clustering of bandits for online recommendation system. In *Proceedings of the 14th ACM conference on recommender systems*, pp. 120–129, 2020.

Xinyang Yi, Shao-Chuan Wang, Ruining He, Hariharan Chandrasekaran, Charles Wu, Lukasz Heldt, Lichan Hong, Minmin Chen, and Ed H Chi. Online matching: A real-time bandit system for large-scale recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 403–414, 2023.

Richard Zhang and Daniel Golovin. Random hypervolume scalarizations for provable multi-objective black box optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11096–11105. PMLR, 2020. URL https://proceedings.mlr.press/v119/zhang20i.html.

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.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021. URL https://jmlr.org/papers/v22/19-753.html.

Eckart Zitzler, Lothar Thiele, Marco Laumanns, Carlos M. Fonseca, and Viviane Grunert da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132, 2003. doi: 10.1109/TEVC.2003.810758.

# A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

The appendix is organized as follows.

1. Appendix A.1 reviews the closest stochastic and adversarial bandit results.
2. Appendix A.2 proves the identities and scalar comparisons in Section 2.
3. Appendix A.3 proves the bandit lower bounds in Section 3, the known-coordinate result cited there, and the full-information extension.
4. Appendix A.4 states detailed guarantees for the comparison algorithms.
5. Appendices A.5–A.8 give the algorithms and upper-bound proofs in the order used in Section 4.
6. Appendix A.9 gives coordinate-dependent refinements of the upper bounds.
7. Appendix B presents the computational study.

## A.1 RELATED WORK

### A.1.1 STOCHASTIC PARETO REGRET

Most work on stochastic multi-objective bandits assumes that each arm has a fixed probability distribution over reward vectors and compares arms through their mean reward vectors. Drugan & Nowe (2013) introduced Pareto UCB1, which forms coordinatewise confidence bounds, constructs an optimistic Pareto set, and selects an arm from that set. Its cumulative Pareto regret adds the Pareto suboptimality gap of the selected arm on each round. This is a round-by-round comparison with the Pareto front of the mean reward vectors.

Several extensions retain this stochastic, mean-based viewpoint. Turgay et al. (2018) define contextual Pareto regret as the sum of the distances from the selected context-arm pairs to the corresponding context-dependent Pareto fronts. Lu et al. (2019) study generalized linear reward models, use confidence bounds to approximate the Pareto front, and obtain a contextual Pareto-regret guarantee. Guan & Xu (2026) ask whether stochastic Pareto regret can follow an easy objective. Under a condition ensuring that an objective-wise maximizer is Pareto optimal, their algorithm searches across objectives and certifies a best arm on an objective with a large mean gap. Repeated samples from fixed reward distributions make such certification possible.

The stochastic regret above is different from the criterion studied here. It sums gaps defined from fixed mean reward vectors or, in contextual models, fixed conditional mean functions. Our criterion instead makes one comparison after $T$ rounds between the learner's cumulative reward vector and the cumulative reward vectors of the fixed arms. This distinction matters under adversarial rewards because there are no fixed arm means or mean gaps to estimate.

### A.1.2 OTHER MULTI-OBJECTIVE CRITERIA

Pareto order is one way to compare vector rewards without first choosing how to trade one objective against another. Other work makes that tradeoff part of the problem. Busa-Fekete et al. (2017b) optimize the Generalized Gini Index, which aggregates the coordinates according to a fairness-sensitive utility. Hüyük & Tekin (2021) study lexicographically ordered and satisficing objectives. Linear and nonlinear scalarizations likewise convert each reward vector into a scalar reward before applying a bandit algorithm (Drugan & Nowe, 2013). These criteria are appropriate when the coordinate preferences are specified, but two different preferences can rank the same Pareto-efficient arms differently.

Multi-objective optimization also evaluates sets of solutions through additive indicators and hypervolume. Additive indicators ask how far one set must be shifted before it weakly covers another, while hypervolume measures the region dominated relative to a chosen reference point (Zitzler et al., 2003; Zhang & Golovin, 2020). Our criterion uses a common additive shift, but it answers a narrower online-learning question. It asks whether one fixed arm improves the learner's cumulative reward in every coordinate. It does not evaluate the coverage or hypervolume of a set of solutions.

Another branch studies Pareto-set identification rather than cumulative regret. Auer et al. (2016) give confidence-based procedures for approximately identifying the Pareto front. Ararat & Tekin (2023) extend this problem to preferences represented by a polyhedral ordering cone and derive cone-dependent sample-complexity bounds. More recent work studies both fixed-budget identification (Kone et al., 2024) and fixed-confidence identification with asymptotically optimal sample complexity for Gaussian arms (Crépon et al., 2024). These methods seek to return the Pareto set, or an approximation to it, after sampling. Our learner instead incurs regret while acting and is compared with fixed arms over the given horizon.

### A.1.3 Adversarial multi-objective bandits and online learning

Jiang et al. (2023) introduce a sequence-wise Pareto suboptimality gap in multi-objective online convex optimization. Their sequence-wise perspective provides conceptual motivation for evaluating the learner's cumulative vector against fixed decisions in hindsight. Our setting uses finitely many arms and observes only the chosen arm's reward vector. We study an explicitly nonnegative fixed-arm comparison; its pathwise representation and bandit guarantees are established here, rather than inferred from online-convex-optimization guarantees.

The closest bandit work is Xu & Klabjan (2023). They define Pareto regret through an equal-shift comparison between the learner's cumulative reward vector and an arm-based Pareto front, and relate it to scalar regret on a coordinate chosen before play. Their MO-KS algorithm runs UCB or EXP3.P on one fixed coordinate when it is told whether rewards are stochastic or adversarial. Their MO-US algorithm modifies EXP3++ to avoid requiring this information. They also give worst-case lower bounds, including an adversarial lower bound based on repeated coordinates.

Our criterion admits a uniform-shift representation using distinct nondominated cumulative arm vectors and an infimum. Proposition A.1 relates it precisely to the incomparability-based criterion, including cases involving ties or an unattained minimum. For the nonnegative criterion studied here, scalar comparisons require the positive part of coordinate regret, not just its signed expectation.

Our lower bounds concern reward sequences with prescribed minimum coordinate loss $L^\star = L_0$. The worst-case order persists at $L_0 = 0$, even though some arm has zero loss in one coordinate, and the complete rate depends on $K$ and $U_0 = T - L_0$. Corollary 3.3 further preserves this rate for an explicit family of prescribed unordered coordinate-loss profiles. These results distinguish knowledge of coordinate-loss values from knowledge of which coordinates attain them.

### A.1.4 Scalar adversarial bandits

The standard adversarial bandit benchmark is regret against the best fixed arm in hindsight. EXP3 and EXP3.P give expected and high-probability guarantees at the usual worst-case scale (Auer et al., 2002b). A separate line of work obtains first-order bounds that become smaller when the best arm's cumulative loss is small (Allenberg et al., 2006; Neu, 2015b; Lykouris et al., 2022; Wei & Luo, 2018). Lower bounds match this dependence up to logarithmic factors in the applicable small-loss regime (Gerchinovitz & Lattimore, 2016). These results motivate our question of whether a small loss in one unknown coordinate also reduces Pareto regret.

The upper bounds in this paper also use scalar guarantees that depend on the best fixed arm's cumulative reward. Poly-INF provides the bound used for our matching result (Audibert & Bubeck, 2010). We additionally analyze EXP3.P, EXP3-IX (Neu, 2015a), Tsallis-INF (Zimmert & Seldin, 2021), EXP3++ (Seldin & Lugosi, 2017), and the scalar algorithms underlying MO-KS and MO-US. This comparison requires some care. A bound on signed expected scalar regret does not by itself bound the expected positive part of scalar regret, which controls our Pareto regret. We therefore state whether each result uses the published algorithm unchanged, uses parameters determined by $U_0$, or changes the reward estimator. The modified EXP3 estimator and the Poly-INF policy are existing scalar constructions; our contribution is their analysis under the Pareto criterion and the matching reward-doubling procedure when $L_0$ is unknown.

With full-information feedback, the scalar comparison becomes the classical prediction-with-expert-advice problem, whose worst-case arm dependence is $\sqrt{\log K}$ rather than $\sqrt{K}$ (Cesa-Bianchi & Lugosi, 2006). Our full-information result shows that this stronger feedback changes the dependence

on the number of arms but does not make an unidentified zero-loss coordinate sufficient for bounded regret.

### A.2 Problem-formulation properties

We prove the formula defining $R_T^{\text{fs}}$, its relation to the original distance, and the comparisons with scalar regret, in that order.

#### A.2.1 The formula for Pareto regret

*Proof of Proposition 2.1.* For a fixed cumulative reward vector $\sigma$, put $m_\sigma = \min_d(\sigma_d - A_{T,d})$, and let $M = \max_{\sigma \in O'} m_\sigma$, which exists because $O'$ is finite and nonempty. If $0 \le \epsilon < M$, choose $\sigma \in O'$ with $m_\sigma = M$. Then $\sigma_d > A_{T,d} + \epsilon$ for every $d$, so this value of $\epsilon$ is not feasible in the infimum. If $\epsilon > \max\{0, M\}$, then for every $\sigma \in O'$ there is a coordinate $d$ with $\sigma_d - A_{T,d} \le M < \epsilon$. Hence $\sigma$ does not weakly dominate $A_T + \epsilon\mathbf{1}$, and $\epsilon$ is feasible. These two implications show that the infimum is $\max\{0, M\} = [M]_+$; whether the boundary itself is feasible does not change its value.

It remains to replace the maximum over $O'$ by a maximum over arms. For every arm $i$, either $G_i \in O'$, or repeatedly following a dominating arm reaches some $\sigma \in O'$ with $\sigma \succeq G_i$, because there are only finitely many distinct arm vectors. Therefore

$$\min_d(G_{i,d} - A_{T,d}) \le \min_d(\sigma_d - A_{T,d}).$$

The reverse inequality between the two maxima holds because every vector in $O'$ is an arm vector. Thus $M = \max_i \min_d(G_{i,d} - A_{T,d})$, proving the identity on every outcome, including those for which the value is zero. □

#### A.2.2 Relation to the original distance

We compare (3) with the uniform-shift incomparability criterion of Xu & Klabjan (2023). Write their front as

$$F_X := \{G_i : i \in [K],\ \nexists j \ne i \text{ with } G_j \succeq G_i\}.$$

This definition excludes an arm whenever another arm weakly dominates it, including when the two arms have the same cumulative reward vector. For example, two arms with cumulative vector $(1, 1)$ are both excluded, leaving $F_X$ empty. Our vector front $O'$ keeps one copy of this vector and is always nonempty.

When $A_T \preceq \sigma$ for every $\sigma \in F_X$, their criterion uses the smallest $\epsilon \ge 0$ for which $A_T + \epsilon\mathbf{1} \parallel \sigma$ for every $\sigma \in F_X$. Strict incomparability can leave this minimum undefined. With $A_T = (0, 0)$ and $F_X = \{(1, 1)\}$, no shift is feasible. Replacing the front by $F_X = \{(1, 2)\}$ gives the open interval $(1, 2)$, which has infimum one but no smallest element. For nonempty $F_X$, define

$$a_- := \max_{\sigma \in F_X} \min_d(\sigma_d - A_{T,d}), \qquad a_+ := \min_{\sigma \in F_X} \max_d(\sigma_d - A_{T,d}). \tag{13}$$

**Proposition A.1** (Comparison)**.** *Suppose $F_X \ne \varnothing$ and $A_T \preceq \sigma$ for every $\sigma \in F_X$. The feasible shifts in the incomparability condition of Xu & Klabjan (2023) form $(a_-, a_+)$. If this interval is nonempty, its infimum is $a_- \le R_T^{\text{fs}}$, and its minimum is not attained. If the cumulative arm vectors are distinct, then $F_X = O'$ and the infimum equals $R_T^{\text{fs}}$. Proposition 2.1 continues to apply when $F_X$ or the feasible interval is empty.*

*Proof of Proposition A.1.* A vector $A_T + \epsilon\mathbf{1}$ is incomparable with $\sigma$ precisely when at least one of its coordinates is strictly smaller and another is strictly larger. Equivalently,

$$\min_d(\sigma_d - A_{T,d}) < \epsilon < \max_d(\sigma_d - A_{T,d}).$$

Intersecting these open intervals over $F_X$ gives $(a_-, a_+)$. Under the proposition's assumptions, $a_- \ge 0$, so requiring $\epsilon \ge 0$ leaves this interval unchanged. If the interval is nonempty, its infimum is $a_-$ and it has no minimum.

Every vector retained in $F_X$ belongs to $O'$, although a vector shared by tied arms can be absent from $F_X$. Hence $F_X \subseteq O'$ and

$$a_- \le \max_{\sigma \in O'} \min_d(\sigma_d - A_{T,d}).$$

Together with $a_- \geq 0$ and Proposition 2.1, this proves $a_- \leq R_T^{\mathrm{fs}}$. If all cumulative arm vectors are distinct, both fronts contain the same vectors, giving $a_- = R_T^{\mathrm{fs}}$. □

### A.2.3 SCALAR COMPARISONS

*Proof of Lemma 4.1.* For $x_{i,d} = G_{i,d} - A_{T,d}$ and any $w \in \Delta_D$,

$$\max_i \min_d x_{i,d} \leq \max_i \sum_d w_d x_{i,d} = R_T^w.$$

Taking positive parts proves the weighted comparison. Taking $w$ supported on coordinate $d'$ gives the coordinate comparison. Finally, all learner rewards are nonnegative, and $G_{i,d} \leq U^\star$ for every $i, d$. Hence $R_T^{\mathrm{fs}} \leq U^\star$ and $\max_i \sum_d w_d G_{i,d} \leq U^\star$.

When a scalar algorithm is used later, its weights are fixed before play. Its scalar observations then lie in $[0, 1]$, use only the selected arm's reward vector, and form a reward sequence fixed before play. The same proof applies after conditioning on weights sampled independently before play. □

## A.3 LOWER-BOUND ANALYSIS

We first establish a scalar bandit estimate, then use a positive-probability version of it to compare the algorithm with one arm in both coordinates.

### A.3.1 A SCALAR BANDIT ESTIMATE

**Lemma A.2** (Hard bandit suffix)**.** *Let $m \geq 1$ and $q \geq 2$ be integers, and put*

$$\alpha := \frac{1}{8} \min\left\{1, \sqrt{q/m}\right\},$$

*the advantage in mean reward that the hidden candidate will carry. Consider $m$ bandit-feedback rounds with $q$ candidate actions and any number of additional actions. The causal learner may begin with an arbitrary random state $Z$, which models whatever it inherits from earlier rounds. Draw $I \sim \mathrm{Unif}[q]$ independently of $Z$. For each $i \in [q]$, let $P_i$ denote the probability distribution of the complete interaction conditional on $I = i$. Let $P_0$ denote the corresponding null distribution, in which no candidate is advantaged. Assume the following.*

1. *The distribution of $Z$ is the same under $P_0, P_1, \ldots, P_q$. Under $P_i$, candidate rewards are independent of $Z$, mutually independent across candidates and rounds, candidate $i$ has Bernoulli mean $1/2 + \alpha$, and every other candidate is fair, that is, Bernoulli with mean $1/2$. Under $P_0$, every candidate is fair.*
2. *Under each fixed distribution $P_i$ or $P_0$, if $\mathcal{F}_{t-1}$ is the complete history before suffix round $t$, including $Z$, then the current candidate-reward vector is independent of $\mathcal{F}_{t-1}$ and of the current action, with the distribution stated in item 1. Each additional action has conditional expected scalar reward at most $1/2$, given $\mathcal{F}_{t-1}$ and its selection.*
3. *Feedback is nonanticipating and contains no unselected current or future candidate reward. For each fixed $i$, the conditional distributions of the observation given the history and selected action coincide under $P_i$ and $P_0$ when the selected action is not $i$. When action $i$ is selected, the observation contains its scalar Bernoulli reward and, conditional on that reward, all remaining feedback has the same conditional distribution under $P_i$ and $P_0$.*

*Under these assumptions, a round on which $i$ is selected contributes exactly its Bernoulli likelihood ratio to the ratio between $P_i$ and $P_0$, and the arbitrary inherited state contributes no likelihood ratio.*

*If $S$ is the learner's cumulative suffix reward and $C_i$ is candidate $i$'s cumulative suffix reward, then every causal learner satisfies*

$$\mathbb{E}\left[\max_{i \in [q]} C_i - S\right] \geq \frac{1}{32} \min\{m, \sqrt{qm}\}, \tag{14}$$

*where the expectation is over the hidden arm, the Bernoulli rewards, the remaining feedback, and the learner's randomization.*

*Proof.* The proof compares every alternative $P_i$ with the null distribution $P_0$, uses relative entropy and Pinsker's inequality, which bounds the total variation distance between two probability distributions by the square root of half their relative entropy, to limit how often the hidden arm can be sampled, and charges $\alpha$ expected reward on every missed pull.

Write $N_i$ for the number of pulls of candidate $i$ and $n_i = \mathbb{E}_0[N_i]$. By the feedback-kernel assumption, item 3, a round contributes nothing to the relative entropy between $P_0$ and $P_i$ unless $i$ is selected, and when $i$ is selected it contributes the divergence $d_0$ between a fair coin and a coin with mean $1/2 + \alpha$. The chain rule for relative entropy, which sums these conditional contributions round by round, therefore gives

$$\mathrm{KL}(P_0 \| P_i) = n_i d_0,$$
$$d_0 := \mathrm{kl}(1/2 \| 1/2 + \alpha) = -\frac{1}{2}\log(1 - 4\alpha^2) \le \frac{32}{15}\alpha^2.$$

For the last inequality, put $x = 4\alpha^2 \le 1/16$ and use $-\log(1-x) \le x/(1-x)$, which gives $d_0 \le 2\alpha^2/(1-4\alpha^2) \le (32/15)\alpha^2$.

Since $0 \le N_i \le m$, replacing $P_0$ by $P_i$ changes the expectation of $N_i$ by at most $m$ times their total variation distance, and Pinsker's inequality bounds that distance by $\sqrt{\mathrm{KL}(P_0\|P_i)/2}$, so

$$\mathbb{E}_i[N_i] \le \mathbb{E}_0[N_i] + m \, \|P_i - P_0\|_{\mathrm{TV}}$$
$$\le n_i + m\sqrt{\frac{\mathrm{KL}(P_0\|P_i)}{2}} = n_i + m\sqrt{\frac{d_0 n_i}{2}}.$$

Because $\sum_i n_i \le m$, averaging over $i$ and applying Cauchy–Schwarz in the form $\sum_i \sqrt{n_i} \le \sqrt{q \sum_i n_i}$ yield

$$\frac{1}{q}\sum_{i=1}^{q} \mathbb{E}_i[N_i] \le \frac{m}{q} + \frac{m}{q}\sqrt{\frac{d_0}{2}} \sum_{i=1}^{q} \sqrt{n_i}$$
$$\le \frac{m}{q} + m\sqrt{\frac{d_0 m}{2q}} \le \frac{m}{q} + \frac{m}{2\sqrt{15}}. \tag{15}$$

For the last step, if $q \le m$, then

$$\sqrt{\frac{d_0 m}{2q}} \le \sqrt{\frac{16\alpha^2 m}{15q}} = \frac{1}{2\sqrt{15}}.$$

If $q > m$, then $\alpha = 1/8$, and the same left-hand side is at most $\sqrt{m/(60q)} < 1/(2\sqrt{15})$.

