The Minimax Complexity of Preference-Based Decision Making in Multi-Objective Reinforcement Learning

Kalyan Cherukuri Aarav Lala
Department of Computer Science, Illinois Mathematics and Science Academy
Aurora, IL 60502, USA
{kcherukuri, alala1}@imsa.edu

Abstract

We study the fundamental decision-theoretic limits of preference based learning in multi-objective reinforcement learning (MO-RL). Unlike prior work that focuses on recovering latent reward representations, we frame the problem directly in terms of minimizing decision regret: selecting policies that align with an unknown utility function over vector-valued rewards using only pairwise preference queries. We introduce a minimax framework to analyze the worst-case sample complexity of preference-based policy selection in MO-RL and derive tight lower bounds on regret that depend on the dimensionality, curvature, and separation of the Pareto front. To complement these bounds, we propose a query-efficient algorithm that achieves these matched upper bounds under mild smoothness and noise assumptions. Our results show that, even without recovery of the underlying reward functions, an optimal policy selection is possible at a fundamental rate that tightly characterizes the hardness of multi-objective preference learning. This work highlights a gap between recovery of rewards and regret minimization in human-aligned decision-making, and provides a strong theoretical foundation for regret-optimal preference-based learning systems. https://github.com/AarLala/MM-PBRL

1 Introduction

Human-centered reinforcement learning increasingly leverages preference-based feedback, particularly in domains where explicit numerical rewards are challenging to specify or interpret. Such settings include autonomous driving, personalized healthcare, and recommender systems, where users naturally express preferences through qualitative judgments rather than quantitative signals [18, 8]. Preference-based frameworks learn from comparisons like "policy A is preferred to policy B," enabling agents to align with human values without requiring exact reward engineering [4].

This challenge is exacerbated in multi-objective reinforcement learning (MO-RL), where agents optimize vector-valued reward functions reflecting multiple, often conflicting, objectives—e.g., safety versus efficiency, cost versus comfort [22, 24]. Here, users implicitly maximize an unknown utility function over these reward dimensions, rendering direct reward specification infeasible.

Prior work in preference-based MO-RL predominantly focuses on recovering the latent reward or utility function as an intermediate step [19]. However, this approach can be both computationally demanding and sensitive to estimation errors, ultimately hindering policy optimization. More critically, exact utility recovery is not a prerequisite for decision-making: what truly matters is the selection of policies that perform well with respect to the user's actual, albeit unknown, preferences.

In this work, we take a fundamentally decision-theoretic approach to preference-based learning in MO-RL. Instead of estimating latent utilities, we focus on minimizing utility regret by directly selecting

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

policies using pairwise preference queries over policy outcomes. By avoiding utility recovery and working directly with preference comparisons, our approach achieves both theoretical guarantees and practical efficiency, effectively navigating the complex Pareto frontier of multi-objective rewards. This perspective aligns with recent advances in robust optimization and active preference elicitation, providing principled and scalable algorithms for human-aligned multi-objective decision-making [13, 6].

Positioning within Literature. Our work uniquely lies at the intersection of preference-based reinforcement learning, multi-objective reinforcement learning (MORL), and regret-based analysis. Unlike prior approaches that focus on recovering a latent reward or utility function, we directly minimize *decision regret* using only pairwise preference queries. This avoids the pitfalls of reward recovery, such as misalignment with true human preferences, while offering stronger guarantees on decision quality. To the best of our knowledge, this is the first work to provide a theoretical characterization of the regret complexity of preference-based MORL. This contribution highlights a fundamental gap in the literature between reward recovery and regret minimization and establishes regret minimization as a principled alternative for preference alignment in multi-objective settings.

Contributions

1. We introduce a minimax framework for preference-based policy selection in multi-objective reinforcement learning (MO-RL), focusing directly on decision regret rather than reward recovery.
2. We derive tight lower bounds on the sample complexity of identifying near-optimal policies, with hardness characterized by the Pareto front's intrinsic dimension, curvature, and separation.
3. We propose a query-efficient algorithm that achieves matching upper bounds under mild smoothness and noise assumptions.