Under $P_i$, a round on which $i$ is not pulled gives the learner a reward of conditional expectation at most $1/2$, by items 1 and 2, while the unobserved reward of candidate $i$ is independent with mean $1/2 + \alpha$. On a round when $i$ is pulled, the learner and the comparator collect the same realized reward. Consequently

$$\mathbb{E}_i[C_i - S] \ge \alpha(m - \mathbb{E}_i[N_i]).$$

Averaging, using (15), and noting $q \ge 2$,

$$\frac{1}{q}\sum_i \mathbb{E}_i[C_i - S] \ge \alpha m \left(1 - \frac{1}{q} - \frac{1}{2\sqrt{15}}\right)$$
$$\ge \frac{1}{32}\min\{m, \sqrt{qm}\}.$$

Indeed, the parenthesized factor is at least $1/2 - 1/(2\sqrt{15}) > 1/4$, while $\alpha m = \min\{m, \sqrt{qm}\}/8$. Finally $\max_i C_i - S \ge C_I - S$ on every play path, and the left side of the previous display is $\mathbb{E}[C_I - S]$ because $I$ is uniform. This proves (14). □

**Lemma A.3** (A positive scalar shortfall with fixed probability)**.** *In Lemma A.2, suppose that every additional action has deterministic scalar reward $1/2$. Put $V = \min\{m, \sqrt{qm}\}$ and $Z = C_I - S$. Then, for every inherited state independent of the suffix experiment,*

$$\mathbb{E}[Z] \ge V/32, \qquad \mathbb{E}[Z^2] \le 3V^2, \qquad \mathbb{P}\{Z \ge V/64\} \ge 1/12288.$$

*Proof.* The proof of Lemma A.2 gives the stated expectation for the hidden arm $I$ itself. Let $X_t$ be candidate $I$'s reward minus the learner's reward on suffix round $t$. Conditional on $I$ and the history before that round, $\mathbb{E}[X_t] = \alpha \mathbf{1}\{a_t \neq I\}$. Therefore

$$M := \sum_{t=1}^{m} \big(X_t - \alpha \mathbf{1}\{a_t \neq I\}\big)$$

is a martingale with $M_0 = 0$. Each difference $X_t$ lies in $[-1, 1]$, so each martingale increment has conditional variance at most one. Martingale orthogonality then gives $\mathbb{E}[M^2] \leq m$, and

$$Z = \alpha(m - N_I) + M, \qquad \mathbb{E}[M^2] \leq m.$$

In both cases $q \leq m$ and $q > m$, the definition of $\alpha$ gives $\alpha m = V/8$. Also $V^2 \geq m$ because $q \geq 2$, and $0 \leq N_I \leq m$. Hence

$$\mathbb{E}[Z^2] \leq 2\alpha^2 m^2 + 2\mathbb{E}[M^2] \leq V^2/32 + 2m \leq 3V^2.$$

Writing $p = \mathbb{P}\{Z \geq V/64\}$, split the positive part of $Z$ at $V/64$ and apply Cauchy–Schwarz on the upper event. This gives

$$V/32 \leq \mathbb{E}[Z] \leq \mathbb{E}[[Z]_+] \leq V/64 + \sqrt{\mathbb{E}[Z^2], p} \leq V/64 + \sqrt{3}\, V \sqrt{p}.$$

Subtracting $V/64$, dividing by $\sqrt{3}V$, and squaring proves $p \geq 1/(64^2 \cdot 3) = 1/12288$. All calculations hold for an arbitrary inherited state, so they also hold conditional on any prefix history independent of the suffix experiment. □

### A.3.2 Proof of Theorem 3.1

*Proof of Theorem 3.1.* We prove a slightly stronger statement covering every integer $T \geq 1$. Let $q = \lfloor K/2 \rfloor \geq 2$ and $W = \min\{T, \sqrt{qT}\}$.

First suppose $W \geq 512$, and set

$$h = \lfloor W/256 \rfloor, \qquad n = T - h, \qquad \alpha = \tfrac{1}{8} \min\{1, \sqrt{q/n}\}.$$

Divide $2q$ arms into groups $A_1, \ldots, A_q$ and $B_1, \ldots, B_q$. If $K$ is odd, there is one additional arm $E$. For the first $h$ rounds their rewards are respectively $(1, 0), (0, 1), (0, 0)$.

Before play, choose a case $w \in \{A, B\}$ uniformly and an independent index $I \in [q]$ uniformly. Generate independent Bernoulli variables $Y_{i,t}$ for the remaining $n$ rounds. Their means are $1/2 + \alpha$ when $i = I$ and $1/2$ otherwise. The later rewards are

| | $A_i$ | $B_i$ | $E$ |
|---|---|---|---|
| $w = A$ | $(1, Y_{i,t})$ | $(0, 1/2)$ | $(0, 1/2)$ |
| $w = B$ | $(1/2, 0)$ | $(Y_{i,t}, 1)$ | $(1/2, 0)$. |

Omit column $E$ when $K = 2q$. All rewards are chosen before the algorithm acts. In case $A$, every $A_i$ receives reward one in coordinate 1 throughout; in case $B$, every $B_i$ does so in coordinate 2. Thus $L^\star = 0$ for every reward sequence in this finite distribution.

The algorithm cannot distinguish the cases during the first $h$ rounds. Condition on its history through those rounds, and let $X_A, X_B, X_E$ be its group pull counts. They sum to $h$. Its initial deficits in the two possible zero-loss coordinates are

$$d_A = h - X_A, \qquad d_B = h - X_B, \qquad d_A + d_B = h + X_E \geq h.$$

The case is still uniform and independent of this history. Therefore $d_w \geq h/2$ with conditional probability at least $1/2$.

We may now reveal the case to the algorithm, which only makes the problem easier. In case $A$, use coordinate 2 as the scalar suffix problem, with the $A_i$ as candidates. In case $B$, use coordinate 1, with the $B_i$ as candidates. The hidden index is independent of the inherited history. The other observed coordinate is deterministic given the group and case, so supplies no extra information about that index. Every noncandidate has scalar reward $1/2$. Lemma A.3 applies. If $C_I$ is the hidden arm's scalar reward and $S$ the algorithm's scalar reward in these $n$ rounds, then, with $V = \min\{n, \sqrt{qn}\}$,

$$\mathbb{P}\{C_I - S \geq V/64 \mid \text{prefix history}, w\} \geq 1/12288.$$

We next verify the four bounds on the chosen lengths. Since $W \geq 512$, we have $W/256 \geq 2$, and hence $h = \lfloor W/256 \rfloor \geq W/512$. Also $h \leq W/256 \leq T/256$, so $n = T - h \geq 255T/256$. Therefore

$$n \geq \frac{255}{256}T, \qquad \sqrt{qn} \geq \sqrt{\frac{255}{256}}\sqrt{qT} \geq \frac{255}{256}\sqrt{qT}.$$

Taking the minimum of the two left-hand terms gives $V \geq (255/256)W$. Finally,

$$\frac{V}{64} \geq \frac{255W}{16384} \geq \frac{2W}{256} \geq 2h.$$

Conditional on every prefix history and the selected case, the suffix event has probability at least $1/12288$. The prefix event has probability at least $1/2$, so their intersection has probability at least $1/24576$. On this intersection, the same hidden arm has cumulative advantage at least $h/2$ in its zero-loss coordinate. That advantage cannot decrease after the prefix, because the arm then receives one on every round. In the other coordinate, its prefix reward is zero and the algorithm's prefix reward is at most $h$, so its final advantage is at least $C_I - S - h \geq h$. Consequently $R_T^{\mathrm{fs}} \geq h/2$ on this event, and

$$\mathbb{E}[R_T^{\mathrm{fs}}] \geq \frac{h}{49152} \geq \frac{W}{25165824}.$$

For $W < 512$, use a simpler finite distribution. Choose an arm uniformly before play, give it reward $(1, 1)$ on round 1 and all other arms $(0, 0)$, and give every arm $(1, 1)$ on all later rounds. The easiest-coordinate loss is zero. The regret is one if the algorithm misses the hidden arm on the first round, so its expectation is $1 - 1/K \geq 3/4$.

For $K \geq 4$, $q = \lfloor K/2 \rfloor \geq K/4$, and therefore

$$W = \min\{T, \sqrt{qT}\} \geq \frac{1}{2}\min\{T, \sqrt{KT}\}.$$

The large-$W$ calculation gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \geq \frac{W}{25165824} \geq \frac{\min\{T, \sqrt{KT}\}}{50331648} \geq 2^{-26}\min\{T, \sqrt{KT}\}.$$

The small-$W$ construction also satisfies this bound because $W < 512$ implies $\min\{T, \sqrt{KT}\} < 1024$, while $3/4 > 2^{-26} \cdot 1024$. For $D > 2$, duplicate coordinate 1. This changes neither the regret nor the information available under bandit feedback.

Finally fix any randomized algorithm, including one told $L^\star = 0$. The average expected regret over our finite distribution has the stated lower bound. At least one deterministic reward sequence in its support therefore has at least that much expected regret, with expectation only over the algorithm's randomness. The support consists entirely of oblivious zero-loss sequences. This gives the claimed quantifiers. The factor lost when the construction is shortened below changes this constant from $2^{-26}$ to $2^{-27}$. □

### A.3.3 When the zero-loss coordinate is known

The lower bound hides which coordinate has zero loss. The next result shows why this uncertainty matters.

**Lemma A.4** (A known zero-loss coordinate)**.** *Let $K, T, D \geq 1$. Suppose a coordinate $d$ with $L_d = 0$ is specified before play, but the arm attaining that loss is unknown. Start with all arms available. On each round, play the first available arm in a fixed ordering, and remove it if its observed reward in coordinate $d$ is less than one. Then*

$$R_T^{\mathrm{fs}} \leq \min\{T, K - 1\}.$$

*When $D = 1$ and $L^\star = 0$, the same conclusion holds without any additional information.*

*Proof.* Because rewards lie in $[0, 1]$, the condition $L_d = 0$ means that some fixed arm receives reward one in coordinate $d$ on every round. This arm is never removed, so the algorithm always has an available arm. Whenever the algorithm receives a reward smaller than one in coordinate $d$, it removes

the arm it just played. No arm can be removed twice, and the zero-loss arm is never removed. There are therefore at most $K-1$ rounds with positive loss in coordinate $d$. Each such loss is at most one, so the algorithm's cumulative loss in that coordinate is at most $\min\{T, K-1\}$.

The best fixed arm receives cumulative reward $T$ in coordinate $d$. Consequently, $R_T^d$ equals the algorithm's cumulative loss in that coordinate. Lemma 4.1 now gives

$$R_T^{\mathrm{fs}} \le [R_T^d]_+ = R_T^d \le \min\{T, K-1\}.$$

If $D = 1$, the only coordinate is known without being supplied separately. □

### A.3.4 Proof of Theorem 3.2

The preceding construction gives zero loss in one coordinate. We now adjust its length and append common rewards to obtain every specified value of $L^\star$.

*Proof of Theorem 3.2.* Fix $U_0 = T - L_0$. If $U_0 = 0$, the claimed lower bound is zero. If $0 < U_0 < 1$, choose a hidden arm uniformly, give it reward $U_0\mathbf{1}$ on round 1, and give all other arm-round pairs reward zero. Every sequence has maximum arm-coordinate cumulative reward $U_0$. The first action misses the hidden arm with probability $1 - 1/K$, giving expected regret $(1-1/K)U_0 \ge 3U_0/4$. Since $U_0 < K$, $\min\{U_0, \sqrt{KU_0}\} = U_0$, so this proves the required bound in this case.

For $U_0 \ge 1$, let $m = \lfloor U_0 \rfloor \ge 1$. Use the zero-loss construction from the proof of Theorem 3.1 for $m$ rounds. If $m < T$, give every arm in every coordinate the same reward $U_0 - m$ on round $m+1$, and zero on all later rounds. If $m = T$, no padding is needed, since then $U_0 = T$. The largest cumulative arm-coordinate reward is exactly $m + (U_0 - m) = U_0$. The added rounds contribute the same reward to every comparator and the algorithm, so every coordinate advantage and $R_T^{\mathrm{fs}}$ are unchanged.

To apply the shorter-horizon bound, simulate the first $m$ actions of any proposed $T$-round algorithm, with its supplied values $T, L_0$ and its arbitrary randomization. This is an admissible $m$-round algorithm. The zero-loss proof covers every $m \ge 1$, not only the horizon range stated in the main theorem. Since $U_0 \ge 1$, $m = \lfloor U_0 \rfloor \ge U_0/2$. Thus $\sqrt{Km} \ge \sqrt{KU_0/2} \ge \frac{1}{2}\sqrt{KU_0}$, and

$$\min\{m, \sqrt{Km}\} \ge \tfrac{1}{2}\min\{U_0, \sqrt{KU_0}\}.$$

Multiplying the zero-loss constant $2^{-26}$ by the factor $1/2$ in the preceding display gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \ge 2^{-27}\min\{U_0, \sqrt{KU_0}\}.$$

As before, averaging over a finite distribution yields a deterministic sequence for each algorithm. Knowing $L_0$ reveals neither the hidden case nor the hidden arm. □

### A.3.5 Proof for a prescribed unordered loss profile

*Proof of Corollary 3.3.* Fix $U_0 = T - L_0$. If $U_0 = 0$, the all-zero sequence has profile $(T, \dots, T)$ and regret zero, as required. Assume first $0 < U_0 < 2$. Before play, choose an arm $I$ uniformly and give it reward $(U_0/2)\mathbf{1}$ on the first round, while every other arm receives zero. Choose a coordinate $d^\star$ independently. On every later round, give every arm reward $U_0/(2(T-1))$ in coordinate $d^\star$ and zero in every other coordinate. All rewards lie in $[0,1]$. The best arm's cumulative reward is $U_0$ in coordinate $d^\star$ and $U_0/2$ in every other coordinate, so the unordered loss profile is exactly $\lambda(L_0)$. Every arm receives the same rewards after the first round, so the final regret is $U_0/2$ if the learner does not choose $I$ on the first round, and zero otherwise. Its expectation is therefore $(1-1/K)U_0/2 \ge 3U_0/8$. Since $U_0 < 2 \le K$, this proves the lower bound for this case.

Now suppose $U_0 \ge 2$, and set $m = \lfloor U_0/2 \rfloor \ge 1$, $N = T - m$. Use the two-coordinate finite distribution in the proof of Theorem 3.1 for the first $m$ rounds. That proof covers every integer horizon $m \ge 1$ and yields expected regret at least $2^{-26}\min\{m, \sqrt{Km}\}$ against any policy. For $D > 2$, set each additional reward coordinate to the average of the first two during these rounds. The learner can compute these averages from the first two coordinates, so the added coordinates give no new information. For each fixed arm, its reward advantage over the learner on an added coordinate is the average of its advantages on the first two coordinates. This average is at least their minimum, so adding these coordinates leaves the Pareto regret over the first $m$ rounds unchanged.

Generate this entire prefix before play, and write

$$M_d := \max_i \sum_{t=1}^{m} r_d^{i,t}.$$

There is a coordinate $d^\star \in \{1, 2\}$ with $M_{d^\star} = m$. When both coordinates satisfy this equality, choose coordinate 1. On each of the last $N$ rounds, give every arm the same reward in coordinate $d$, equal to

$$b_d := \begin{cases} (U_0 - m)/N, & d = d^\star, \\ (U_0/2 - M_d)/N, & d \neq d^\star. \end{cases}$$

The inequalities $0 \le M_d \le m \le U_0/2$, $U_0 - m \le T - m = N$, and $U_0/2 \le T - m = N$ imply $0 \le b_d \le 1$ for all $d$, so the appended rewards are valid. The final best-arm cumulative reward is exactly $U_0$ on $d^\star$ and $U_0/2$ elsewhere. Every realization therefore belongs to $\mathcal{C}(\lambda(L_0))$.

We choose the entire reward sequence before play, including the rewards added after round $m$, so the adversary is oblivious. In coordinate $d$, these later rounds add the same amount $Nb_d$ to the learner's reward and to every arm's reward. Each arm's advantage therefore stays the same on every coordinate, and the final Pareto regret equals the regret over the first $m$ rounds.

Any $T$-round policy that knows the unordered profile can also be run for just the first $m$ rounds. Knowing $T, L_0, \lambda(L_0)$ gives the learner no information about the hidden case, the hidden arm, or the generated rewards, because these values are fixed independently of the construction. The zero-loss lower bound therefore applies to these first $m$ rounds. Since $m = \lfloor U_0/2 \rfloor \ge U_0/4$,

$$\min\{m, \sqrt{Km}\} \ge \frac{1}{4} \min\{U_0, \sqrt{KU_0}\}.$$

The expected regret over the finite construction is at least $2^{-28} \min\{U_0, \sqrt{KU_0}\}$. For every randomized policy, at least one deterministic sequence in this finite construction gives at least this expected regret, with expectation over the policy's randomness. This proves the minimax lower bound. The upper bound follows from Theorem 4.3, since the learner can compute $L_0 = \min_d \lambda_d$ from the profile and every sequence in the profile class has $L^\star = L_0$. □

### A.3.6 Full-information extension

Let $\mathcal{V}^{\mathrm{fi}}_{K,D,T}(L_0)$ denote the value in (6) when the algorithm observes every arm's reward vector after each round. As in (6), the algorithm may know $L_0$ before play.

**Theorem A.5** (Full-information minimax value). *For $K \ge 4$, $T \ge 6$, $D \ge 2$, and every $L_0 \in [0, T]$, put $U_0 = T - L_0$. Then*

$$\mathcal{V}^{\mathrm{fi}}_{K,D,T}(L_0) = \Theta\Big(\min\{U_0, \sqrt{U_0 \log K}\}\Big).$$

We prove the zero-loss lower bound first and then obtain every value of $L_0$ by shortening and padding the construction. The matching upper bound uses exponential weights on one coordinate fixed before play.

**Lemma A.6** (Zero-loss lower bound with full information). *For $K \ge 4$, $T \ge 1$, and $D \ge 2$, every full-information algorithm has a deterministic reward sequence with $L^\star = 0$ such that*

$$\mathbb{E}[R_T^{\mathrm{fs}}] \ge 2^{-17} \min\{T, \sqrt{T \log K}\}.$$

*The algorithm may know that $L^\star = 0$ before play.*

*Full-information lower bound when $L_0 = 0$.* We again prove the zero-loss statement for every integer $T \ge 1$. Put $q = \lfloor K/2 \rfloor$, $b = \lfloor \log_2 q \rfloor \ge 1$, and $W_f = \sqrt{T \min\{T, b\}}$. If $W_f \ge 128$, let

$$h = \lfloor W_f/64 \rfloor, \quad n = T - h, \quad \ell = \min\{n, b\}, \quad \tau = \lfloor n/\ell \rfloor, \quad V = \sqrt{\ell n}.$$

Split $2q$ arms into groups $A_1, \ldots, A_q$ and $B_1, \ldots, B_q$, with one unused arm $E$ when $K$ is odd. During the first $h$ rounds, give every $A_i$ reward $(1, 0)$, every $B_i$ reward $(0, 1)$, and $E$ reward $(0, 0)$. Independently choose a case $w \in \{A, B\}$ uniformly before play. In case $A$, coordinate 1 will be the

zero-loss coordinate of group $A$; in case $B$, coordinate 2 will be the zero-loss coordinate of group $B$. The prefix reward matrices are identical in the two cases, so full information does not reveal $w$. If $X_A, X_B, X_E$ are the algorithm's prefix pull counts, then its deficits in the two possible zero-loss coordinates are

$$d_A = h - X_A, \qquad d_B = h - X_B, \qquad d_A + d_B = h + X_E \geq h.$$

Conditional on the prefix history, the uniform case therefore satisfies $d_w \geq h/2$ with probability at least $1/2$. We reveal $w$ after the prefix, which can only help the algorithm.

Within the group that keeps the zero-loss coordinate, label $2^\ell \leq q$ candidate arms by all sign vectors in $\{-1, 1\}^\ell$. Duplicate any labels to fill the remaining positions in this group. Partition $\ell\tau$ suffix rounds into $\ell$ blocks of length $\tau$. On block $j$, let candidate $i$'s other coordinate be

$$Y_{i,t} = (1 + \sigma_j(i)\epsilon_t)/2,$$

where the $\epsilon_t$ are independent uniform signs generated before play. For the remaining suffix rounds use $Y_{i,t} = 1/2$. The zero-loss coordinate is one for all candidates and zero outside their group. Every noncandidate has scalar reward $1/2$ in the other coordinate. Thus every support sequence still has $L^\star = 0$.

Write $S_j = \sum_{t \text{ in block } j} \epsilon_t$. In each block $j$, choose the sign $\sigma_j$ to agree with the sign of $S_j$, using either sign when $S_j = 0$. One candidate has this entire sign vector, so the best candidate's scalar suffix reward is $n/2 + Q$, where

$$Q = \tfrac{1}{2} \sum_{j=1}^{\ell} |S_j|.$$

No other candidate can obtain more in any block, so this maximum is attained. For a suffix round $t$ in block $j$, define

$$\xi_t := \begin{cases} \sigma_j(i)\epsilon_t/2, & \text{if the algorithm selects candidate } i, \\ 0, & \text{if it selects a noncandidate.} \end{cases}$$

Its scalar reward on that round is $1/2 + \xi_t$. The selected arm and its sign are fixed before the independent sign $\epsilon_t$ is revealed, so $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\xi_t^2 \mid \mathcal{F}_{t-1}] \leq 1/4$. Full information arrives only after the action and does not change these conditional statements. Thus $M := \sum_t \xi_t$ is a martingale, the algorithm's scalar suffix reward is $n/2 + M$, and martingale orthogonality gives $\mathbb{E}[M^2] \leq n/4$.

For completeness, a sum $S_j$ of $\tau$ independent signs satisfies $\mathbb{E}[S_j^2] = \tau$ and $\mathbb{E}[S_j^4] = 3\tau^2 - 2\tau \leq 3\tau^2$. Hölder's inequality gives $\mathbb{E}[S_j^2] \leq (\mathbb{E}[|S_j|])^{2/3}(\mathbb{E}[S_j^4])^{1/3}$. Rearranging and using the two moments yields

$$\mathbb{E}[|S_j|] \geq \frac{(\mathbb{E}[S_j^2])^{3/2}}{(\mathbb{E}[S_j^4])^{1/2}} \geq \sqrt{\frac{\tau}{3}}.$$

Because $n/\ell \geq 1$, the inequality $\lfloor x \rfloor \geq x/2$ for $x \geq 1$ gives $\ell\tau \geq n/2$. Hence

$$\mathbb{E}[Q] \geq \frac{\ell\sqrt{\tau}}{2\sqrt{3}} \geq \frac{\sqrt{\ell n}}{2\sqrt{6}} > V/5.$$

By Cauchy–Schwarz, $Q^2 = (\sum_j |S_j|)^2/4 \leq (\ell/4)\sum_j S_j^2$. Therefore $\mathbb{E}[Q^2] \leq \ell^2\tau/4 \leq \ell n/4 = V^2/4$. Set $Z = Q - M$, the best candidate's scalar suffix advantage. Then

$$\mathbb{E}[Z] \geq V/5, \qquad \mathbb{E}[Z^2] \leq 2\mathbb{E}[Q^2] + 2\mathbb{E}[M^2] \leq V^2,$$

because $V^2 = \ell n \geq n$. Let $p = \mathbb{P}\{Z \geq V/10\}$. Splitting $[Z]_+$ at $V/10$ and applying Cauchy–Schwarz on the upper event gives

$$\frac{V}{5} \leq \mathbb{E}[Z] \leq \mathbb{E}[[Z]_+] \leq \frac{V}{10} + \sqrt{\mathbb{E}[Z^2], p} \leq \frac{V}{10} + V\sqrt{p}.$$

Subtracting $V/10$ and dividing by $V > 0$ gives $1/10 \leq \sqrt{p}$. Squaring both sides gives $p \geq 1/100$. These inequalities hold uniformly conditional on the prefix history and the revealed case.

Since $W_f \geq 128$, $W_f/64 \geq 2$, and therefore $h = \lfloor W_f/64 \rfloor \geq W_f/128$. Also $h \leq W_f/64 \leq T/64$, so $n \geq 63T/64$. This also gives

$$\min\{n, b\} \geq \frac{63}{64} \min\{T, b\}.$$

Indeed, if $b \leq n$, the left side is $b$, which is at least the right side. If $b > n$, then the left side is $n \geq 63T/64$, while the right side is at most $63T/64$. Multiplying this inequality by $n \geq 63T/64$ and taking square roots yields

$$V = \sqrt{n \min\{n, b\}} \geq \frac{63}{64} \sqrt{T \min\{T, b\}} = \frac{63}{64} W_f.$$

Since $h \leq W_f/64$ and $63/640 > 1/32$, we have $V/10 \geq 2h$. On the joint event that the initial deficit is at least $h/2$ and $Z \geq V/10$, choose a scalar-maximizing candidate. This same arm has final advantage at least $h/2$ in the zero-loss coordinate and at least $Z - h \geq h$ in the other coordinate. The two events have joint probability at least $(1/2)(1/100) = 1/200$, so

$$\mathbb{E}[R_T^{\mathrm{fs}}] \geq \frac{h}{400} \geq \frac{W_f}{51200}.$$

To compare $W_f$ with the claimed rate, first note that $K \leq 3q$. For $q \geq 2$,

$$b = \lfloor \log_2 q \rfloor \geq \tfrac{1}{2} \log_2 q = \frac{\log q}{2 \log 2}.$$

Moreover, $\log K \leq \log(3q) \leq (\log 6/\log 2) \log q$, so $b \geq (\log K)/(2 \log 6) \geq (\log K)/4$. Therefore

$$W_f = \sqrt{T \min\{T, b\}} \geq \tfrac{1}{2} \sqrt{T \min\{T, \log K\}} = \tfrac{1}{2} \min\{T, \sqrt{T \log K}\}.$$

If $W_f < 128$, the one-round hidden-arm construction in the bandit proof gives expected regret $1 - 1/K \geq 3/4$ even with full information, because the reward row is revealed only after the first action. The target rate is then smaller than 256, and $3/4 > 2^{-17} \cdot 256$. In the large case, $W_f/51200 \geq 2^{-17} \min\{T, \sqrt{T \log K}\}$. Both cases therefore give the stated constant $2^{-17}$. Duplication of a coordinate handles $D > 2$, and averaging over the finite reward distribution gives the deterministic-sequence statement. □

We next extend the construction from zero loss to a specified value of $L_0$.

**Lemma A.7** (Full-information lower bound at a specified loss)**.** *Under the conditions of Theorem A.5, every full-information algorithm has a deterministic reward sequence with $L^\star = L_0$ such that*

$$\mathbb{E}[R_T^{\mathrm{fs}}] \geq 2^{-20} \min\{U_0, \sqrt{U_0 \log K}\}.$$

*Full-information lower bound for a specified $L_0$.* For $U_0 \geq 1$, apply the preceding full-information zero-loss construction for $m = \lfloor U_0 \rfloor$ rounds, and append common rewards of total $U_0 - m$ as in the proof of Theorem 3.2. The cumulative advantages are unchanged and the easiest-coordinate loss is exactly $L_0$. Since $m \geq U_0/2$, we have both $m \geq U_0/2$ and $\sqrt{m \log K} \geq \frac{1}{2}\sqrt{U_0 \log K}$. Therefore

$$\min\{m, \sqrt{m \log K}\} \geq \tfrac{1}{2} \min\{U_0, \sqrt{U_0 \log K}\}.$$

For $0 < U_0 < 1$, use the one-round hidden reward $U_0\mathbf{1}$ from that proof. The expected regret is $(1 - 1/K)U_0 \geq 3U_0/4$. Since $K \geq 4$ implies $\log K > 1 > U_0$, the target minimum is $U_0$ in this range. When $U_0 = 0$, the target lower bound is zero. The first case gives constant $2^{-18}$, and the second gives constant $3/4$, so the smaller constant $2^{-20}$ is valid in all three cases. □

For the upper bound, exponential weights is run on one coordinate fixed before play.

**Lemma A.8** (Full-information upper bound at a specified loss)**.** *Under the conditions of Theorem A.5, there is a full-information algorithm that knows $L_0$ and satisfies, on every reward sequence with $L^\star = L_0$,*

$$\mathbb{E}[R_T^{\mathrm{fs}}] \leq 4 \min\{U_0, \sqrt{U_0 \log K}\}.$$

*Full-information upper bound for a specified $L_0$.* The learner may use the class parameter $L_0$, hence $U_0$, and it attends to a single coordinate. Fix a coordinate $d'$ before play and a deterministic reward sequence in the exact class $L^\star = L_0$, and write

$$g_{i,t} := r_{d'}^{i,t}, \qquad U_{d'} = \max_i \sum_{t=1}^{T} g_{i,t} \le U_0,$$

where the inequality holds because $U_{d'} \le U^\star = T - L^\star = U_0$ on the class. If $U_0 \le \log K$, play any fixed arm. The pathwise cap (9), which reads $R_T^{\mathrm{fs}} \le U_0$ on the class, gives $\mathbb{E}[R_T^{\mathrm{fs}}] \le U_0 = \min\{U_0, \sqrt{U_0 \log K}\}$, the equality because $U_0 \le \log K$ makes $U_0$ the smaller term, and this includes $U_0 = 0$.

Suppose now that $U_0 > \log K$. Run exponential weights on the rewards of coordinate $d'$, that is, the Hedge algorithm applied to gains rather than losses, with its learning rate set using the reward scale $U_0$ instead of the horizon. Before round $t$, set

$$p_{i,t} = \frac{\exp(\eta_0 \sum_{s<t} g_{i,s})}{\sum_{j=1}^{K} \exp(\eta_0 \sum_{s<t} g_{j,s})}, \qquad \eta_0 := \sqrt{\frac{\log K}{2U_0}} < 1,$$

and draw $a_t \sim p_t$ with fresh randomness. The update uses the full reward row and not the sampled action. Consequently, for each fixed reward sequence, the sequence $(p_t)_{t \le T}$ and the mixed cumulative gain

$$V := \sum_{t=1}^{T} p_t \cdot g_t,$$

the reward that the distributions $p_t$ collect in expectation, are deterministic.

Let $W_t = \sum_i \exp(\eta_0 \sum_{s<t} g_{i,s})$ be the total weight before round $t$, so $W_1 = K$. Convexity gives $e^{\eta_0 x} \le 1 + (e^{\eta_0} - 1)x$ for $x \in [0, 1]$. Hence

$$\begin{aligned} \frac{W_{t+1}}{W_t} &= \sum_{i=1}^{K} p_{i,t} e^{\eta_0 g_{i,t}} \\ &\le 1 + (e^{\eta_0} - 1) p_t \cdot g_t. \end{aligned}$$

Using $\log(1 + x) \le x$ and summing over $t$ gives

$$\log \frac{W_{T+1}}{W_1} \le (e^{\eta_0} - 1) \sum_{t=1}^{T} p_t \cdot g_t = (e^{\eta_0} - 1) V.$$

If $i^\star$ is a best arm, its term in $W_{T+1}$ equals $e^{\eta_0 U_{d'}}$. Therefore

$$\eta_0 U_{d'} - \log K \le \log \frac{W_{T+1}}{W_1} \le (e^{\eta_0} - 1) V.$$

Subtracting $\eta_0 V$, dividing by $\eta_0$, and using $e^x \le 1 + x + x^2$ for $x \in [0, 1]$ gives

$$U_{d'} - V \le \frac{\log K}{\eta_0} + \eta_0 V. \tag{16}$$

Put

$$X := \sum_{t=1}^{T} g_{a_t,t}, \qquad M := V - X,$$

so that $X$ is the learner's realized reward in coordinate $d'$ and $M$ is the sampling noise. Each increment $p_t \cdot g_t - g_{a_t,t}$ has mean zero given the history $\mathcal{F}_{t-1}$ before round $t$, so $M$ is a mean-zero martingale. Martingale orthogonality and $g_{i,t}^2 \le g_{i,t}$ give

$$\mathbb{E}[M^2] = \sum_{t=1}^{T} \mathbb{E}[\mathrm{Var}(g_{a_t,t} \mid \mathcal{F}_{t-1})] \le \sum_{t=1}^{T} \mathbb{E}[p_t \cdot g_t^2] \le V. \tag{17}$$

In particular, since $M$ has mean zero, its positive and negative parts have equal expectation, and the Cauchy–Schwarz inequality gives $\mathbb{E}[[M]_+] = \frac{1}{2}\mathbb{E}[|M|] \le \frac{1}{2}\sqrt{V}$.

If $V \le 2U_0$, the mixed gain stays within the reward scale and the potential bound controls the regret. Then (9), which gives $R_T^{\mathrm{fs}} \le [U_{d'} - X]_+ = [(U_{d'} - V) + M]_+$, the bound (16), and the positive-part triangle inequality $[x+y]_+ \le [x]_+ + [y]_+$ give

$$\begin{aligned}\mathbb{E}[R_T^{\mathrm{fs}}] &\le [U_{d'} - V]_+ + \mathbb{E}[[M]_+] \\ &\le \frac{\log K}{\eta_0} + 2\eta_0 U_0 + \frac{1}{2}\sqrt{2U_0} \\ &= 2\sqrt{2U_0 \log K} + \sqrt{U_0/2} < 4\sqrt{U_0 \log K},\end{aligned}$$

where the last inequality follows from $2\sqrt{2} + 1/\sqrt{2\log K} < 4$ for $K \ge 4$.

If $V > 2U_0$, the mixed gain is so far ahead of every fixed arm that only a large sampling deviation can make the regret positive. Let $h := V - U_{d'}$ be this lead, so $h > V/2$ because $U_{d'} \le U_0 < V/2$. Pathwise, $R_T^{\mathrm{fs}} \le [U_{d'} - X]_+ = [M - h]_+$. For $z < h$, $[z-h]_+ = 0$. For $z \ge h$, the inequality $(z-2h)^2 \ge 0$ is equivalent to $z - h \le z^2/(4h)$. Thus $[z-h]_+ \le z^2/(4h)$ for every real $z$ and $h > 0$. Apply this inequality with $z = M$, and then use (17), to obtain

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \frac{\mathbb{E}[M^2]}{4h} \le \frac{V}{4(V - U_{d'})} < \frac{1}{2} < \sqrt{U_0 \log K}.$$

The final inequality holds because $U_0 \log K > (\log K)^2 \ge (\log 4)^2 > 1/4$. Combining the two cases with the pathwise cap $R_T^{\mathrm{fs}} \le U_0$ proves

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le 4\min\{U_0, \sqrt{U_0 \log K}\}.$$

The policy uses $U_0 = T - L_0$, which the learner may know because the conditional minimax value lets the learner use the class parameter $L_0$ (Section 2.2). □

*Proof of Theorem A.5.* Lemma A.7 says that for every full-information algorithm there is a reward sequence in the class $L^\star = L_0$ on which its expected regret is at least $2^{-20}\min\{U_0, \sqrt{U_0 \log K}\}$. Taking the supremum over that class and then the infimum over algorithms therefore gives the same lower bound on $\mathcal{V}^{\mathrm{fi}}_{K,D,T}(L_0)$. Lemma A.8 provides one algorithm whose expected regret is at most $4\min\{U_0, \sqrt{U_0 \log K}\}$ on every sequence in the class. Taking the same supremum and infimum gives the upper bound. The two inequalities prove the theorem. □

### A.4 Detailed comparison guarantees

We first give the guarantees for EXP3 and its variants. Each method uses its original update rule, and its parameters are chosen without knowing $L_0$.

**Proposition A.9** (EXP3 variants when $L_0$ is unknown)**.** *Fix $K \ge 2, T \ge 2, D \ge 1$ and choose $d'$ before play. On every reward sequence with $L^\star = L_0$,*

$$\begin{aligned}
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, \sqrt{KT\log(KT)}\} && \textit{for EXP3.P and adversarial MO-KS},\\
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, \sqrt{KT\log K}\} && \textit{for EXP3-IX},\\
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, \sqrt{KL_{d'}\log(KT)} + K\log(KT)\} && \textit{for GREEN-IX},\\
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, (K\log K)^{1/3}T^{2/3}\} && \textit{for gain-based EXP3}.
\end{aligned} \tag{18}$$

*Original EXP3.1 satisfies*

$$\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\left\{U_0, \sqrt{KU_0\log K} + \left(\frac{KU_0^3}{\log K}\right)^{1/4} + K\log K\right\}.$$

The first four bounds depend on $T$ or on the chosen coordinate's loss $L_{d'}$. The EXP3.1 bound depends on $U_0$, with extra terms beyond $\sqrt{KU_0\log K}$. These bounds can therefore be larger than the bound in Theorem 4.5.

For Tsallis-INF, EXP3++, and MO-US, the available guarantees bound signed expected scalar regret. To bound the expectation of its positive part, we add a term measuring the extra reward from choosing the best arm separately on each round.

**Proposition A.10** (Original Tsallis-INF, EXP3++, and MO-US)**.** *For a coordinate $d'$ fixed before play, define*

$$S_{d'} := \sum_{t=1}^{T} \max_i r_{d'}^{i,t} - U_{d'}.$$

*The quantity $S_{d'}$ measures how much more reward we could obtain by choosing the best arm on each round than by choosing the best fixed arm. Each method satisfies*

$$\begin{aligned}
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, \sqrt{KT} + S_{d'}\} && \textit{for Tsallis-INF (IW)},\\
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, \sqrt{KT} + K\log T + S_{d'}\} && \textit{for Tsallis-INF (RV)},\\
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, \sqrt{KT\log K} + S_{d'}\} && \textit{for EXP3++},\\
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, \sqrt{KT\log K} + K + S_{d'}\} && \textit{for MO-US}.
\end{aligned} \tag{19}$$

*The same bounds hold with $L_{d'}$ in place of $S_{d'}$, because $S_{d'} \le L_{d'}$.*

The algorithms do not need to know $S_{d'}$ or $L_{d'}$. These quantities appear in the bounds and can be larger than $\min\{U_0, \sqrt{KU_0}\}$, even when $L^\star = L_0$. For example, if $D = 1$, $T$ is divisible by $K$, and the arms take turns receiving the only unit reward, then $U_0 = T/K$ and $S_{d'} = T - T/K$. The bound that includes $S_{d'}$ can therefore be larger than the optimal rate in this case.

We also consider versions that restart with fresh randomness after blocks of a fixed length. Their bounds depend on $K$ and $T$, so we no longer need the term $S_{d'}$.

**Proposition A.11** (Fixed-block versions of Tsallis-INF, EXP3++, and MO-US)**.** *Algorithm 6, with block length $h = \min\{T, \lceil\sqrt{KT}\rceil\}$, does not use $L_0$ and satisfies*

$$\begin{aligned}
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, K^{1/4}T^{3/4}\} && \textit{for Tsallis-INF (IW)},\\
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, K^{1/4}T^{3/4}\log(2T)\} && \textit{for Tsallis-INF (RV) and MO-US},\\
&\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, K^{1/4}T^{3/4}\sqrt{\log K}\} && \textit{for EXP3++}.
\end{aligned} \tag{20}$$

The $T^{3/4}$ terms can be larger than the optimal rate, so the bounds for these restarted methods are weaker than the bound in Theorem 4.5.

The bounds also hold when the algorithms use a fixed weighted average of coordinates. For GREEN-IX, Tsallis-INF, EXP3++, and MO-US, we then compute the loss or switching term from these weighted rewards.

### A.5 Algorithm specifications

This section gives pseudocode only for algorithms introduced or modified in Section 4. The order follows the main text. Every method uses rewards already observed, and the coordinate or fixed weighted average is chosen before play.

#### A.5.1 Scalar Poly-INF

Both optimal algorithms use the scalar policy below. Its scalar input is $g_{i,t} = r_{d'}^{i,t}$ for a coordinate $d'$ fixed before play. The policy and its logarithmic reward estimator are from Audibert & Bubeck (2010).

**Algorithm 1** POLY-INF($B$) with a logarithmic reward estimator

**Require:** Number of arms $K \geq 2$, scale parameter $B \geq 81K$

Set $\eta \leftarrow 2\sqrt{B}$, $\gamma \leftarrow 3\sqrt{K/B}$, and $\lambda \leftarrow 1/\sqrt{2KB}$.
Define $\psi : (-\infty, 0) \to (0, \infty)$ by $\psi(x) = (\eta/(-x))^2 + \gamma/K$.
Set $V_{i,0} \leftarrow 0$ and $p_{i,1} \leftarrow 1/K$ for all $i$.
**for** local rounds $s = 1, 2, \ldots$ **do**
    Draw $I_s \sim p_s$, receive reward $g_{I_s,s} \in [0, 1]$.
    Set $v_{i,s} \leftarrow 0$ for every $i \neq I_s$, and set

$$v_{I_s,s} \leftarrow -\frac{1}{\lambda} \log\left(1 - \frac{\lambda g_{I_s,s}}{p_{I_s,s}}\right).$$

    Set $V_{i,s} \leftarrow V_{i,s-1} + v_{i,s}$ for every $i$.
    Find the unique $z_s > \max_i V_{i,s}$ such that $\sum_{i=1}^{K} \psi(V_{i,s} - z_s) = 1$.
    Set $p_{i,s+1} \leftarrow \psi(V_{i,s} - z_s)$ for every $i$.
**end for**

The estimator and normalization in Algorithm 1 are well defined. Since $p_{I_s,s} \geq \gamma/K$,

$$\frac{\lambda g_{I_s,s}}{p_{I_s,s}} \leq \frac{\lambda K}{\gamma} = \frac{1}{3\sqrt{2}} < 1,$$

so the logarithm has a positive argument. The map $c \mapsto \sum_i \psi(V_{i,s} - c)$ is continuous and strictly decreasing on $(\max_i V_{i,s}, \infty)$, diverges at the left endpoint, and tends to $\gamma \leq 1/3$ at infinity. Hence $z_s$ exists and is unique. It lies in

$$\left(\max_i V_{i,s} + \eta,\ \max_i V_{i,s} + \eta\sqrt{K/(1-\gamma)}\right].$$

At the left endpoint the largest summand exceeds one. At the right endpoint each summand is at most $(1-\gamma)/K + \gamma/K$, so bisection on this interval computes $z_s$.

### A.5.2 ALGORITHMS USED WHEN $L_0$ IS KNOWN

The first algorithm receives $B = U_0 = T - L_0$. It runs Poly-INF when the supplied reward scale is large enough for its guarantee and otherwise uses the pathwise bound $R_T^{\mathrm{fs}} \leq U_0$.