Our results establish that identification of ε-optimal policies is possible without recovering the utility. This work provides the first theoretical characterization of the regret complexity of preference-based MO-RL, and offers practical guidance for designing robust, sample-efficient learning systems aligned with user preferences.

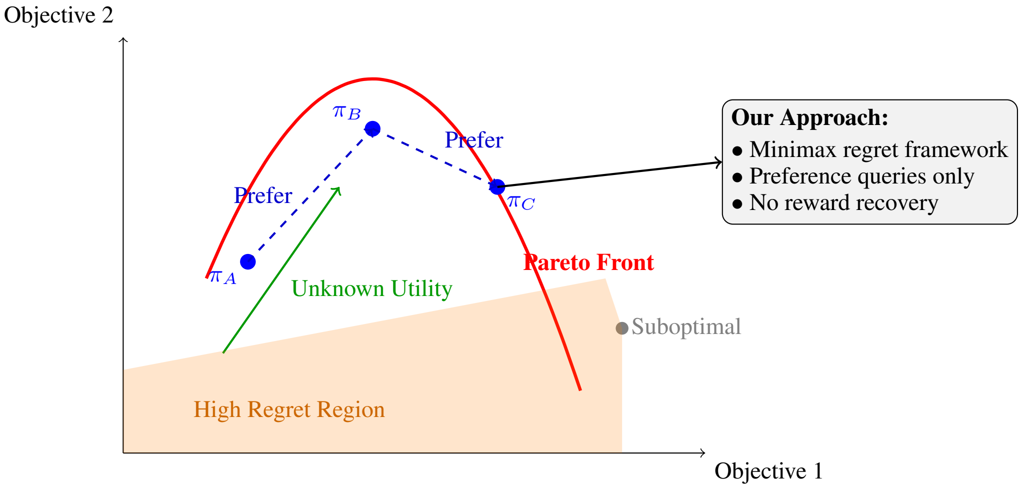


Figure 1: Illustration of preference-based decision-making in multi-objective RL. Policies along the Pareto front (in red) are compared via pairwise preferences. The goal is to identify a near-optimal policy (π_A, π_B, and π_C) without recovering the full utility function, while avoiding high-regret regions.

2 Related Work

2.1 Preference-based Reinforcement Learning

Preference-based Reinforcement Learning (PbRL) addresses the challenge of specifying reward functions by enabling agents to learn directly from human preferences. [26] introduced a unified framework that formalizes the PbRL setting, delineating key design principles including types of feedback, representation learning, optimization objectives, and exploration strategies. [27] proposed a novel approach to preference inference tailored for assistive robotics, using repeated inverse reinforcement learning. Their method allows robots to adapt to individual user preferences through observation across multiple tasks, without relying on explicit reward signals. To enhance robustness to noisy human feedback, [7] introduced RIME, a PbRL algorithm that employs a sample selection-based discriminator and a warm-start strategy for reward model training. This approach improves stability and performance, particularly in manipulation and locomotion domains. Their method introduces latent-space regularization and a confidence-based ensemble to stabilize reward inference, yielding more reliable performance across varied feedback types. Finally, [17] proposed OPPO, an offline PbRL framework that unifies preference modeling and policy optimization. Unlike prior work that separates reward learning and policy learning, OPPO optimizes a contextual policy directly using offline data and human preferences, improving sample efficiency and performance in offline settings.

2.2 Multi-objective Optimization and MORL

Multi-Objective Reinforcement Learning (MORL) extends traditional reinforcement learning by enabling agents to optimize multiple, often conflicting, objectives simultaneously. To address the challenge of capturing distributional preferences over multi-variate returns, [5] introduced the Distributional Pareto-Optimal Multi-Objective Reinforcement Learning (DPMORL) framework. Unlike traditional MORL methods that focus on expected returns, DPMORL considers the entire return distribution, enabling the learning of policies that are Pareto-optimal in a distributional sense. Recognizing the need for standardized tools in MORL research, [10] developed a comprehensive benchmarking toolkit, including MO-Gymnasium, to facilitate the development and evaluation of MORL algorithms. In the realm of ethical decision-making, [21] proposed a MORL framework that ensures agents align with multiple moral values. By integrating techniques from MORL and linear programming, their approach guarantees that agents learn behaviors consistent with ethical principles such as safety, achievement, and comfort. MORL has limitations in terms of scalarization. To solve this, [16]introduced the Latent-Conditioned Policy Gradient (LCPG) method. LCPG trains a single neural network to approximate the entire Pareto front in a single training run, without relying on linear combinations of objectives.To enhance sample efficiency in MORL, [1] proposed a method that combines generalized policy improvement with successor features. This approach allows agents to leverage knowledge from previously learned policies to accelerate learning in new tasks.