**Algorithm 2** Fixed-coordinate policy when $L_0$ is known

**Require:** Arms $K \geq 2$, horizon $T$, scale parameter $B \in [0, T]$, precommitted coordinate $d' \in [D]$ (e.g., $d' = 1$)

**if** $B \geq 81K$ **then**
    Start a fresh copy of POLY-INF($B$) (Algorithm 1).
    **for** $t = 1, \ldots, T$ **do**
        Draw $a_t$ from the arm distribution maintained by the current copy and observe $r^{a_t,t}$.
        Feed back the scalar reward $r_{d'}^{a_t,t}$.
    **end for**
**else**
    Play arm 1 on every round.
**end if**

Section 4.1 next considers modified EXP3 and gain-based EXP3. Modified EXP3 uses the logarithmic estimator below. Gain-based EXP3 keeps the original update and changes only the theoretically chosen exploration parameter.

**Algorithm 3** EXP3 with logarithmic reward estimates

**Require:** Arms $K \geq 2$, horizon $T$, reward bound $B = U_0 = T - L_0$, coordinate $d'$ chosen before play

Set $H = \log(3K)$.
**if** $B \leq 4KH$ **then**
  Play one fixed arm on every round.
**else**
  Set $\gamma = 2\sqrt{KH/B}$, $\eta = 2\sqrt{H/(KB)}$, and $\beta = \sqrt{H/(2KB)}$.
  Initialize $w_{i,1} = 1$ for every arm $i$.
  **for** $t = 1, \ldots, T$ **do**
    Set $q_{i,t} = w_{i,t} / \sum_j w_{j,t}$ and $p_{i,t} = (1-\gamma)q_{i,t} + \gamma/K$ for every $i$.
    Draw $a_t \sim p_t$ and observe $g_t = r_{d'}^{a_t,t}$.
    Set $v_{a_t,t} = -\beta^{-1}\log(1 - \beta g_t / p_{a_t,t})$ and $v_{i,t} = 0$ for $i \neq a_t$.
    Set $w_{i,t+1} = w_{i,t}\exp(\eta v_{i,t})$ for every $i$.
  **end for**
**end if**

**Algorithm 4** Gain-based EXP3 with a supplied exploration parameter

**Require:** $K, T, \gamma \in (0, 1]$, a coordinate or fixed weights

Set $\eta = \gamma/K$ and $w_i = 1$ for all $i$.
**for** $t = 1, \ldots, T$ **do**
  Set $q_i = w_i / \sum_j w_j$ and $p_i = (1-\gamma)q_i + \gamma/K$.
  Draw $a_t \sim p$ and observe its scalar reward $g_{a_t,t}$.
  Set $\widehat{g}_{i,t} = g_{a_t,t}\mathbf{1}\{i = a_t\}/p_i$.
  Update $w_i \leftarrow w_i \exp(\eta \widehat{g}_{i,t})$ for all $i$.
**end for**

EXP3.P, EXP3-IX, and the adversarial branch of MO-KS retain their published updates. Their parameter choices and the resulting bounds are given with their proofs in Appendix A.6.7.

### A.5.3 Algorithms used when $L_0$ is unknown

Reward-doubling Poly-INF starts from $B = 81K$ and doubles this value only when the reward collected by the current copy reaches $B$.

**Algorithm 5** Reward-doubling Poly-INF

**Require:** Arms $K \geq 2$, horizon $T$, precommitted coordinate $d' \in [D]$ (e.g., $d' = 1$)

Set $e \leftarrow 0$, $B_e \leftarrow 81K$, $Y_e \leftarrow 0$, and start a fresh, independently randomized Poly-INF($B_e$).
**for** $t = 1, \ldots, T$ **do**
  Draw $a_t$ from the current copy, observe $g_t = r_{d'}^{a_t,t}$, update the copy, and set $Y_e \leftarrow Y_e + g_t$.
  **if** $Y_e \geq B_e$ and $t < T$ **then**
    Set $e \leftarrow e + 1$, $B_e \leftarrow 2B_{e-1}$, $Y_e \leftarrow 0$, and start a fresh independent Poly-INF($B_e$) on round $t + 1$.
  **end if**
**end for**

The unknown-$L_0$ comparison also considers fixed-block versions of Tsallis-INF, EXP3++, and MO-US. The wrapper below restarts the selected published algorithm independently in each block.

**Algorithm 6** Fixed-block restarts of a scalar algorithm

**Require:** Scalar algorithm $\mathcal{A}$, $K, T$, integer $h \in \{1, \dots, T\}$, a coordinate or fixed weights

1: **for** blocks of consecutive rounds of length $h$, with a shorter last block **do**
2: Initialize a fresh copy of $\mathcal{A}$, with independent random choices and its local time set to one.
3: If $\mathcal{A}$ needs a horizon, supply the length of this block.
4: Run that copy for the entire block, supplying only its chosen arm's scalar reward or loss on each round.
5: **end for**

### A.6 Upper bounds when $L_0$ is known

We follow Section 4.1. We first record the scalar reduction used by all comparison algorithms and the Poly-INF guarantee used by both optimal algorithms. We then prove Theorem 4.3 and analyze the methods in Proposition 4.4 in their main-text order.

#### A.6.1 Common scalar reduction

The scalar algorithms compared in Section 4 observe only $g_{i,t} = r_{d'}^{i,t}$ in the coordinate fixed before play. A method written for losses instead receives $\ell_{i,t} = 1 - g_{i,t}$. These conventions give the same realized scalar regret,

$$R_T^{d'} = \max_i \sum_t g_{i,t} - \sum_t g_{a_t,t} = \sum_t \ell_{a_t,t} - \min_i \sum_t \ell_{i,t}.$$

The best-arm loss in this scalar problem is $L_{d'}$, not necessarily $L^\star$. On the exact class, $U_0 = T - L_0$ and

$$0 \le R_T^{\mathrm{fs}} \le \min\{U_0, [R_T^{d'}]_+\}.$$

This inequality is the basis for all upper-bound conversions below.

#### A.6.2 Signed regret and positive regret

Fix the scalar rewards $g_{i,t} \in [0,1]$ before play. Write $G_i = \sum_t g_{i,t}$, $G^\star = \max_i G_i$, $A = \sum_t g_{a_t,t}$, and $L_g = T - G^\star$. For one coordinate, $L_g = L_{d'}$. For a fixed weighted average it is the best-arm loss for that average, not necessarily $L^\star$. Throughout these comparisons $G^\star \le U_0$.

**Lemma A.12** (Conversion with a bound on negative regret)**.** *Suppose a scalar algorithm satisfies* $\mathbb{E}[G^\star - A] \le B$*, where* $B \ge 0$*. Define* $S_g = \sum_t \max_i g_{i,t} - G^\star$*. Then*

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \min\{U_0, B + S_g\} \le \min\{U_0, B + L_g\}.$$

*Proof.* Put $R = G^\star - A$. Since $A \le \sum_t \max_i g_{i,t}$, $[-R]_+ \le S_g \le T - G^\star = L_g$. The identity $[R]_+ = R + [-R]_+$ gives

$$\mathbb{E}[[R]_+] = \mathbb{E}[R] + \mathbb{E}[[-R]_+] \le B + S_g \le B + L_g.$$

Separately, the pathwise cap gives $\mathbb{E}[R_T^{\mathrm{fs}}] \le U_0$. Combining this cap with $R_T^{\mathrm{fs}} \le [R]_+$ proves both minima in the lemma. □

The extra term cannot simply be dropped. It measures how much an algorithm could gain by switching arms, relative to the best fixed arm. It is zero if one arm has the largest scalar reward on every round. The lemma gives a valid bound for an unchanged algorithm, but in general does not give a useful bound uniform over $L^\star = L_0$.

#### A.6.3 The scalar Poly-INF bound

Integrating the scalar tail bound over confidence levels gives the positive-part expectation bound used by Algorithm 2. For reward doubling, we use the deterministic inequality within the source proof and control the estimation error across all epochs together. Neither that inequality nor the full scalar tail bound requires the scale parameter $B$ to exceed the best arm's reward. This matters because an epoch

can start with a value of $B$ below that reward. The assumption that all rewards are fixed before play makes the full-horizon best arm and the scale index used below deterministic.

For a value $B$, POLY-INF($B$) denotes the policy in Algorithm 1, with exponent 2 and parameters $(\eta, \gamma, \lambda) = (2\sqrt{B},\ 3\sqrt{K/B},\ 1/\sqrt{2KB})$. The next lemma states the scalar guarantee used by both algorithms.

**Lemma A.13** (Positive-regret bound for POLY-INF)**.** *For a deterministic $K$-armed scalar reward sequence $g_{i,t} \in [0,1]$ of length $n \geq K$, write $G_n^\star := \max_i \sum_{t=1}^n g_{i,t}$ and $R_n := G_n^\star - \sum_{t=1}^n g_{a_t,t}$. For every $B \geq 81K$ and every $\delta \in (0,1)$, the policy* POLY-INF($B$)*, which does not depend on $\delta$, satisfies*

$$\mathbb{P}\left\{ R_n > 4.5 G_n^\star \sqrt{\frac{K}{B}} + 4\sqrt{KB} + \sqrt{2KB}\log(\delta^{-1}) \right\} \leq \delta. \tag{21}$$

*If $G_n^\star \leq B$, then*

$$\mathbb{E}[[R_n]_+] \leq (8.5 + \sqrt{2})\sqrt{KB} < 10\sqrt{KB}. \tag{22}$$

*Proof of Lemma A.13.* The deterministic-adversary part of Audibert & Bubeck (2010, Theorem 18, Eq. (18)) considers a fixed gain matrix in $[0,1]^{K\times n}$, where $n \geq K \geq 2$, and a parameter $G_0 \geq 81K$. With

$$q = 2, \qquad \eta = 2\sqrt{G_0}, \qquad \gamma = 3\sqrt{K/G_0}, \qquad \beta_{\mathrm{AB}} = \frac{1}{\sqrt{2KG_0}},$$

that theorem gives, with probability at least $1-\delta$,

$$R_n^{\mathrm{AB}} \leq 4.5 G_{\max}^{\mathrm{AB}} \sqrt{\frac{K}{G_0}} + 4\sqrt{KG_0} + \sqrt{2KG_0}\log(\delta^{-1}).$$

The source statement does not require $G_{\max}^{\mathrm{AB}} \leq G_0$. Its notation corresponds to ours as

$$G_0 \mapsto B, \quad G_{\max}^{\mathrm{AB}} \mapsto G_n^\star, \quad I_t^{\mathrm{AB}} \mapsto a_t, \quad R_n^{\mathrm{AB}} \mapsto G_n^\star - \sum_{t=1}^n g_{a_t,t}, \quad \beta_{\mathrm{AB}} \mapsto \lambda.$$

The update in Algorithm 1 is the same policy after these substitutions. It depends on $K$ and $B$, but not on $n$ or $\delta$, so every confidence level concerns the same algorithm. The source bound is therefore exactly (21).

It remains to prove (22). If $G_n^\star \leq B$, the first two terms in (21) are at most $4.5\sqrt{KB} + 4\sqrt{KB} = 8.5\sqrt{KB}$. Set $a = 8.5\sqrt{KB}$ and $b = \sqrt{2KB}$. For $x > a$, substitute $\delta = \exp(-(x-a)/b) \in (0,1)$ into (21) to obtain $\mathbb{P}\{R_n > x\} \leq \exp(-(x-a)/b)$. The layer-cake identity, which writes the expectation of a nonnegative random variable as the integral of its tail probabilities, gives

$$\mathbb{E}[[R_n]_+] = \int_0^\infty \mathbb{P}\{R_n > x\}\, dx \leq a + \int_a^\infty e^{-(x-a)/b}\, dx = a + b,$$

which is (22). □

Lemma A.13 speaks about a scalar bandit problem, so we record how its quantities read once the scalar rewards are those of the precommitted coordinate. When Algorithm 2 runs POLY-INF($B$) on the scalar rewards $g_{i,t} = r_{d'}^{i,t}$ over the full horizon $n = T$, the scalar quantities in Lemma A.13 are

$$G_T^\star = \max_i \sum_{t=1}^T r_{d'}^{i,t} \leq U^\star, \qquad R_T = R_T^{d'}, \qquad R_T^{\mathrm{fs}} \leq [R_T^{d'}]_+. \tag{23}$$

The first identity is the definition of scalar regret for this coordinate. The inequality $G_T^\star \leq U^\star$ follows because $U^\star$ is the maximum cumulative reward over all arm-coordinate pairs. The last inequality is the coordinate case of Lemma 4.1.

### A.6.4 KNOWN-SCALE POLY-INF

We now apply the scalar guarantee with the supplied value $U_0$.

*Proof of Theorem 4.3.* Knowing $L_0$ determines $B = U_0$ for POLY-INF and also gives a pathwise cap. The two branches of Algorithm 2 use one of these facts each. On the class $L^\star = L_0$, (5) gives $\max_i \sum_t r_{d'}^{i,t} \le U^\star = U_0$ for every coordinate $d'$, in particular for the coordinate $d'$ fixed before play on which Algorithm 2 runs, so the best cumulative reward the policy faces is at most the scale $B = U_0$, and the pathwise bound (9) gives $R_T^{\mathrm{fs}} \le U_0$ on every reward sequence in the class and every play path. If $U_0 \ge 81K$, Algorithm 2 runs POLY-INF$(U_0)$ over the full horizon $T \ge U_0 \ge K$, so Lemma A.13 applies, and its expectation bound (22) together with the reduction (23) gives $\mathbb{E}[R_T^{\mathrm{fs}}] < 10\sqrt{KU_0}$, where $\sqrt{KU_0} = \min\{U_0, \sqrt{KU_0}\}$ because $U_0 \ge K$. If $U_0 < 81K$, Algorithm 2 plays a fixed arm and the pathwise bound yields

$$R_T^{\mathrm{fs}} \le U_0 \le 9 \min\{U_0, \sqrt{KU_0}\}.$$

When $U_0 \le K$, the minimum is $U_0$, so the inequality holds with factor one. When $K < U_0 < 81K$, the minimum is $\sqrt{KU_0}$, and dividing $U_0$ by $\sqrt{KU_0}$ gives $\sqrt{U_0/K} < 9$. In both branches $\mathbb{E}[R_T^{\mathrm{fs}}] \le 10 \min\{U_0, \sqrt{KU_0}\}$, which proves the theorem. □

The remaining proofs follow Proposition 4.4. They analyze modified EXP3, gain-based EXP3, EXP3.P and MO-KS, and EXP3-IX in that order.

### A.6.5 EXP3 WITH LOGARITHMIC REWARD ESTIMATES

Algorithm 3 uses the exponential-weights policy and logarithmic reward estimate of Audibert & Bubeck (2010, Theorem 21). The fixed-arm branch covers the values of $B$ excluded by that theorem. It changes neither the guarantee in Proposition 4.4 nor the information available to the algorithm. Only the chosen arm's reward is used in each update.

*Modified-EXP3 part of Proposition 4.4.* We prove the explicit bound

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \min\left\{U_0, (3+\sqrt{2})\sqrt{KU_0 \log(3K)}\right\}.$$

Write $B = U_0$, $H = \log(3K)$, and $G_T^\star = \max_i \sum_t g_{i,t} = U_{d'} \le B$. If $B = 0$, the bound $R_T^{\mathrm{fs}} \le B$ proves the claim. If $0 < B \le 4KH$, the same bound applies to the fixed-arm rule, and

$$B \le 2\sqrt{KBH} \le (3+\sqrt{2})\sqrt{KBH}.$$

Thus the minimum in this explicit bound equals $B$ in this case.

Suppose now that $B > 4KH$, so $0 < \gamma < 1$. The logarithmic estimate is well defined because

$$0 \le \frac{\beta g_t}{p_{a_t,t}} \le \frac{\beta K}{\gamma} = \frac{1}{2\sqrt{2}} < 1.$$

With $G_0 = B$, Algorithm 3 is the policy in Audibert & Bubeck (2010, Theorem 21). Indeed, normalizing the exponential part of its probability rule gives $p_{i,t} = (1-\gamma)w_{i,t}/\sum_j w_{j,t} + \gamma/K$. For every $\delta \in (0,1)$, that theorem gives

$$\mathbb{P}\left\{R_T^{d'} > \frac{5}{2}G_T^\star\sqrt{\frac{KH}{B}} + \frac{1}{2}\sqrt{KBH} + \sqrt{\frac{2KB}{H}}\log\frac{K}{\delta}\right\} \le \delta.$$

The parameters do not depend on $\delta$, so all these bounds concern the same algorithm. Since $G_T^\star \le B$, define

$$s = \sqrt{KBH}, \qquad b = \sqrt{\frac{2KB}{H}}, \qquad a = 3s + b\log K.$$

Substituting $G_T^\star \le B$ and $\log(K/\delta) = \log K + \log(1/\delta)$ into the preceding display gives $\mathbb{P}\{R_T^{d'} > a + b\log(1/\delta)\} \le \delta$. For $x > a$, set $\delta = e^{-(x-a)/b}$. Then $\mathbb{P}\{R_T^{d'} > x\} \le e^{-(x-a)/b}$, so

$$\begin{aligned}\mathbb{E}[[R_T^{d'}]_+] &= \int_0^\infty \mathbb{P}\{R_T^{d'} > x\}\,dx \\ &\le a + \int_a^\infty e^{-(x-a)/b}\,dx = 3s + b(\log K + 1) \\ &\le (3+\sqrt{2})s.\end{aligned}$$

The last inequality uses $\log K + 1 \le \log(3K) = H$. Finally, $R_T^{\mathrm{fs}} \le U_0$ and $R_T^{\mathrm{fs}} \le [R_T^{d'}]_+$ prove the modified-EXP3 bound in Proposition 4.4. □

### A.6.6 Original gain-based EXP3

This is the gain-based EXP3 update of Auer et al. (2002b). Only the choice of $\gamma$ changes between the parameter settings below; no bias or logarithmic transformation is added to its reward estimate.

*Gain-based EXP3 bounds in Propositions 4.4 and A.9.* Choose a deterministic $k \in \arg\max_i G_i$. If $\gamma = 1$, the claimed bound follows from $R_T^{\mathrm{fs}} \le U_0$, so suppose $\gamma < 1$. Since $0 \le \eta\widehat{g}_{i,t} \le 1$, $e^x \le 1 + x + (e-2)x^2$ on this interval. If $W_t = \sum_i w_{i,t}$, then

$$\begin{aligned}\log\frac{W_{t+1}}{W_t} &= \log\sum_i q_{i,t}e^{\eta\widehat{g}_{i,t}} \\ &\le \log\bigl(1 + \eta\langle q_t, \widehat{g}_t\rangle + (e-2)\eta^2\langle q_t, \widehat{g}_t^2\rangle\bigr) \\ &\le \eta\langle q_t, \widehat{g}_t\rangle + (e-2)\eta^2\langle q_t, \widehat{g}_t^2\rangle.\end{aligned}$$

For the comparator arm $k$, $W_{T+1} \ge \exp(\eta\sum_t \widehat{g}_{k,t})$, while $W_1 = K$. Summing the preceding upper bound, applying this lower bound, and dividing by $\eta$ gives

$$\sum_t \widehat{g}_{k,t} - \sum_t \langle q_t, \widehat{g}_t\rangle \le \frac{\log K}{\eta} + (e-2)\eta\sum_t \langle q_t, \widehat{g}_t^2\rangle.$$

Here $\langle p_t, \widehat{g}_t\rangle = g_{a_t,t}$ and $(1-\gamma)q_{i,t} \le p_{i,t}$. Also,

$$A = \sum_t \langle p_t, \widehat{g}_t\rangle = (1-\gamma)\sum_t \langle q_t, \widehat{g}_t\rangle + \frac{\gamma}{K}\sum_{t,i}\widehat{g}_{i,t}.$$

Multiply the potential inequality by $1-\gamma$, use $\eta = \gamma/K$, and drop the final nonpositive term $-(\gamma/K)\sum_{t,i}\widehat{g}_{i,t}$. This yields

$$G^\star - A \le \gamma G^\star + (1-\gamma)M + \frac{(1-\gamma)K\log K}{\gamma} + Z,$$

where

$$M = \sum_t (g_{k,t} - \widehat{g}_{k,t}), \qquad Z = (e-2)\frac{\gamma}{K}\sum_{t,i} p_{i,t}\widehat{g}_{i,t}^2 \ge 0.$$

Conditional unbiasedness and martingale orthogonality give

$$\mathbb{E}[M] = 0, \qquad \mathbb{E}[M^2] = \sum_t g_{k,t}^2\mathbb{E}[1/p_{k,t} - 1] \le \frac{K}{\gamma}\sum_t g_{k,t}^2 \le \frac{KU_0}{\gamma}.$$

For any integrable mean-zero variable, $\mathbb{E}[[M]_+] = \mathbb{E}[|M|]/2$. Also, since each arm's total scalar reward is at most $U_0$,

$$\mathbb{E}[Z] = (e-2)\frac{\gamma}{K}\sum_{t,i} g_{i,t}^2 \le (e-2)\frac{\gamma}{K}\sum_i G_i \le (e-2)\gamma U_0.$$

Take positive parts before expectation, use Cauchy–Schwarz for $M$, and drop factors $1-\gamma \le 1$. Together with $R_T^{\mathrm{fs}} \le U_0$, this gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \min\left\{U_0, (e-1)\gamma U_0 + \frac{K\log K}{\gamma} + \frac{1}{2}\sqrt{\frac{KU_0}{\gamma}}\right\}. \tag{24}$$

This is the parameter-dependent EXP3 bound used in both comparison propositions.

For $V \geq U_0$, take $\gamma_V = \min\{1, (K \log K/V)^{1/3}\}$, with $\gamma_0 = 1$. If $V \leq K \log K$, the bound $U_0$ suffices. Otherwise, write $C = K \log K \leq V$, so $\gamma_V = (C/V)^{1/3}$. Since $U_0 \leq V$, the three terms satisfy

$$\begin{aligned}(e-1)\gamma_V U_0 &\leq (e-1)C^{1/3}V^{2/3},\\ \frac{C}{\gamma_V} &= C^{2/3}V^{1/3} \leq C^{1/3}V^{2/3},\\ \frac{1}{2}\sqrt{\frac{KU_0}{\gamma_V}} &\leq \frac{1}{2}K^{1/3}(\log K)^{-1/6}V^{2/3} \leq \frac{1}{2\sqrt{\log 2}}C^{1/3}V^{2/3}.\end{aligned}$$

Thus their sum is $O((K \log K)^{1/3}V^{2/3})$. Setting $V = T$ proves the unknown-$L_0$ rate; setting $V = U_0$ proves the known-$L_0$ rate. The displayed parameter-dependent bound also applies to the original square-root parameter choice. That choice is used in the saved experiments. □

### A.6.7 EXP3.P, MO-KS, AND EXP3-IX WITH KNOWN $L_0$

We next check whether supplying $L_0$ improves the parameter-dependent bounds for EXP3.P or EXP3-IX. This gives the EXP3.P bound with $U_0$-based parameters in Proposition 4.4 and explains why setting the EXP3-IX learning rate using $U_0$ does not improve the bound in our analysis. Neither calculation replaces the method's reward or loss estimate.

**Lemma A.14** (EXP3.P and EXP3-IX when $L_0$ is known)**.** *Fix $K \geq 2, T \geq 2, D \geq 1$, $L_0 \in [0, T]$, and a coordinate $d'$ before play. Put $U_0 = T - L_0$. On every reward sequence with $L^\star = L_0$, the following bounds hold.*

1. *Run EXP3.P on rewards in coordinate $d'$, with*

$$q = \log\big(KT(T+1)^2\big), \qquad \alpha = 2\sqrt{q}, \qquad \gamma = \min\left\{\frac{3}{5}, 3\sqrt{\frac{K \log K}{5U_0}}\right\}.$$

*At $U_0 = 0$, set $\gamma = 3/5$. Then*

$$\mathbb{E}[R_T^{\mathrm{fs}}] \leq \min\left\{U_0,\ 2\sqrt{5KU_0 \log K} + 4\sqrt{KTq} + 8q + \frac{U_0}{(T+1)^2}\right\}. \tag{25}$$

*The same result applies to the adversarial branch of MO-KS with this parameter choice. In particular, this is an $O(\min\{U_0, \sqrt{KT \log(KT)}\})$ upper bound.*

2. *Run EXP3-IX on losses $\ell_{i,t} = 1 - r_{d'}^{i,t}$, with any constant $\eta > 0$ and $\gamma = \eta/2$. Then*

$$\mathbb{E}[R_T^{\mathrm{fs}}] \leq \min\left\{U_0,\ \frac{2\log K + 1 + \log 2}{\eta} + \eta KT + 1 + \log 2\right\}. \tag{26}$$

*In particular, the choice $\eta = \sqrt{2 \log K/(KT)}$ gives $O(\min\{U_0, \sqrt{KT \log K}\})$.*

*Proof.* For EXP3.P, the sampling probabilities and updates are

$$p_{i,t} = (1-\gamma)\frac{w_{i,t}}{\sum_j w_{j,t}} + \frac{\gamma}{K}, \qquad w_{i,t+1} = w_{i,t} \exp\left[\frac{\gamma}{3K}\left(\frac{g_{i,t}\mathbf{1}\{a_t = i\}}{p_{i,t}} + \frac{\alpha}{p_{i,t}\sqrt{KT}}\right)\right],$$

with $w_{i,1} = 1$. The common initial factor in the source algorithm cancels from its probabilities. Set $\delta_0 = (T+1)^{-2}$, so $q = \log(KT/\delta_0)$. The inequality $q \leq KT$ holds for $K, T \geq 2$: $KT - \log(KT(T+1)^2)$ is positive at $(2,2)$ and increasing in each variable on this range. Thus $\alpha \leq 2\sqrt{KT}$, as required by Auer et al. (2002b, Lemmas 6.1 and 6.2).

Let $G_T^\star = U_{d'} \leq U_0$, and let $\widehat{U}$ be the largest upper confidence estimate in those lemmas. With probability at least $1 - \delta_0$, $\widehat{U} \geq G_T^\star$, and the collected reward $G_T^{\mathrm{alg}}$ satisfies

$$G_T^{\mathrm{alg}} \geq \left(1 - \frac{5\gamma}{3}\right)\widehat{U} - \frac{3K \log K}{\gamma} - 2\alpha\sqrt{KT} - 2\alpha^2.$$

Since $\gamma \le 3/5$, this gives on the same event

$$R_T^{d'} \le \frac{5\gamma}{3} U_0 + \frac{3K \log K}{\gamma} + 4\sqrt{KTq} + 8q.$$

If $U_0 \ge 5K \log K$, the chosen $\gamma$ makes the first two terms sum to $2\sqrt{5KU_0 \log K}$. On the complementary event, $R_T^{\mathrm{fs}} \le U_0$, so its contribution to the expectation is at most $U_0\delta_0$. This proves (25) in this range. If $0 < U_0 < 5K \log K$, then $2\sqrt{5KU_0 \log K} > 2U_0$, so the displayed minimum is $U_0$, which bounds every algorithm's regret. The case $U_0 = 0$ follows from the same bound. MO-KS runs this scalar method on a coordinate chosen before play (Xu & Klabjan, 2023, Algorithm 1), so the argument also applies after conditioning on that choice.

To simplify (25), first note that

$$\log(KT) \le q = \log\big(KT(T+1)^2\big) \le 3\log(KT),$$

because $T + 1 \le KT$ for $K, T \ge 2$. Moreover, $q \le KT$, as verified above. Hence

$$\sqrt{KU_0 \log K} \le \sqrt{KTq}, \qquad q \le \sqrt{KTq}, \qquad \frac{U_0}{(T+1)^2} \le 1 \le \sqrt{KTq}.$$

Every term outside $U_0$ in (25) is therefore at most a numerical constant times $\sqrt{KT \log(KT)}$. This proves the compact upper bound stated in the lemma.

For EXP3-IX, the parameter-dependent inequality in the proof of Neu (2015a, Theorem 1), with $\gamma = \eta/2$, gives for every $\delta \in (0, 1)$

$$\mathbb{P}\left\{ R_T^{d'} > \frac{2\log K}{\eta} + \eta KT + \left(\frac{1}{\eta} + 1\right) \log \frac{2}{\delta} \right\} \le \delta.$$

For fixed $\eta$, all confidence levels concern the same algorithm. Put $c = 2\log K/\eta + \eta KT$ and $d = 1/\eta + 1$. For $x > c + d\log 2$, choosing $\delta = 2e^{-(x-c)/d}$ gives $\mathbb{P}\{R_T^{d'} > x\} \le e^{-(x-c-d\log 2)/d}$. Thus

$$\mathbb{E}[[R_T^{d'}]_+] \le c + d\log 2 + \int_{c+d\log 2}^{\infty} e^{-(x-c-d\log 2)/d}\, dx = c + (1 + \log 2)d.$$

Combining this with $R_T^{\mathrm{fs}} \le \min\{U_0, [R_T^{d'}]_+\}$ proves (26). The $\eta$-dependent part of this bound is $(2\log K + 1 + \log 2)/\eta + \eta KT$, whose minimizing value is $\eta = \sqrt{(2\log K + 1 + \log 2)/(KT)}$. It does not depend on $U_0$, which is why supplying $L_0$ does not improve this EXP3-IX bound. □

### A.7 Upper bounds when $L_0$ is unknown

We now follow Section 4.2. The optimal reward-doubling method comes first, followed by the horizon-scale and fixed-action baselines and then Propositions A.9, A.10, and A.11.

#### A.7.1 Reward-doubling Poly-INF

Restart times depend on the rewards collected by the algorithm, whereas Lemma A.13 concerns a fixed number of rounds. We therefore use the deterministic inequality within the proof of Audibert & Bubeck (2010, Appendix C.9) and control estimation error across all epochs with one supermartingale. Keeping the same comparator arm in every epoch makes the regrets add up to the full scalar regret. When that regret is positive, the best arm's cumulative reward bounds the largest epoch scale. This bound lets us choose one supermartingale weight that is valid throughout those epochs, without taking a union bound over restart times. We first prove a high-probability bound, then integrate it to obtain Theorem 4.5.

**Lemma A.15** (Explicit high-probability form of Theorem 4.5)**.** *For $K \ge 2, T \ge 2, D \ge 1$ and $\delta \in (0, 1)$, put $A(\delta) := 48 + \sqrt{2}\log(\delta^{-1})$. On every deterministic reward sequence fixed before play, Algorithm 5 satisfies, with probability at least $1 - \delta$,*

$$R_T^{\mathrm{fs}} \le [R_T^{d'}]_+ \le \min\Big\{ U^\star,\ A(\delta)\big(\sqrt{KU_{d'}} + 9K\big) \Big\}, \tag{27}$$

*where $d'$ is the coordinate fixed before play in Algorithm 5 and $U_{d'}$ is the best cumulative reward in that coordinate.*

Fix the coordinate $d'$ of Algorithm 5 and write $g_{i,t} := r_{d'}^{i,t} \in [0,1]$. The reward sequence is fixed before play, so the smallest index $k^\star$ attaining $\max_{i\in[K]} \sum_{t=1}^{T} g_{i,t}$ is a deterministic arm, with $U_{d'} = \sum_{t=1}^{T} g_{k^\star,t} \le U^\star$, and by (2) $R_T^{d'} = \sum_{t=1}^{T}(g_{k^\star,t} - g_{a_t,t})$. The started epochs are $e = 0, \dots, e_{\max}$ with $B_e = 81K\,2^e$. Epoch $e$ starts at round $t_e$, occupies the rounds $\mathcal{T}_e = \{t_e, \dots, t_{e+1} - 1\}$ with $t_{e_{\max}+1} := T + 1$, and has length $n_e := |\mathcal{T}_e|$. Its POLY-INF($B_e$) copy uses $(\eta_e, \gamma_e, \lambda_e) = (2\sqrt{B_e},\, 3\sqrt{K/B_e},\, 1/\sqrt{2KB_e})$. If the threshold of epoch $e_{\max}$ is crossed on round $T$, no further epoch is created, because the restart condition requires $t < T$. Write $e(t)$ for the epoch of round $t$, $p_t$ for the arm distribution of the copy active on round $t$, and $v_{i,t}$ for that copy's estimate, so that $v_{a_t,t} = -\lambda_{e(t)}^{-1} \log(1 - \lambda_{e(t)} g_{a_t,t}/p_{a_t,t})$ and $v_{i,t} = 0$ for $i \ne a_t$. The normalization of Algorithm 1 gives $p_{i,t} \ge \gamma_{e(t)}/K$, hence $\lambda_{e(t)} g_{i,t}/p_{i,t} \le \lambda_{e(t)} K/\gamma_{e(t)} = 1/(3\sqrt{2})$.

Let $\mathcal{F}_t$ be the sigma-field generated by $a_1, \dots, a_t$, with $\mathcal{F}_0$ trivial. Because the reward sequence is deterministic, the epoch index $e(t)$, the vector $p_t$, and each event $\{t \in \mathcal{T}_e\}$ are $\mathcal{F}_{t-1}$-measurable, since $\{t \in \mathcal{T}_e\} = \{t_e \le t < t_{e+1}\}$ is decided by the rewards collected before round $t$. Each copy draws its action from $p_t$ with randomness independent of the past, so $\mathbb{P}\{a_t = i \mid \mathcal{F}_{t-1}\} = p_{i,t}$ for every $i$. This is the only property of the learner's randomization used below.

For every started epoch $e$ put

$$Y_e := \sum_{t\in\mathcal{T}_e} g_{a_t,t}, \qquad G_e := \sum_{t\in\mathcal{T}_e} g_{k^\star,t}, \qquad V_i^{(e)} := \sum_{t\in\mathcal{T}_e} v_{i,t},$$

$$r_e := G_e - Y_e, \qquad D_e := G_e - V_{k^\star}^{(e)}.$$

We call $r_e$ the epoch regret against $k^\star$ and $D_e$ the epoch deficit of the estimator of $k^\star$. Both may be negative. Because the epochs partition $[T]$ and the comparator is the same arm in every epoch,

$$R_T^{d'} = \sum_{e=0}^{e_{\max}} r_e, \qquad \sum_{e=0}^{e_{\max}} D_e = \sum_{t=1}^{T} \big(g_{k^\star,t} - v_{k^\star,t}\big). \tag{28}$$

Every started epoch satisfies $Y_e \le B_e + 1$, because a crossing adds one reward in $[0,1]$ to a total below $B_e$, while an epoch that reaches round $T$ without crossing has $Y_e < B_e$. Every epoch $e < e_{\max}$ was completed, so $Y_e \ge B_e$. Doubling keeps the sum of the square roots of the scales within a constant factor of the last one,

$$\sum_{e=0}^{e_{\max}} \sqrt{B_e} = \sqrt{B_{e_{\max}}} \sum_{s=0}^{e_{\max}} 2^{-s/2} \le (2+\sqrt{2})\sqrt{B_{e_{\max}}}. \tag{29}$$

The last inequality uses $\sum_{s=0}^{\infty} 2^{-s/2} = (1 - 2^{-1/2})^{-1} = 2 + \sqrt{2}$. If $R_T^{d'} > 0$, then $\sum_{e\le e_{\max}} Y_e < U_{d'}$, and since rewards are nonnegative,

$$B_{e_{\max}} - 81K = \sum_{e<e_{\max}} B_e \le \sum_{e<e_{\max}} Y_e < U_{d'}, \tag{30}$$

so that $2^{e_{\max}} < 1 + U_{d'}/(81K)$. Define the deterministic index and scale

$$e^\star := \left\lfloor \log_2\!\Big(1 + \frac{U_{d'}}{81K}\Big) \right\rfloor, \qquad B_{e^\star} = 81K\,2^{e^\star} \le U_{d'} + 81K. \tag{31}$$

Then $e_{\max} \le e^\star$ on the event $\{R_T^{d'} > 0\}$, and $\sqrt{KB_{e^\star}} \le \sqrt{KU_{d'}} + 9K$ by $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $\sqrt{81K^2} = 9K$.

**Lemma A.16** (Epoch inequality)**.** *For every started epoch $e$, on every play path,*

$$r_e \le D_e + 14\sqrt{KB_e}. \tag{32}$$

*Proof.* Write $s_e := \sqrt{K/B_e} \le 1/9$, so that $\gamma_e = 3s_e$, $\lambda_e K = s_e/\sqrt{2}$, and $s_e B_e = \sqrt{KB_e}$. Define

$$x_0 := \frac{1}{3\sqrt{2}}, \qquad c_{\mathrm{AB}} := \frac{-\log(1-x_0)}{x_0}, \qquad \nu := \frac{1}{x_0} + \frac{1}{\log(1-x_0)},$$

and, with $y_e := (K/B_e)^{1/4}$,

$$\mu_e := \exp\left\{\sqrt{3}c_{\mathrm{AB}}y_e\left(1-\frac{c_{\mathrm{AB}}y_e}{4\sqrt{3}}\right)^{-2}\right\}, \qquad \xi_e := \left(\frac{c_{\mathrm{AB}}\mu_e(1+\mu_e)}{12}\right)^2.$$

These are the constants in the pathwise Poly-INF inequality used below after setting its exponent to 2.

The copy of epoch $e$ is POLY-INF$(B_e)$ started from $V=0$ and run for $n_e$ rounds

on the realized reward segment $(g_{i,t})_{t\in\mathcal{T}_e}$. The segment starts and ends at random rounds. The

inequality below is applied on each play path with the realized value of $n_e$, which is legitimate because it is a deterministic statement for every number of rounds $n \geq 1$; Theorem 2 of Audibert & Bubeck (2010) is stated for every positive integer time. Its estimates have the form $v_{i,t} = (c_t/p_{i,t})\mathbf{1}\{i = a_t\}$ with

$$c_t = -\frac{p_{a_t,t}}{\lambda_e}\log\Big(1-\frac{\lambda_e g_{a_t,t}}{p_{a_t,t}}\Big) = g_{a_t,t}\cdot\frac{-\log(1-x_t)}{x_t}, \qquad x_t := \frac{\lambda_e g_{a_t,t}}{p_{a_t,t}} \in \Big[0,\frac{1}{3\sqrt{2}}\Big],$$

and $c_t = 0$ when $g_{a_t,t} = 0$. The map $x \mapsto -\log(1-x)/x$ is increasing on $(0,1)$ and $g_{a_t,t} \leq 1$, so $0 \leq c_t \leq c_{\mathrm{AB}} < 1.1405$. Moreover,

$$2\eta_e\sqrt{\gamma_e/K} = 4\sqrt{3}\,(B_e/K)^{1/4} \geq 12\sqrt{3} > c_{\mathrm{AB}}.$$

These are the hypotheses of Audibert & Bubeck (2010, Theorem 2, Eq. (8)) with $q=2$, $\eta=\eta_e$, $\gamma=\gamma_e$, and $c=c_{\mathrm{AB}}$. That statement concerns an arbitrary finite sequence of nonnegative reals together with the probabilities that the POLY-INF recursion computes from it, holds on every play path, and imposes no lower bound on the number of rounds. It gives

$$(1-\gamma_e)\max_i V_i^{(e)} - (1+\gamma_e\xi_e)\sum_{t\in\mathcal{T}_e} p_{a_t,t}v_{a_t,t} \leq 2\eta_e\sqrt{K} = 4\sqrt{KB_e}. \tag{33}$$

This is Eq. (39) in the source's Appendix C.9 and is the only external inequality used in this epoch proof.