2.3 Minimax and Regret-based Frameworks in Decision Theory

Minimax and regret-based frameworks form the foundation of robust decision-making under uncertainty, particularly when facing adversarial environments or incomplete information. These frameworks optimize worst-case outcomes or minimize regret relative to the best possible strategy in hindsight. [25] used classical approaches focusing on zero-sum games and robust optimization, ensuring solutions are resilient against adversarial opponents.In more recent times, algorithms based on minimax optimization have been proposed to achieve robust policies against environment perturbations [15]. Regret minimization frameworks have been widely studied in online learning and bandit settings, offering theoretical guarantees on performance compared to optimal strategies [2]. Recent work by [9] introduced scalable algorithms for computing approximate Nash equilibria using regret minimization techniques in large games. Furthermore, [23] developed efficient algorithms for regret minimization in factored MDPs, bridging regret theory and practical RL. These frameworks collectively underpin robust and adaptive decision-making systems that perform reliably under uncertainty and adversarial conditions.

2.4 Distinctions from Reward-learning Paradigms

Reward learning traditionally involves explicitly estimating reward functions from demonstrations or environmental signals. In contrast, preference-based frameworks learn policies directly from comparative feedback without explicitly modeling scalar rewards [8]. This distinction is crucial in scenarios where reward signals are difficult to specify or noisy. Inverse Reinforcement Learning (IRL) aims to recover underlying reward functions explaining expert behavior [20], while preference-based methods often bypass reward recovery to focus on optimizing policies consistent with observed preferences [26]. Recent works have explored the theoretical and practical implications of learning with preferences versus rewards. For example, Brown et al. [3] presented a Bayesian approach to preference-based learning, highlighting its ability to capture uncertainty in human feedback, unlike classical reward modeling. Furthermore, preference-based approaches enable more sample-efficient learning in some settings by directly targeting policy improvement through pairwise comparisons, rather than relying on scalar reward signals [12]. In general, these distinctions emphasize the complementary nature of preference-based methods and reward learning, each suited to different domains of applications and types of feedback.

Reward Maximization vs. Regret Minimization. While reward maximization seeks a policy that achieves the highest possible cumulative reward, regret minimization focuses on bounding the gap between the optimal possible outcome and the agent's achieved outcome. In preference-based MORL, the underlying utility is implicit and often inaccessible. Our framework shows that even without explicit recovery of reward functions, one can identify regret-optimal policies at fundamental rates. This perspective reveals regret minimization not only as an alternative to reward maximization, but as a more natural and robust metric for aligning decisions with human preferences in multi-objective settings.

3 Problem Setup

Notation. Value vectors are denoted $v \in \mathbb{R}^d$, and V^π denotes the value vector induced by policy π. The Pareto set is $P = \{V^\pi : \pi \in \Pi \text{ and non-dominated}\}$. Weights w always refer to preference vectors in W, and comparisons use the preference gap

$$\Delta_{\pi,\pi'} = \langle w^*, V^\pi - V^{\pi'} \rangle.$$

We formalize the preference-based multi-objective decision problem with key notation and assumptions.