We next derive a pathwise bound on the learner's mixed estimate. Fix $t \in \mathcal{T}_e$ and write $p = p_{a_t,t}$, $g = g_{a_t,t}$, $x = x_t$. If $g = 0$ then $v_{a_t,t} = 0$ and (34) below is trivial, so let $g > 0$, hence $x > 0$. The map $\phi(y) = 1/y + 1/\log(1-y)$ is increasing on $(0,1)$, because $\phi'(y) \geq 0$ is equivalent, with $u = -\log(1-y) > 0$, to $2\sinh(u/2) \geq u$. Hence $\phi(x) \leq \phi(x_0) = \nu < 0.5224$. Multiplying $\phi(x) \leq \nu$ by $-g\log(1-x) \geq 0$ gives $-(g/x)\log(1-x) \leq g - \nu g\log(1-x)$. Since $p\,v_{a_t,t} = -(p/\lambda_e)\log(1-x) = -(g/x)\log(1-x)$, $-\log(1-x) = \lambda_e v_{a_t,t}$, and $g \leq 1$,

$$\sum_{i=1}^K p_{i,t}v_{i,t} = p\,v_{a_t,t} \leq g_{a_t,t} + \nu\lambda_e\, g\, v_{a_t,t} \leq g_{a_t,t} + \nu\lambda_e\sum_{i=1}^K v_{i,t}. \tag{34}$$

Summing (34) over $t \in \mathcal{T}_e$, using $\sum_i V_i^{(e)} \leq K\max_i V_i^{(e)}$, which holds because every $v_{i,t}$ is nonnegative, and inserting into (33) gives

$$a_e\max_i V_i^{(e)} \leq (1+\gamma_e\xi_e)Y_e + 4\sqrt{KB_e}, \qquad a_e := 1-\gamma_e-(1+\gamma_e\xi_e)\nu\lambda_e K. \tag{35}$$

We now bound the constants. The functions defining $\mu_e$ and $\xi_e$ are increasing in $y_e$, and $y_e \leq 1/3$ because $B_e \geq 81K$. The preceding formulas give

$$\begin{aligned}
c_{\mathrm{AB}} &= \frac{-\log(1-1/(3\sqrt{2}))}{1/(3\sqrt{2})} = 1.140412\ldots < 1.1405,\\
\nu &= 3\sqrt{2} + \frac{1}{\log(1-1/(3\sqrt{2}))} = 0.522372\ldots < 0.5224,\\
\mu_e &\leq \exp\left\{\frac{\sqrt{3}c_{\mathrm{AB}}}{3}\left(1-\frac{c_{\mathrm{AB}}}{12\sqrt{3}}\right)^{-2}\right\} = 2.089831\ldots < 2.090,\\
\xi_e &\leq \left(\frac{c_{\mathrm{AB}}(2.089831\ldots)(1+2.089831\ldots)}{12}\right)^2 = 0.376576\ldots < 0.3766.
\end{aligned}$$

The quantity $a_e$ is decreasing in $\xi_e$, $\nu$, and $s_e$, so

$$a_e \geq 1 - \frac{1}{3} - \Big(1 + \frac{0.3766}{3}\Big)\frac{0.5224}{\sqrt{2}} \cdot \frac{1}{9} > 0.6204,$$

and

$$1 + \gamma_e \xi_e - a_e = s_e\Big[3(1+\xi_e) + (1+\gamma_e\xi_e)\frac{\nu}{\sqrt{2}}\Big] \leq s_e\big[3(1.3766) + (1.1256)(0.3694)\big] \leq 4.546\, s_e.$$

Since $a_e > 0$ and $\max_i V_i^{(e)} \geq V_{k^\star}^{(e)} = G_e - D_e = Y_e + r_e - D_e$, inequality (35) gives $a_e(r_e - D_e) \leq (1 + \gamma_e\xi_e - a_e)Y_e + 4\sqrt{KB_e}$, that is,

$$r_e \leq D_e + \beta_e Y_e + \frac{4}{a_e}\sqrt{KB_e}, \qquad \beta_e := \frac{1 + \gamma_e\xi_e - a_e}{a_e} \leq \frac{4.546}{0.6204}\, s_e \leq 7.33\, s_e.$$

Insert $Y_e \leq B_e + 1$, $s_e B_e = \sqrt{KB_e}$, and $s_e \leq 1/9$,

$$r_e \leq D_e + \Big(7.33 + \frac{4}{0.6204}\Big)\sqrt{KB_e} + \frac{7.33}{9} \leq D_e + 13.78\sqrt{KB_e} + 0.815 \leq D_e + 14\sqrt{KB_e},$$

where the last step uses $\sqrt{KB_e} \geq 9K \geq 18$, so that $0.815 \leq 0.046\sqrt{KB_e}$. □

**Lemma A.17** (Monotone exponential supermartingale)**.** *For every round $t \in [T]$ and every deterministic $\theta > 0$, on the $\mathcal{F}_{t-1}$-measurable event $\{\theta \leq \lambda_{e(t)}\}$,*

$$\mathbb{E}\big[\exp\big(\theta(g_{k^\star,t} - v_{k^\star,t})\big) \,\big|\, \mathcal{F}_{t-1}\big] \leq 1.$$

*Proof.* Write $\lambda = \lambda_{e(t)}$, $p = p_{k^\star,t}$, $g = g_{k^\star,t}$, $x = \lambda g/p \in [0, 1/(3\sqrt{2})]$, and $a = \theta/\lambda$, which lies in $(0, 1]$ on the event considered. Given $\mathcal{F}_{t-1}$, the estimate $v_{k^\star,t}$ equals $-\lambda^{-1}\log(1-x)$ with probability $p$ and 0 otherwise, so

$$\mathbb{E}\big[e^{-\theta v_{k^\star,t}} \,\big|\, \mathcal{F}_{t-1}\big] = p(1-x)^a + (1-p) \leq p(1-ax) + (1-p) = 1 - \theta g \leq e^{-\theta g},$$

where the first inequality is Bernoulli's inequality $(1-x)^a \leq 1 - ax$ for $a \in (0,1]$ and $x \in [0,1)$, the concavity of $x \mapsto (1-x)^a$, and the middle equality uses $apx = \theta g$. Multiplying by $e^{\theta g}$ proves the claim. □

Let

$$\theta^\star := \lambda_{e^\star} = \frac{1}{\sqrt{2KB_{e^\star}}}, \qquad \tau := \begin{cases} t_{e^\star+1} - 1 & \text{if epoch } e^\star + 1 \text{ starts,} \\ T & \text{otherwise,} \end{cases} \tag{36}$$

and $M_0 := 1$,

$$M_t := \exp\Big(\theta^\star \sum_{s=1}^{t\wedge\tau} \big(g_{k^\star,s} - v_{k^\star,s}\big)\Big), \qquad t = 1, \ldots, T. \tag{37}$$

**Lemma A.18** (One maximal inequality)**.** *$(M_t)_{0\leq t\leq T}$ is a nonnegative supermartingale for $(\mathcal{F}_t)_{0\leq t\leq T}$ with $M_0 = 1$. Consequently, for every $\delta \in (0,1)$, the event*

$$\mathcal{E}_\delta := \Big\{\sum_{s=1}^{t}\big(g_{k^\star,s} - v_{k^\star,s}\big) \leq \sqrt{2KB_{e^\star}}\,\log(\delta^{-1}) \text{ for every } t \leq \tau\Big\} \tag{38}$$

*has probability at least $1 - \delta$.*

*Proof.* For $t < T$, the event $\{\tau \leq t\}$ occurs exactly when epoch $e^\star + 1$ starts by round $t+1$. This is determined by the rewards collected through round $t$, so the event belongs to $\mathcal{F}_t$. For $t = T$, the event is the whole sample space. Thus $\tau$ is a stopping time for $(\mathcal{F}_t)$. On $\{\tau \leq t-1\} \in \mathcal{F}_{t-1}$ we have $M_t = M_{t-1}$. On $\{\tau \geq t\}$, round $t$ lies in an epoch $e(t) \leq e^\star$, so $\{\tau \geq t\}$ is contained in $\{\theta^\star \leq \lambda_{e(t)}\}$ and Lemma A.17 with $\theta = \theta^\star$ gives $\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}$. Since $v_{k^\star,s} \geq 0$ and $g_{k^\star,s} \leq 1$, we have $0 \leq M_t \leq e^{\theta^\star T}$, so each $M_t$ is integrable and the supermartingale property holds. Let $\sigma := \min\{t \leq T : M_t \geq \delta^{-1}\}$, with $\sigma := T$ if no such $t$ exists. Then

$\sigma$ is a bounded stopping time, the optional stopping theorem gives $\mathbb{E}[M_\sigma] \le \mathbb{E}[M_0] = 1$, and $\{\max_{0\le t\le T} M_t \ge \delta^{-1}\} = \{M_\sigma \ge \delta^{-1}\}$, so Markov's inequality gives

$$\mathbb{P}\Big\{\max_{0\le t\le T} M_t \ge \delta^{-1}\Big\} \le \delta,$$

which is Ville's maximal inequality for nonnegative supermartingales. On the complement, $\theta^\star \sum_{s\le t}(g_{k^\star,s} - v_{k^\star,s}) < \log(\delta^{-1})$ for every $t \le \tau$, which is (38). □

The event $\mathcal{E}_\delta$ controls the total estimation error across the epochs up to $e^\star$. The next proof uses it at $t = T$ when $\tau = T$. There is no union bound over epochs or prefix lengths, and no continuation of an epoch beyond its end.

*Proof of Lemma A.15.* Fix $\delta \in (0,1)$ and work on the event $\mathcal{E}_\delta$ of Lemma A.18, which has probability at least $1-\delta$. On every play path $[R_T^{d'}]_+ \le U_{d'} \le U^\star$, because the learner's cumulative reward in coordinate $d'$ is nonnegative, and $R_T^{\mathrm{fs}} \le [R_T^{d'}]_+$ by (9). It remains to bound $[R_T^{d'}]_+$ on $\mathcal{E}_\delta$. If $R_T^{d'} \le 0$ there is nothing to prove, so suppose $R_T^{d'} > 0$. Then $e_{\max} \le e^\star$ by (31), epoch $e^\star + 1$ never starts, and $\tau = T$. Lemma A.16 applies to every started epoch, so by (28), (29), and $B_{e_{\max}} \le B_{e^\star}$,

$$\begin{aligned} R_T^{d'} = \sum_{e=0}^{e_{\max}} r_e &\le \sum_{t=1}^{T} \big(g_{k^\star,t} - v_{k^\star,t}\big) + 14 \sum_{e=0}^{e_{\max}} \sqrt{KB_e} \\ &\le \sqrt{2KB_{e^\star}}\,\log(\delta^{-1}) + 14(2+\sqrt{2})\sqrt{KB_{e^\star}} \le \big(47.8 + \sqrt{2}\log(\delta^{-1})\big)\sqrt{KB_{e^\star}}, \end{aligned} \tag{39}$$

where the comparator sum is the sum in (38) at $t = T = \tau$, and $14(2+\sqrt{2}) < 47.8$. With $\sqrt{KB_{e^\star}} \le \sqrt{KU_{d'}} + 9K$ from (31), on $\mathcal{E}_\delta$,

$$[R_T^{d'}]_+ \le \big(47.8 + \sqrt{2}\log(\delta^{-1})\big)\big(\sqrt{KU_{d'}} + 9K\big) \le A(\delta)\big(\sqrt{KU_{d'}} + 9K\big), \tag{40}$$

which proves (27). □

*Proof of Theorem 4.5.* Put $a := 47.8\sqrt{KB_{e^\star}}$ and $b := \sqrt{2KB_{e^\star}}$, both deterministic. On $\mathcal{E}_\delta$, a nonpositive $R_T^{d'}$ has zero positive part, while a positive $R_T^{d'}$ satisfies (39). Thus, for $x > a$, the choice $\delta := \exp(-(x-a)/b) \in (0,1)$ gives $[R_T^{d'}]_+ \le x$ on $\mathcal{E}_\delta$, and hence $\mathbb{P}\{[R_T^{d'}]_+ > x\} \le \mathbb{P}\{\mathcal{E}_\delta^c\} \le e^{-(x-a)/b}$. The layer-cake identity gives

$$\begin{aligned} \mathbb{E}\Big[[R_T^{d'}]_+\Big] &= \int_0^\infty \mathbb{P}\{[R_T^{d'}]_+ > x\}\,dx \le a + \int_a^\infty e^{-(x-a)/b}\,dx = a + b \\ &= (47.8 + \sqrt{2})\sqrt{KB_{e^\star}} < 50\big(\sqrt{KU_{d'}} + 9K\big). \end{aligned}$$

Together with $[R_T^{d'}]_+ \le U^\star$ on every play path and $R_T^{\mathrm{fs}} \le [R_T^{d'}]_+$, this gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \min\{U^\star, 50(\sqrt{KU_{d'}} + 9K)\}.$$

To obtain (11), use $U_{d'} \le U^\star$. If $U^\star \le 81K$, the minimum is at most $U^\star$, and $U^\star \le 9\min\{U^\star, \sqrt{KU^\star}\}$, because $U^\star$ is the minimum when $U^\star \le K$, while $U^\star = \sqrt{U^\star}\sqrt{U^\star} \le 9\sqrt{K}\sqrt{U^\star}$ when $K < U^\star \le 81K$. If $U^\star > 81K$, then $9K < \sqrt{KU^\star}$, so $50(\sqrt{KU_{d'}} + 9K) < 100\sqrt{KU^\star} = 100\min\{U^\star, \sqrt{KU^\star}\}$. On a reward sequence with $L^\star = L_0$ we have $U^\star = U_0 = T - L_0$ by (5), which gives the last claim. □

The absence of logarithmic factors in $K$ or $T$ comes from controlling all estimation errors together. The index $e^\star$ is deterministic, and its weight $\lambda_{e^\star}$ is valid in every earlier epoch. Thus the proof needs neither a union over arms nor a union over restart times.

*Proof of the horizon-scale Poly-INF comparison.* The two branches of Algorithm 2 with $B = T$ are handled separately. If $T \ge 81K$, the best cumulative reward in coordinate $d'$ is at most $T = B$, so (9) and (22) give $\mathbb{E}[R_T^{\mathrm{fs}}] < 10\sqrt{KT} = 10W_{K,T}$, where $W_{K,T} = \sqrt{KT}$ because $T > K$. If $T < 81K$, the algorithm plays arm 1 throughout, and the pathwise cap $R_T^{\mathrm{fs}} \le U^\star \le T$ is at most $9W_{K,T}$. Indeed, $W_{K,T} = T$ when $T \le K$, while for $K < T < 81K$, $T/W_{K,T} = \sqrt{T/K} < 9$. □

### A.7.2 UNIFORM AND FIXED-ARM PLAY

*Simple-baseline bounds.* Fix a coordinate and a deterministic best arm $k$ for that coordinate, with total reward $G^\star \le U_0$. Since every reward is nonnegative, the algorithm's collected reward is at least the reward it collects on rounds when it plays $k$. Hence, pathwise,

$$R_T^{\mathrm{fs}} \le [G^\star - A]_+ \le G^\star - \sum_t g_{k,t}\mathbf{1}\{a_t = k\}.$$

Uniform play selects $k$ with probability $1/K$ on every round, so its expected regret is at most $(1-1/K)G^\star \le (1-1/K)U_0$. The pathwise bound $R_T^{\mathrm{fs}} \le U_0$ also covers specified fixed-arm play, including the endpoint $U_0 = 0$. □

### A.7.3 EXP3.P, MO-KS, EXP3-IX, AND GREEN-IX

*High-probability part of Proposition A.9.* First consider EXP3.P. Set $\delta_0 = (T+1)^{-2}$, and use

$$\gamma = \min\left\{\frac{3}{5}, 2\sqrt{\frac{3K\log K}{5T}}\right\}, \qquad \alpha = 2\sqrt{\log(KT/\delta_0)}.$$

Theorem 6.3 of Auer et al. (2002b) bounds the realized scalar regret by

$$b_{K,T} = 4\sqrt{KT\log(KT/\delta_0)} + 4\sqrt{\frac{5}{3}KT\log K} + 8\log(KT/\delta_0)$$

with probability at least $1-\delta_0$. On that event, $R_T^{\mathrm{fs}} \le b_{K,T}$; on its complement, $R_T^{\mathrm{fs}} \le U_0 \le T$. Thus

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le (1-\delta_0)b_{K,T} + \delta_0 U_0 \le b_{K,T} + T\delta_0.$$

Combining this inequality with the pathwise bound $R_T^{\mathrm{fs}} \le U_0$ gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \min\{U_0, b_{K,T} + T\delta_0\} \lesssim \min\{U_0, \sqrt{KT\log(KT)}\}.$$

Here $q = \log(KT(T+1)^2) \le KT$ and $q \le 3\log(KT)$ for $K, T \ge 2$, so the additive $q$ term is absorbed by $\sqrt{KTq}$. The parameter $\delta_0$ is fixed before play, so this argument does not integrate bounds for different algorithms. The adversarial branch of MO-KS is this scalar EXP3.P procedure on a fixed coordinate (Xu & Klabjan, 2023, Algorithm 1); choosing its parameters as above gives the same conclusion. Randomizing the coordinate before play also preserves this bound by conditioning on that choice.

For EXP3-IX, set $\eta = 2\gamma = \sqrt{2\log K/(KT)}$, use the loss estimate

$$\widetilde{\ell}_{i,t} = \frac{\ell_{i,t}\mathbf{1}\{a_t = i\}}{p_{i,t} + \gamma}, \qquad p_{i,t} \propto \exp\left(-\eta \sum_{s<t} \widetilde{\ell}_{i,s}\right).$$

Theorem 1 of Neu (2015a) gives, for the same algorithm at every $\delta \in (0,1)$,

$$\mathbb{P}\{R_T^{d'} > a + b\log(2/\delta)\} \le \delta, \quad a = 2\sqrt{2KT\log K}, \qquad b = \sqrt{\frac{2KT}{\log K}} + 1.$$

Put $y = a + b\log 2$. For $x > y$, choose $\delta = 2e^{-(x-a)/b} \in (0,1)$. The preceding probability bound then gives $\mathbb{P}\{R_T^{d'} > x\} \le e^{-(x-y)/b}$. Therefore

$$\mathbb{E}[[R_T^{d'}]_+] \le y + \int_y^\infty e^{-(x-y)/b}\,dx = a + (1+\log 2)b.$$

For $K, T \ge 2$ this is $O(\sqrt{KT\log K})$. Combining it with the bound by $U_0$ proves the EXP3-IX bound in Proposition A.9.

For GREEN-IX, apply Corollary 4.2 of Lykouris et al. (2022) to the losses in coordinate $d'$, with $\delta_0 = (T+1)^{-2}$. Its number of arms, denoted $d$ in that source, is our $K$. The adaptive loss-doubling procedure gives, with probability at least $1-\delta_0$,

$$R_T^{d'} \lesssim \sqrt{KL_{d'}\log(KT)} + K\log(KT),$$

where $\log(K/\delta_0) = \log(K(T+1)^2) \le 3\log(KT)$ for $K, T \ge 2$. The loss-doubling overhead in Lemma 3.8 of the same source enters through $\log(1+\log(L_{d'}+1))$, which is also absorbed into $O(\log(KT))$ since $L_{d'} \le T$. The procedure does not need $L_{d'}$ or $L_0$ as an input. Let $b_G$ denote the displayed right side, including the numerical constant in the cited result. On the event, which has probability at least $1-\delta_0$, $R_T^{\mathrm{fs}} \le [R_T^{d'}]_+ \le b_G$; on the complement, $R_T^{\mathrm{fs}} \le U_0$. Hence

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \min\{U_0, b_G + U_0\delta_0\}.$$

Since $U_0\delta_0 \le T/(T+1)^2 < 1$, this proves the GREEN-IX bound in Proposition A.9. Taking a positive part introduces no extra loss-level dependence. □

### A.7.4 GAIN-BASED EXP3

The calculation in Appendix A.6.6 proves (24) for every supplied $V \ge U_0$. Choosing $V = T$ gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\{U_0, (K\log K)^{1/3}T^{2/3}\},$$

which is the gain-based EXP3 statement in Proposition A.9.

### A.7.5 ORIGINAL EXP3.1

EXP3.1 (Auer et al., 2002b) restarts the gain-based EXP3 weights, but retains its cumulative reward estimates across epochs. For completeness, set

$$c = \frac{K\log K}{e-1}, \qquad b_r = c4^r, \qquad \gamma_r = 2^{-r}.$$

Epoch $r$ starts fresh EXP3 weights with parameter $\gamma_r$, but keeps the cumulative importance-weighted reward estimates $\widehat{G}_i$ from all earlier epochs, initially zero at $r = 0$. Before every action, including the first action of an epoch, check whether $\max_i \widehat{G}_i \le b_r - K/\gamma_r$. If this fails, increment $r$, reset the weights, and repeat the check without taking an action or changing the retained estimates. Otherwise, take one EXP3 action and add its importance-weighted reward to $\widehat{G}_i$. This is the pre-action check in Auer et al. (2002b, Figure 2); it permits empty epochs. The retained estimates, rather than collected rewards, determine the restart times.

*Proof of the EXP3.1 part of Proposition A.9.* Write $\ell = \log K$, $c = K\ell/(e-1)$. All the inequalities in the first part of the proof hold on every realized history and at every prefix of that history. Let $r$ be the epoch of the last completed action, $x = \max_i \widehat{G}_i$, and $A$ the collected reward at that prefix.

Within a nonempty epoch $s$, every action begins with $\widehat{G}_i \le b_s - K/\gamma_s$. Since its estimate increment is at most $K/\gamma_s$, every global estimate immediately after that action is at most $b_s$, including at the end of the epoch or any prefix of it. Nonnegativity gives the same bound for the epoch increment, denoted $\widehat{G}_{i,s}$. The EXP3 potential argument above, with $p_{i,t}\widehat{g}_{i,t}^2 \le \widehat{g}_{i,t}$, gives for every arm $j$

$$A_s \ge \widehat{G}_{j,s} - (e-1)\gamma_s b_s - K\ell/\gamma_s = \widehat{G}_{j,s} - 2K\ell\, 2^s.$$

Empty epochs satisfy this inequality as well. This reproduces the epoch inequality of Auer et al. (2002b, Lemma 4.2) and also shows explicitly why it holds for a truncated last epoch. Summing over epochs and maximizing over $j$ yields

$$A \ge x - 4K\ell\, 2^r. \tag{41}$$

If $r \ge 1$, leaving epoch $r-1$ required $x > c4^{r-1} - K2^{r-1}$. Put $y = 2^r$. This inequality is $cy^2 - 2Ky - 4x < 0$, so $y$ is smaller than its positive root:

$$y < \frac{K + \sqrt{K^2 + 4cx}}{c} \le \frac{2K}{c} + 2\sqrt{\frac{x}{c}},$$

where the last inequality uses $\sqrt{K^2+4cx} \le K + 2\sqrt{cx}$. When $r = 0$, $2^r = 1$. Combining the two cases gives

$$2^r \le 1 + 2K/c + 2\sqrt{x/c}. \tag{42}$$

Consequently

$$A \ge x - a\sqrt{x} - b, \qquad a = 8\sqrt{(e-1)K\ell}, \quad b = 4K\ell + 8(e-1)K. \tag{43}$$

Indeed, substituting (42) into (41) gives

$$A \ge x - 4K\ell - \frac{8K^2\ell}{c} - \frac{8K\ell}{\sqrt{c}}\sqrt{x},$$

and $c = K\ell/(e-1)$ gives the displayed values of $a$ and $b$. For $0 \le y \le x$, $x - a\sqrt{x} \ge y - a\sqrt{y} - a^2/4$. To see this, write $x = y + z$ with $z \ge 0$, use $\sqrt{y+z} \le \sqrt{y} + \sqrt{z}$, and note that

$$z - a\sqrt{z} = (\sqrt{z} - a/2)^2 - a^2/4 \ge -a^2/4.$$

Thus (43) implies, for each fixed comparator $k$,

$$A \ge \widehat{G}_k - a\sqrt{\widehat{G}_k} - b - a^2/4. \tag{44}$$

We next bound the estimation error without assuming that the algorithm knows $U_0$. Fix a deterministic best arm $k$, and let $\tau$ be the first round at which collected reward reaches $G^\star$, or $T$ if this never occurs. If $G^\star = 0$, nonnegative rewards imply that every arm receives zero reward on every round, so $R_T^{\mathrm{fs}} = 0$. We therefore assume $G^\star > 0$. This stopping time is used only in the proof. Nonnegative rewards imply

$$[G^\star - A_T]_+ = [\sum_{t \le \tau} g_{k,t} - A_\tau]_+, \qquad A_\tau \le U_0 + 1.$$

Indeed, if the threshold is reached, then $A_\tau \ge G^\star$ and $A_T \ge A_\tau$, so both positive parts are zero. If it is not reached, then $\tau = T$, and the two expressions are identical. Moreover, immediately before a crossing the collected reward is smaller than $G^\star \le U_0$, and the crossing reward is at most one; hence $A_\tau \le U_0 + 1$. From (43) and $a\sqrt{x} \le (x + a^2)/2$, at this stopped prefix

$$x \le H := 2(U_0 + 1) + a^2 + 2b.$$

Epoch indices only increase. Hence every action up to $\tau$ occurs in an epoch $u \le r$, and its sampling probability satisfies $p_{k,t} \ge \gamma_u/K = 1/(K2^u) \ge 1/(K2^r)$. By (42), $2^r \le s$, where the deterministic quantity $s$ is

$$s = 1 + 2K/c + 2\sqrt{H/c}.$$

The stopped estimation error $M_\tau = \sum_{t \le \tau}(g_{k,t} - \widehat{g}_{k,t})$ has mean zero. Its increments include the predictable indicator $\mathbf{1}\{t \le \tau\}$, so conditional unbiasedness and orthogonality give

$$\mathbb{E}[M_\tau^2] \le Ks \sum_t g_{k,t}^2 \le KsU_0, \qquad \mathbb{E}[\widehat{G}_{k,\tau}] = \mathbb{E}\left[\sum_{t \le \tau} g_{k,t}\right] \le U_0.$$

At the stopped prefix, (44) gives

$$\sum_{t \le \tau} g_{k,t} - A_\tau \le M_\tau + a\sqrt{\widehat{G}_{k,\tau}} + b + a^2/4.$$

Taking positive parts and expectations, using $\mathbb{E}[[M_\tau]_+] \le \frac{1}{2}\sqrt{\mathbb{E}[M_\tau^2]}$, and applying Jensen's inequality to $\sqrt{\widehat{G}_{k,\tau}}$ proves

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \min\{U_0, a\sqrt{U_0} + b + a^2/4 + \tfrac{1}{2}\sqrt{KsU_0}\}. \tag{45}$$

For $K \ge 2$, $\ell = \log K \ge \log 2$, $a^2 = 64(e-1)K\ell$, and $b = 4K\ell + 8(e-1)K = O(K\ell)$. Consequently,

$$H = O(U_0 + K\ell), \qquad s = 1 + \frac{2(e-1)}{\ell} + 2\sqrt{\frac{H(e-1)}{K\ell}} = O\left(1 + \sqrt{\frac{U_0}{K\ell}}\right).$$

Since $s = O(1 + \sqrt{U_0/(K\ell)})$, taking square roots gives $\sqrt{s} = O(1 + (U_0/(K\ell))^{1/4})$. Therefore

$$\sqrt{KsU_0} = O\left(\sqrt{KU_0} + \left(\frac{KU_0^3}{\ell}\right)^{1/4}\right).$$

The term $a\sqrt{U_0}$ is $O(\sqrt{KU_0\ell})$, while $b + a^2/4 = O(K\ell)$. Since $\sqrt{KU_0} \le (\log 2)^{-1/2}\sqrt{KU_0\ell}$, substitution in (45) gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \lesssim \min\left\{U_0, \sqrt{KU_0 \log K} + \left(\frac{KU_0^3}{\log K}\right)^{1/4} + K \log K\right\},$$

which is the EXP3.1 bound in Proposition A.9. This proof concerns the original EXP3.1 procedure, not reward-doubling Poly-INF and not a modified EXP3 estimator. □

### A.7.6 TSALLIS-INF, EXP3++, AND MO-US

*Proof of Proposition A.10.* Use Tsallis-INF with parameter $\alpha = 1/2$, symmetric regularization, and the estimators in Zimmert & Seldin (2021, Theorem 1). Its learning rates are $2/\sqrt{t}$ for importance-weighted (IW) estimates and $4/\sqrt{t}$ for reduced-variance (RV) estimates. For deterministic losses the cited pseudo-regret bounds are signed expected regret bounds. Write $\overline{B}_{\mathcal{A}}$ for the following explicit bounds. For Tsallis-INF,

$$\overline{B}_{\mathrm{IW}}(n) = 4\sqrt{Kn} + 1,$$
$$\overline{B}_{\mathrm{RV}}(n) = 2\sqrt{Kn} + 10K \log n + 16.$$

Lemma A.12 proves the stated positive-regret bounds for these unchanged algorithms.

For EXP3++, Seldin & Lugosi (2017, Theorem 1) gives $\overline{B}_{++}(n) = 4\sqrt{Kn \log K}$. The assumptions on its exploration parameters are important. Set

$$\eta_t = \tfrac{1}{2}\sqrt{\log K/(Kt)}, \quad q_{i,t} \propto e^{-\eta_t \widehat{L}_{i,t-1}}, \quad p_{i,t} = (1-\epsilon_t)q_{i,t} + \epsilon_{i,t}, \quad \epsilon_t = \sum_i \epsilon_{i,t},$$

where $0 \le \epsilon_{i,t} \le \min\{1/(2K), \eta_t\}$ is determined from past observations. Its loss estimate is $\widehat{\ell}_{i,t} = \ell_{i,t}\mathbf{1}\{a_t = i\}/p_{i,t}$. The adversarial argument does not require the empirical gaps used to choose these exploration probabilities to be accurate.

For completeness, this argument also covers the initialization used by MO-US (Xu & Klabjan, 2023, Algorithm 2). Charge at most $K$ regret for its initial pulls (or $T$ if $T < K$). Its initial cumulative estimates lie in $[0, 1]$, so their range costs at most one additional unit in the exponential-weights comparison. We now derive the needed inequality on the remaining rounds.

Let $H_{i,0} \in [0, 1]$ be the estimate after initialization, let $H_{i,t} = H_{i,0} + \sum_{s=1}^{t} \widehat{\ell}_{i,s}$, and define

$$F_t := -\frac{1}{\eta_t} \log\left(\frac{1}{K}\sum_i e^{-\eta_t H_{i,t-1}}\right).$$

We set $\eta_{T+1} = \eta_T$ when defining the potential after the final update. Conditional on the initialization, $H_{i,0}$ is fixed. At a fixed value of $\eta_t$, put

$$u_t := \eta_t\langle q_t, \widehat{\ell}_t\rangle - \frac{\eta_t^2}{2}\langle q_t, \widehat{\ell}_t^2\rangle.$$

The inequality $e^{-x} \le 1 - x + x^2/2$ gives $\sum_i q_{i,t}e^{-\eta_t\widehat{\ell}_{i,t}} \le 1 - u_t$. The sum on the left is positive, so $u_t < 1$, and $-\log(1 - u_t) \ge u_t$. Consequently,

$$-\frac{1}{\eta_t}\log\sum_i q_{i,t}e^{-\eta_t\widehat{\ell}_{i,t}} \ge \langle q_t, \widehat{\ell}_t\rangle - \frac{\eta_t}{2}\langle q_t, \widehat{\ell}_t^2\rangle.$$

For a fixed nonnegative vector $H$, the function

$$\eta \longmapsto -\frac{1}{\eta}\log\left(\frac{1}{K}\sum_i e^{-\eta H_i}\right)$$

is nonincreasing. One way to see this is to write the derivative as $-\mathrm{KL}(q_\eta \| \mathrm{Unif}[K])/\eta^2 \le 0$, where $q_{\eta,i} \propto e^{-\eta H_i}$. Since $\eta_{t+1} \le \eta_t$, replacing $\eta_t$ by $\eta_{t+1}$ after the update can only increase $F_{t+1}$. Therefore

$$\langle q_t, \widehat{\ell}_t \rangle \le F_{t+1} - F_t + \frac{\eta_t}{2} \langle q_t, \widehat{\ell}_t^2 \rangle.$$

Summing this inequality telescopes. Moreover,

$$F_{T+1} \le H_{k,T} + \frac{\log K}{\eta_T}, \qquad F_1 \ge \min_i H_{i,0},$$

so $H_{k,0} - F_1 \le 1$. Subtracting $\sum_t \widehat{\ell}_{k,t} = H_{k,T} - H_{k,0}$ gives

$$\sum_t \langle q_t, \widehat{\ell}_t \rangle - \sum_t \widehat{\ell}_{k,t} \le 1 + \frac{\log K}{\eta_T} + \frac{1}{2} \sum_t \eta_t \sum_i q_{i,t} \widehat{\ell}_{i,t}^2.$$

Since $\epsilon_t \le 1/2$, $p_{i,t} \ge q_{i,t}/2$. Thus the conditional expectation of the quadratic term is

$$\mathbb{E}_t \left[ \sum_i q_{i,t} \widehat{\ell}_{i,t}^2 \right] = \sum_i \frac{q_{i,t} \ell_{i,t}^2}{p_{i,t}} \le 2K.$$

Moreover $\mathbb{E}_t[\ell_{a_t,t} - \langle q_t, \widehat{\ell}_t \rangle] = \langle p_t - q_t, \ell_t \rangle \le \epsilon_t \le K\eta_t$. The first inequality follows from $p_{i,t} - q_{i,t} = \epsilon_{i,t} - \epsilon_t q_{i,t}$ and $\ell_{i,t} \in [0,1]$. Unbiasedness gives $\mathbb{E}_t[\widehat{\ell}_{k,t}] = \ell_{k,t}$. Finally, $\sum_{t \le T} t^{-1/2} \le 2\sqrt{T}$, and hence

$$\sum_{t \le T} \eta_t \le \sqrt{\frac{T \log K}{K}}.$$

Taking expectations in the potential inequality and adding the sampling difference therefore gives

$$\mathbb{E}[R_T^{d'}] \le \overline{B}_{\mathrm{MO}}(T) := 4\sqrt{KT \log K} + K + 1.$$

This also covers any of MO-US's empirical exploration choices satisfying the displayed bounds. A zero denominator in the exploration expression is interpreted as $+\infty$ before taking its minimum, so the probabilities remain well defined. Apply Lemma A.12 once more. Substituting the four explicit bounds gives the corresponding statements in Proposition A.10 for $K \ge 2, n \ge 1$. This argument does not replace $L_{d'}$ by $L_0$. □

### A.7.7 Independent restarts in fixed blocks

**Lemma A.19** (Positive regret after independent block restarts)**.** *Suppose $\mathcal{A}$ has signed expected regret at most $B_{\mathcal{A}}(n) \ge 0$ on every deterministic scalar sequence of length $n$. For the block lengths $n_1, \dots, n_m$ in Algorithm 6,*

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \min \left\{ U_0, \sum_{j=1}^m B_{\mathcal{A}}(n_j) + \frac{1}{2} \sqrt{\min \left\{ \frac{1}{4} \sum_j n_j^2, hKU_0 \right\}} \right\}. \tag{46}$$

*Proof.* Fix a best scalar arm $k$ for the entire sequence. Let $A_j$ be the collected reward in block $j$, $G_{k,j}$ that arm's reward there, and $X_j = G_{k,j} - A_j$. Each $\mathbb{E}[X_j]$ is at most $B_{\mathcal{A}}(n_j)$, since a block's best arm earns at least $G_{k,j}$. The blocks and reward sequence are fixed, and the copies use independent random choices without shared state. Hence the $X_j$ are independent, even though actions within a block are adaptive. For $X = \sum_j X_j$, put $Y = X - \mathbb{E}[X]$. Since $Y$ has mean zero, $\mathbb{E}[[Y]_+] = \mathbb{E}[|Y|]/2 \le \sqrt{\mathrm{Var}(X)}/2$. Also $\mathrm{Var}(X) = \sum_j \mathrm{Var}(X_j) = \sum_j \mathrm{Var}(A_j)$, because the $G_{k,j}$ are deterministic. Therefore

$$\mathbb{E}[[X]_+] \le [\mathbb{E}[X]]_+ + \mathbb{E}[[X - \mathbb{E}[X]]_+] \le \sum_j B_{\mathcal{A}}(n_j) + \tfrac{1}{2} \sqrt{\sum_j \mathrm{Var}(A_j)}.$$

Since $A_j \in [0, n_j]$, $\mathrm{Var}(A_j) \le n_j^2/4$. Also $A_j^2 \le hA_j$, and pathwise $\sum_j A_j \le \sum_{i,t} g_{i,t} \le KU_0$. Therefore

$$\sum_j \mathrm{Var}(A_j) \le \sum_j \mathbb{E}[A_j^2] \le h\,\mathbb{E}\left[\sum_j A_j\right] \le hKU_0.$$

Combining the two variance bounds and $R_T^{\mathrm{fs}} \le \min\{U_0, [X]_+\}$ proves the result. □

*Proof of Proposition A.11.* Use the four explicit bounds $\overline{B}_{\mathcal{A}}$ established in Section A.7.6 as $B_{\mathcal{A}}$ in Lemma A.19. Let $h = \min\{T, \lceil\sqrt{KT}\rceil\}$. If $K \ge T$, $K^{1/4}T^{3/4} \ge T$, and the bound $U_0$ suffices. For $K < T$, there are at most $T/h + 1$ blocks, and

$$\sum_j \sqrt{Kn_j} \le \sqrt{K}(T/\sqrt{h} + \sqrt{h}), \qquad \sum_j \mathrm{Var}(A_j) \le hT/4.$$

Here

$$\sqrt{KT} \le h \le \sqrt{KT} + 1 \le 2\sqrt{KT}.$$

The lower bound on $h$ gives

$$\sqrt{K}\,\frac{T}{\sqrt{h}} \le K^{1/4}T^{3/4}.$$

The upper bound on $h$, together with $K < T$, gives

$$\sqrt{Kh} \le \sqrt{2}\,K^{3/4}T^{1/4} \le \sqrt{2}\,K^{1/4}T^{3/4}, \qquad \sqrt{hT} \le \sqrt{2}\,K^{1/4}T^{3/4}.$$

Thus the square-root terms in (46) are $O(K^{1/4}T^{3/4})$.

The number of blocks is at most

$$\frac{T}{h} + 1 \le \sqrt{\frac{T}{K}} + 1 \le 2\sqrt{\frac{T}{K}}.$$

For IW, the additive constant in each block therefore contributes at most $O(\sqrt{T/K})$, which is $O(K^{1/4}T^{3/4})$. For RV and MO-US, the non-square-root terms sum to

$$O\left(K\log(eT)\left(\frac{T}{h} + 1\right)\right) = O(\sqrt{KT}\log(eT)) = O(K^{1/4}T^{3/4}\log(2T)).$$

For EXP3++, multiplying the square-root calculation by $\sqrt{\log K}$ gives $O(K^{1/4}T^{3/4}\sqrt{\log K})$. Substitution into (46) proves all four bounds in Proposition A.11. These statements concern the independently restarted copies in Algorithm 6. □

For a fixed weighted average, replace each scalar reward by $g_{i,t} = \sum_d w_d r_d^{i,t}$. The same scalar proofs apply, with best reward at most $U_0$. For GREEN-IX the relevant loss is $\min_i \sum_t (1 - g_{i,t})$, which need not equal $L^\star$. Thus averaging is valid for the algorithms with a positive-regret guarantee. This is not a consequence of a signed expected-regret bound. If all coordinates coincide, Pareto regret is already the positive part of scalar regret, so that distinction remains necessary.

### A.8 Proof of the minimax theorem

*Proof of Theorem 4.6.* For the exact class $L^\star = L_0$, Theorem 3.2 applies to every algorithm, including algorithms that know $L_0$. It gives

$$\mathcal{V}_{K,D,T}(L_0) \ge 2^{-27}\min\{U_0, \sqrt{KU_0}\}.$$

Algorithm 2 may use $L_0$, and Theorem 4.3 gives

$$\mathcal{V}_{K,D,T}(L_0) \le 10\min\{U_0, \sqrt{KU_0}\}.$$

Together these inequalities prove (12). Algorithm 5 does not know $L_0$, and Theorem 4.5 gives the uniform classwise bound

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le 100\min\{U_0, \sqrt{KU_0}\}$$

for every $L_0$, so the same order is attainable without this input.

At $L_0 = 0, U_0 = T$, and the preceding two-sided bound is $\Theta(W_{K,T})$. For the unrestricted class of reward sequences, the zero-loss class is a subset, so Theorem 3.1 gives the same lower bound. For every unrestricted sequence, $0 \le U^\star \le T$, and the function $u \mapsto \min\{u, \sqrt{Ku}\}$ is nondecreasing. Applying Theorem 4.5 with $U^\star = T - L^\star$ therefore gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le 100 \min\{U^\star, \sqrt{KU^\star}\} \le 100 \min\{T, \sqrt{KT}\} = 100 W_{K,T}.$$

This proves the unrestricted upper bound. The displayed numerical constants are not claimed to be sharp. □

### A.9 Coordinate-dependent refinements of the upper bounds

Theorem 4.6 matches the upper and lower bounds after taking the worst case over sequences with $L^\star = L_0$. It does not assert tightness on every sequence. We now retain the chosen coordinate's loss $L_d$ to obtain finer upper bounds. Throughout, $d$ is fixed before play, and the existing notation gives $U_d = T - L_d \le U_0 = T - L_0$.

**Corollary A.20** (Coordinate-dependent bounds)**.** *Fix a deterministic reward sequence of length $T \ge 1$. For any policy and any coordinate $d$,*

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \mathbb{E}[[R_T^d]_+] \le \min\left\{T - L_d,\ \frac{T - L_d}{T}\left(L_d + \mathbb{E}[R_T^d]\right)\right\}.$$

*Proof.* The learner's cumulative loss $R_T^d + L_d$ lies in $[0, T]$, so

$$[R_T^d]_+ \le \frac{T - L_d}{T}(R_T^d + L_d), \qquad [R_T^d]_+ \le T - L_d.$$

Taking expectations and applying Lemma 4.1 proves the claim. □

**Corollary A.21** (Refinements of Table 1)**.** *For $K, T \ge 2$ and $D \ge 1$, the policies in Table 1, run on coordinate $d$ with their stated information and parameters, satisfy the following bounds. Every minimum combines guarantees for the same policy.*

| Method | Upper bound on $\mathbb{E}[R_T^{\mathrm{fs}}]$ (up to constants) |
|---|---|
| Poly-INF | $\min\{U_d, \sqrt{KU_0}\}$ |
| Modified EXP3 | $\min\{U_d, \sqrt{KU_0 \log(3K)}\}$ |
| Gain-based EXP3 | $\min\{U_d, (K \log K)^{1/3} V^{2/3}\}$, $V = U_0$ or $T$ |
| Reward-doubling Poly-INF | $\min\{U_d, \sqrt{KU_d}\}$ |
| Poly-INF with $B = T$ | $\min\{U_d, \sqrt{KT}\}$ |
| EXP3.P and adversarial MO-KS | $\min\{U_d, \sqrt{KT \log(KT)}\}$ |
| EXP3-IX | $\min\{U_d, \sqrt{KT \log K}\}$ |
| GREEN-IX | $\min\{U_d, \sqrt{KL_d \log(KT)} + K \log(KT)\}$ |
| Original EXP3.1 | $\min\{U_d, \sqrt{KU_d \log K} + (KU_d^3/\log K)^{1/4} + K \log K\}$ |
| Tsallis-INF (IW) | $\min\{U_d, \sqrt{KT} + S_d,$ $\frac{U_d}{T}(L_d + \sqrt{KT})\}$ |
| Tsallis-INF (RV) | $\min\{U_d, \sqrt{KT} + K \log T + S_d,$ $\frac{U_d}{T}(L_d + \sqrt{KT} + K \log T)\}$ |
| EXP3++ | $\min\{U_d, \sqrt{KT \log K} + S_d,$ $\frac{U_d}{T}(L_d + \sqrt{KT \log K})\}$ |
| MO-US | $\min\{U_d, \sqrt{KT \log K} + K + S_d,$ $\frac{U_d}{T}(L_d + \sqrt{KT \log K} + K)\}$ |
| Fixed-block Tsallis-INF (IW) | $\min\{U_d, K^{1/4} T^{3/4}\}$ |
| Fixed-block Tsallis-INF (RV) and MO-US | $\min\{U_d, K^{1/4} T^{3/4} \log(2T)\}$ |
| Fixed-block EXP3++ | $\min\{U_d, K^{1/4} T^{3/4} \sqrt{\log K}\}$ |

*Proof.* Fix the coordinate $d$ used by the policy. Its scalar regret is

$$R_T^d = U_d - \sum_{t=1}^{T} r_d^{a_t,t}.$$

The collected reward is nonnegative, so Lemma 4.1 gives, on every play path,

$$R_T^{\mathrm{fs}} \le [R_T^d]_+ \le U_d.$$

When $U_d = 0$, the cap proves every row. Assume $U_d > 0$ for the remaining argument. The cap can be combined with any other upper bound for the same policy: if $\mathbb{E}[R_T^{\mathrm{fs}}] \le Cb$, then $\mathbb{E}[R_T^{\mathrm{fs}}] \le \min\{U_d, Cb\} \le \max\{1, C\} \min\{U_d, b\}$.