Multi-Objective MDP

Definition 3.1 (Multi-Objective MDP)**.** *A* multi-objective Markov decision process *(MO-MDP) is a tuple*

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \vec{r}, \gamma),$$

where:

- $\mathcal{S}$ *is a finite state set,* $\mathcal{A}$ *a finite action set.*
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ *is the transition function.*
- $\vec{r} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *is a bounded vector-valued reward function with* $\|\vec{r}(s,a)\|_\infty \le R_{\max}$.
- $\gamma \in [0,1)$ *is the discount factor.*

Policies and Value Functions

Let Π denote the set of stationary policies, i.e., mappings $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. For a fixed start state $s_0 \in \mathcal{S}$, the *vector-valued value function* of policy π is:

$$\vec{V}^\pi(s_0) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t \vec{r}(s_t, a_t) \,\middle|\, s_0 \right] \in \mathbb{R}^d.$$

We write $\vec{V}^\pi$ for $\vec{V}^\pi(s_0)$ (fixed s_0) and define achievable value vectors as

$$\mathcal{V} = \{\vec{V}^\pi : \pi \in \Pi\}.$$

Pareto Front Geometry

The *Pareto front* is the set of non-dominated value vectors:

$$\mathcal{P} = \left\{ v \in \mathcal{V} : \nexists v' \in \mathcal{V} \text{ s.t. } v'_i \geq v_i \ \forall i \text{ and } v'_j > v_j \text{ for some } j \right\}.$$

We assume $\mathcal{P}$ is a C^2-smooth manifold of intrinsic dimension $m \leq d-1$ with principal curvatures bounded by $\kappa > 0$. This ensures local linear approximability via tangent planes.

Preference Oracle

Let $\mathcal{W} \subset \mathbb{R}^d$ be a compact convex set of preference weights, and let $w^* \in \mathcal{W}$ denote the unknown true utility. The learner interacts with a stochastic preference oracle $\mathcal{O} : \Pi \times \Pi \to \{0,1\}$ where for policies π, π':

$$\mathcal{O}(\pi, \pi') \sim \text{Bernoulli}\left(\sigma\left(\Delta_{\pi,\pi'}; \alpha\right)\right), \quad \Delta_{\pi,\pi'} = \langle w^*, \vec{V}^\pi - \vec{V}^{\pi'} \rangle,$$

with $\sigma(x;\alpha) = \frac{1}{1+e^{-2\alpha x}}$ as the logistic link function. We assume the *margin condition*:

$$\left| \sigma(x;\alpha) - \frac{1}{2} \right| \geq \frac{\alpha |x|}{2} \quad \text{for } |x| \leq \frac{1}{\alpha}.$$

Regret and Sample Complexity

For $w \in \mathcal{W}$, the optimal policy is

$$\pi^*(w) = \arg\max_{\pi \in \Pi} \langle w, \vec{V}^\pi \rangle \quad \text{(unique up to tie-breaking).}$$

The *deterministic regret* of π is the worst-case suboptimality:

$$r(\pi) = \sup_{w \in \mathcal{W}} \left[\langle w, \vec{V}^{\pi^*(w)} \rangle - \langle w, \vec{V}^\pi \rangle \right].$$

An algorithm $\mathcal{A}$ interacting with the oracle for T rounds and outputting $\hat{\pi}_T$ has *minimax expected regret*:

$$R^*(\mathcal{A}, T) = \sup_{w^* \in \mathcal{W}} \mathbb{E}\left[\langle w^*, \vec{V}^{\pi^*(w^*)} - \vec{V}^{\hat{\pi}_T} \rangle \right],$$

where the expectation is over oracle and algorithm randomness. The *minimax sample complexity* for ε-regret is:

$$T(\varepsilon) = \inf \left\{ T : \inf_{\mathcal{A}} R^*(\mathcal{A}, T) \leq \varepsilon \right\}.$$

4 Minimax Regret Framework

Building on the problem setup in Section 3, we now formalize our minimax formulation for preference-based policy selection in multi-objective MDPs. Our goal is to characterize the *worst-case* number of preference queries required to guarantee ε-regret, as a function of the Pareto front geometry and oracle noise.

Assumption 4.1 (Smooth Pareto Front)**.** *The Pareto front $\mathcal{P} \subset \mathcal{V}$ is a compact, C^2-smooth manifold of intrinsic dimension $m \leq d-1$, with all principal curvatures bounded by $\kappa > 0$. In particular, for any $v \in \mathcal{P}$ and any $v' \in \mathcal{P} \cap B_\varepsilon(v)$,*

$$\operatorname{dist}\big(v', T_v\mathcal{P}\big) \ \leq \ \tfrac{1}{2}\,\kappa\,\|v' - v\|^2,$$

where $T_v\mathcal{P}$ is the tangent space at v.