For known-scale Poly-INF and Poly-INF with $B = T$, take $B = U_0$ and $B = T$, respectively. If $B \ge 81K$, then $T \ge B \ge K$ and the best scalar reward satisfies $U_d \le B$. Equation (22) therefore gives $\mathbb{E}[[R_T^d]_+] < 10\sqrt{KB}$. If $B < 81K$, the fixed-arm branch instead satisfies $R_T^{\mathrm{fs}} \le U_d \le B < 9\sqrt{KB}$. Combining these bounds with the cap proves both Poly-INF rows. For modified EXP3, the proof in Appendix A.6.5 gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le (3 + \sqrt{2})\sqrt{KU_0 \log(3K)}.$$

The gain-based EXP3 calculation in Appendix A.6.6 gives $O((K \log K)^{1/3} V^{2/3})$ for the stated choice $V = U_0$ or $V = T$. Likewise, Lemma A.14 and Appendix A.7.3 give the displayed learning terms for EXP3.P, adversarial MO-KS, and EXP3-IX. Adding the cap $U_d$ proves these rows. In particular, this argument does not replace $U_0$ by $U_d$ in the parameters of a known-scale algorithm.

For GREEN-IX, the proof in Appendix A.7.3 uses the fixed confidence level $\delta_0 = (T+1)^{-2}$. On its successful event,

$$[R_T^d]_+ \le C\big(\sqrt{KL_d \log(KT)} + K \log(KT)\big)$$

for an absolute constant $C$; on the complementary event it is at most $U_d$. Splitting the expectation between these two events gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le \min\left\{U_d,\, C\big(\sqrt{KL_d \log(KT)} + K \log(KT)\big) + \frac{U_d}{(T+1)^2}\right\}.$$

The last term is smaller than one because $U_d \le T$, and is absorbed by $K \log(KT)$. This proves the GREEN-IX row without requiring the policy to know $L_d$.

Reward-doubling Poly-INF permits a stronger refinement because its positive-regret analysis already retains $U_d$. The tail integration in the proof of Theorem 4.5 gives

$$\mathbb{E}[[R_T^d]_+] < 50\big(\sqrt{KU_d} + 9K\big).$$

If $U_d \le K$, the cap equals $\min\{U_d, \sqrt{KU_d}\}$. If $K < U_d \le 81K$, then $U_d \le 9\sqrt{KU_d}$, so the cap is at most nine times that minimum. Finally, if $U_d > 81K$, then $9K < \sqrt{KU_d}$, and the preceding expectation bound is smaller than $100\sqrt{KU_d}$. Thus, in every case,

$$\mathbb{E}[R_T^{\mathrm{fs}}] \le 100 \min\{U_d, \sqrt{KU_d}\}.$$

Here $U_d$ enters only the analysis; the restart rule remains Algorithm 5.

For original EXP3.1, use the notation of Appendix A.7.5, and let $k$ be a deterministic best arm in coordinate $d$, so that $G^\star = U_d$. Stop at the first round $\tau$ when collected reward reaches $U_d$, or at $T$ if this never occurs. Nonnegative rewards imply

$$[R_T^d]_+ = [\sum_{t \le \tau} g_{k,t} - A_\tau]_+, \qquad A_\tau \le U_d + 1.$$

If the threshold is reached, both positive parts are zero; otherwise $\tau = T$. Thus the prefix estimates in that appendix, with its unchanged constants $a, b, c$, give

$$H = 2(U_d + 1) + a^2 + 2b, \qquad s = 1 + \frac{2K}{c} + 2\sqrt{\frac{H}{c}}.$$

Up to $\tau$, the proof in Appendix A.7.5 gives $p_{k,t} \geq 1/(Ks)$. The comparator's total reward is $U_d$, so its reward accumulated up to $\tau$ is at most $U_d$. The stopped estimation error is a mean-zero martingale sum, and these bounds give

$$\mathbb{E}[M_\tau^2] \leq KsU_d, \qquad \mathbb{E}[\widehat{G}_{k,\tau}] \leq U_d.$$

Taking positive parts in (44), using $\mathbb{E}[[M_\tau]_+] \leq \frac{1}{2}\sqrt{\mathbb{E}[M_\tau^2]}$, and applying Jensen's inequality therefore yields

$$\mathbb{E}[R_T^{\mathrm{fs}}] \leq \min\left\{U_d,\ a\sqrt{U_d} + b + \frac{a^2}{4} + \frac{1}{2}\sqrt{KsU_d}\right\}.$$

Since $a = O(\sqrt{K\log K})$, $b = O(K\log K)$, and $c = K\log K/(e-1)$, we have

$$H = O(U_d + K\log K), \qquad s = O\left(1 + \sqrt{\frac{U_d}{K\log K}}\right),$$

and taking square roots in the bound on $s$ gives

$$\sqrt{KsU_d} = O\left(\sqrt{KU_d} + \left(\frac{KU_d^3}{\log K}\right)^{1/4}\right).$$

Substitution gives the stated EXP3.1 row. The stopping time and these deterministic bounds are proof devices; no threshold in the implemented EXP3.1 policy is changed.

For the four methods with signed expected-regret guarantees, Appendix A.7.6 establishes $\mathbb{E}[R_T^d] \leq \overline{B}_{\mathcal{A}}(T)$, where its explicit bounds are

$$\begin{aligned}
\overline{B}_{\mathrm{IW}}(T) &= 4\sqrt{KT} + 1,\\
\overline{B}_{\mathrm{RV}}(T) &= 2\sqrt{KT} + 10K\log T + 16,\\
\overline{B}_{++}(T) &= 4\sqrt{KT\log K},\\
\overline{B}_{\mathrm{MO}}(T) &= 4\sqrt{KT\log K} + K + 1.
\end{aligned}$$

A signed expectation alone does not bound the expectation of the positive part. For every play path,

$$-R_T^d = \sum_t r_d^{a_t,t} - U_d \leq \sum_t \max_i r_d^{i,t} - U_d = S_d.$$

Since $S_d \geq 0$ and $[R_T^d]_+ = R_T^d + [-R_T^d]_+$, it follows that

$$\mathbb{E}[[R_T^d]_+] \leq \overline{B}_{\mathcal{A}}(T) + S_d.$$

For the second conversion, apply Corollary A.20 and the same signed expectation bound to obtain

$$\mathbb{E}[[R_T^d]_+] \leq \frac{U_d}{T}\big(L_d + \overline{B}_{\mathcal{A}}(T)\big).$$

Together with the cap they give the explicit common form

$$\mathbb{E}[R_T^{\mathrm{fs}}] \leq \min\left\{U_d,\ \overline{B}_{\mathcal{A}}(T) + S_d,\ \frac{U_d}{T}\big(L_d + \overline{B}_{\mathcal{A}}(T)\big)\right\}.$$

Substituting the four bounds above and absorbing numerical constants for $K, T \geq 2$ proves the Tsallis-INF, EXP3++, and MO-US rows.

Finally, consider the independently restarted policies. In Lemma A.19, each arm's total scalar reward is at most $U_d$, so the same proof gives

$$\mathbb{E}[R_T^{\mathrm{fs}}] \leq \min\left\{U_d,\ \sum_j \overline{B}_{\mathcal{A}}(n_j) + \frac{1}{2}\sqrt{\min\left\{\frac{1}{4}\sum_j n_j^2,\ hKU_d\right\}}\right\}.$$

The independence used here follows from the fixed blocks and fresh randomness; no independence of actions within a block is assumed. Keep the prescribed block length $h = \min\{T, \lceil\sqrt{KT}\rceil\}$. For $K \geq T$, the cap already implies all the displayed rates up to constants. For $K < T$,

$$\sqrt{KT} \leq h \leq 2\sqrt{KT}, \qquad \sum_j \sqrt{Kn_j} \leq \sqrt{K}\left(\frac{T}{\sqrt{h}} + \sqrt{h}\right) \leq (1+\sqrt{2})K^{1/4}T^{3/4}.$$

Also $\sum_j n_j^2 \leq hT$, so the variance contribution is at most $\frac{1}{4}\sqrt{hT} = O(K^{1/4}T^{3/4})$. There are at most $2\sqrt{T/K}$ blocks. Thus the additive constants in the IW bound are absorbed by this rate, whereas the additional terms for RV and MO-US sum to at most $O(\sqrt{KT}\log(2T))$, which is $O(K^{1/4}T^{3/4}\log(2T))$. The EXP3++ bound contributes the factor $\sqrt{\log K}$ to the square-root calculation. These substitutions prove the final three rows with the stated block length. □

*Remark* A.22 (Comparison with the uniform minimax bound). The following comparisons concern the displayed rate expressions, with numerical constants suppressed. For known-scale Poly-INF, the minimum is $\min\{T - L_d, \sqrt{K(T-L_0)}\}$. Its cap is no larger than the learning term when $T - L_d \leq \sqrt{K(T-L_0)}$; otherwise the learning term is smaller. Compared with its uniform rate, the expression is strictly smaller when

$$T - L_d < \min\{T - L_0, \sqrt{K(T-L_0)}\}.$$

The cap holds for every policy on every coordinate, regardless of which coordinate the policy observes. Its improvement alone therefore does not demonstrate a benefit from coordinate selection.
For reward-doubling Poly-INF, the learning term also depends on the coordinate supplied to the policy, and

$$\min\{T - L_d, \sqrt{K(T-L_d)}\} \leq \min\{T - L_0, \sqrt{K(T-L_0)}\}.$$

These expressions coincide when $L_d = L_0$, and the left side is strictly smaller when $L_d > L_0$. Within the coordinate-dependent minimum, the cap is no larger when $T - L_d \leq K$, and the square-root term is no larger when $T - L_d \geq K$.
A small-loss learning term gives a different comparison:

$$\sqrt{KL_d} < \min\{T - L_0, \sqrt{K(T-L_0)}\} \quad \Longleftrightarrow \quad L_d < \min\{T - L_0, (T-L_0)^2/K\}.$$

When $T - L_0 \geq K$, this reduces to $L_d < T - L_0$; in this regime, at least one coordinate satisfies the condition exactly when $L_0 < T/2$. For GREEN-IX, the complete learning term is $\sqrt{KL_d\log(KT)} + K\log(KT)$, so the logarithmic and additive terms must also be included before claiming a tighter rate. This comparison does not give Poly-INF a small-loss guarantee or justify taking a minimum across different policies. None of these upper-bound comparisons establishes an ordering of actual regrets or contradicts the worst-case lower bound over $L^\star = L_0$.

# B Computational experiments

## B.1 Design and evaluation

We approximate a worst-case comparison by selecting a high-regret reward sequence separately for each policy and evaluating that policy on its selected sequence. Within each setting, the eight simulated policies search the same 76 distinct sequences, whose coordinate-loss profiles are identical up to coordinate order. Specifically, the best-arm cumulative rewards on the two coordinates are $U_0$ and $U_0/2$, giving losses $T - U_0$ and $T - U_0/2$. We use $D = 2$, $T = 256U_0$, and the five combinations of $K$ and $U_0/K$ listed in Table 2. Regret is normalized by its common cap $B = U_0/2$; smaller values indicate better performance.

The sequences vary how rewards are distributed over time and across arms. In a phase sequence, the first $U_0$ rounds are divided into intervals with constant rewards; all later rewards are zero. In each active round, a hidden arm receives 1 on the higher-reward coordinate and $1/2$ on the lower-reward coordinate. Competing arms receive zero on the latter. On the former, members of a competing group receive 1, and the group remains fixed, shrinks, or changes membership across phases. Additional constructions vary positive background rewards and initial periods with equal rewards across arms on

this coordinate. All memberships and rewards are fixed before play. A phase singleton gives rewards only to the hidden arm, while a diluted singleton spreads that arm's rewards over all $T$ rounds, with values $U_0/T$ and $U_0/(2T)$ per round. Both singleton constructions appear in both coordinate orders; the other phase sequences place the higher-reward coordinate first.

The known-profile and unknown-scale comparisons use the method groups in Table 2. Only known-profile Poly-INF and modified EXP3 receive $U_0$. Every scalar policy uses coordinate 1, fixed before play. We compare tuned versions of the two proposed Poly-INF methods against six simulated baselines at their original defaults (Audibert & Bubeck, 2010; Auer et al., 2002b; Neu, 2015a; Zimmert & Seldin, 2021), and against uniform play, whose value is computed analytically.

For each proposed method, a prior development study tested $c \in \{1/4, 1/2, 1, 2, 4\}$ at $(K, U_0) = (128, 65{,}536)$ and $(512, 65{,}536)$. It selected the value minimizing the largest normalized search mean over candidates and both settings. Both methods selected $c = 4$, which divides their scale parameter $\eta$ by four, including after reward-doubling resets.

After fixing the policy parameters, search evaluates every sequence with 8 runs per policy. Validation uses 64 independent runs per evaluated sequence and selects each policy's sequence with the largest validation mean among those evaluated. The proposed methods validate all 76 sequences. The baselines validate a shared subset: the union of the four highest-regret sequences from searches of all eight policies at their original defaults, predetermined coverage sequences, and every newly added sequence.

### B.2 Results

Table 2 and Figure 2 compare normalized mean regret on each policy's selected sequence. Both proposed methods have lower means than every included baseline in all five settings: 25 comparisons in the known-profile group and 30 in the unknown-scale group. The baselines with the lowest means are modified EXP3 and EXP3.1, respectively. Relative to these baselines, the reductions are 31.0%–37.5% for known-profile Poly-INF and 5.0%–24.1% for reward-doubling Poly-INF.

| Policy / $K$ | 128 | 512 | 512 | 1024 | 2048 |
|---|---|---|---|---|---|
| $U_0/K$ | 512 | 128 | 512 | 128 | 128 |
| *(a) Known profile* | | | | | |
| Known-profile Poly-INF | **0.3329** | **0.6588** | **0.3423** | **0.6713** | **0.6737** |
| Modified EXP3 | 0.4825 | 0.9686* | 0.5479 | 0.9849* | 0.9925* |
| EXP3-IX | 0.9903* | 0.9978* | 0.9976* | 0.9989* | 0.9995* |
| Tsallis-INF (IW) | 0.9862* | 0.9976* | 0.9961* | 0.9987* | 0.9993* |
| Tsallis-INF (RV) | 0.9634* | 0.9962* | 0.9837* | 0.9979* | 0.9991* |
| Uniform (exact) | 0.9922* | 0.9980* | 0.9980* | 0.9990* | 0.9995* |
| *(b) Unknown scale* | | | | | |
| Reward-doubling Poly-INF | **0.5726** | **0.7580** | **0.5965** | **0.7594** | **0.7497** |
| EXP3.1 | 0.6028 | 0.9685 | 0.7383 | 0.9808 | 0.9874 |
| Horizon Poly-INF | 0.9890* | 0.9978 | 0.9973* | 0.9989 | 0.9994* |
| EXP3-IX | 0.9903* | 0.9978 | 0.9976* | 0.9989 | 0.9995* |
| Tsallis-INF (IW) | 0.9862* | 0.9976 | 0.9961* | 0.9987 | 0.9993* |
| Tsallis-INF (RV) | 0.9634* | 0.9962 | 0.9837* | 0.9979 | 0.9991* |
| Uniform (exact) | 0.9922* | 0.9980* | 0.9980* | 0.9990* | 0.9995* |

Table 2: Mean Pareto regret divided by $U_0/2$. Stars indicate simultaneous 95% evidence of improvement by the proposed method; uniform is exact.

The selected sequences differ across methods: both proposed methods select phase sequences, while EXP3-IX and both Tsallis-INF variants select the diluted singleton with lower-reward-coordinate feedback in every setting. These choices show why evaluating all methods on one common sequence could miss difficulties encountered by individual policies.

Uniform play has the exact normalized reference $1-1/K$. To see this, let $X$ be the reward contributed by a best arm on the lower-reward coordinate on rounds when that arm is selected. Then $R_T^{\mathrm{fs}} \le B - X$ and $\mathbb{E}_{\mathrm{unif}} X = B/K$, giving expected regret at most $(1 - 1/K)B$. A singleton attains this bound.

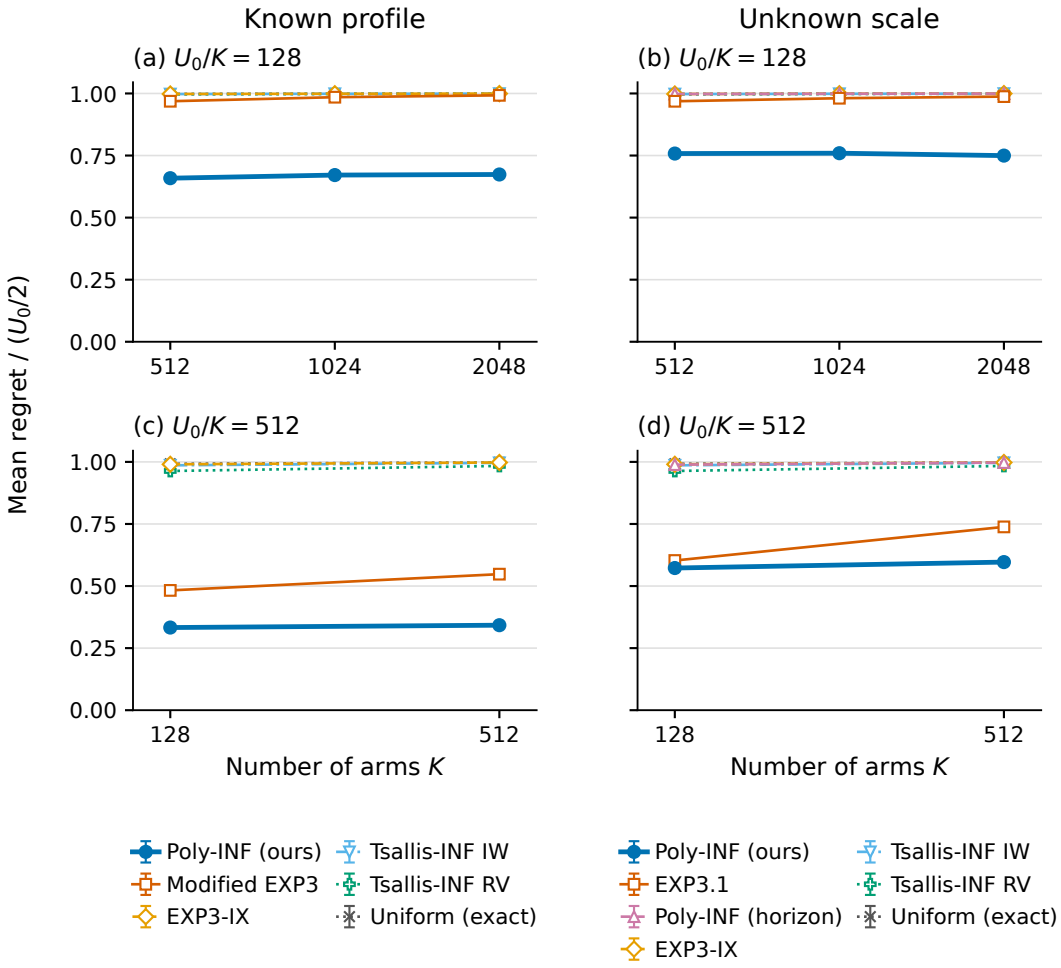


Figure 2: Normalized mean Pareto regret on each policy's selected instance: known profile (left) and unknown scale (right).