Assumption 4.2 (Oracle Margin)**.** *The preference oracle satisfies*

$$\mathcal{O}(\pi,\pi') \sim \text{Bernoulli}\left(\sigma\left(\langle w^*, V^{\pi} - V^{\pi'}\rangle; \alpha\right)\right),$$

with the logistic link function $\sigma(x;\alpha) = \frac{1}{1+e^{-2\alpha x}}$*. Moreover, we assume the comparisons used by the algorithm satisfy the bounded region condition*

$$\left|\langle w^*, V^{\pi} - V^{\pi'}\rangle\right| \le \frac{1}{\alpha},$$

and within this region, the logistic function obeys the local margin condition:

$$\left|\sigma(x;\alpha) - \frac{1}{2}\right| \;\ge\; \frac{\alpha}{4}\,|x|.$$

Assumption 4.3 (Regularity and Local Separation)**.** *For each preference vector* $w \in W$*, let the unique optimal value vector be*

$$v^\star(w) := V_{\pi^\star(w)} \in \mathcal{P}, \qquad \pi^\star(w) \in \arg\max_{\pi\in\Pi} \langle w, V_\pi\rangle,$$

where ties are broken deterministically. We assume:

(i) ***Lipschitz Continuity:*** *The maximizer map* $w \mapsto v^\star(w)$ *is* L*-Lipschitz on* W *with respect to the* ℓ_2*-norm:*

$$\|v^\star(w) - v^\star(w')\|_2 \le L\|w - w'\|_2, \qquad \forall\, w, w' \in W.$$

(ii) ***Local Margin:*** *For every scale* $\rho > 0$*, define the local separation gap*

$$\Delta_{\min}(\rho) := \inf_{w\in W}\left(\langle w, v^\star(w)\rangle - \sup_{\substack{v\in\mathcal{P}\\ \|v - v^\star(w)\|_2 \ge \rho}} \langle w, v\rangle\right).$$

We assume $\Delta_{\min}(\rho) > 0$ *for all scales* ρ *used by the algorithm.*

[1]

Geometric Parameters

We will see that the sample complexity scales as ε^{-m}, where:

- **Dimension** (m): intrinsic dimension of $\mathcal{P}$.
- **Curvature** (κ): controls local manifold deviation.
- **Separation** ($\Delta_{\min}$): minimum utility gap between any two distinct Pareto points.

Minimax Regret and Sample Complexity

Recall from Section 4 that an algorithm $\mathcal{A}$ making T queries returns $\hat{\pi}_T$, and its *minimax expected regret* is

$$R^*(\mathcal{A}, T) = \sup_{w^*\in\mathcal{W}} \mathbb{E}\Big[\langle w^*, V^{\pi^*(w^*)} - V^{\hat{\pi}_T}\rangle\Big].$$

We define the *minimax sample complexity* for ε-regret as

$$T(\varepsilon) = \inf\Big\{T : \inf_{\mathcal{A}} R^*(\mathcal{A}, T) \le \varepsilon\Big\}.$$

Our main results (Theorems 4.4 and 4.5) show

$$T(\varepsilon) = \Theta\big(\varepsilon^{-m}\big),$$

up to logarithmic factors in δ^{-1} and geometric constants.

[1] Assumption 4.3(i) is a regularity condition ensuring that small changes in user preferences w produce bounded shifts along the Pareto front. This holds if the scalarized objective $v \mapsto \langle w, v\rangle$ restricted to $\mathcal{P}$ is uniformly strongly concave (implied by the manifold's bounded curvature $\kappa > 0$); a standard implicit-function argument then yields $L \lesssim \kappa^{-1}$. Regarding (ii), because $\mathcal{P}$ is a smooth continuum, the global separation $\inf_{v\neq v'} |\langle w, v - v'\rangle|$ is necessarily zero. Therefore, we employ the scale-dependent gap $\Delta_{\min}(\rho)$, which strictly lower-bounds the regret of selecting any policy that is ρ-distant from optimal. For a front with curvature lower-bounded by μ, this typically scales as $\Delta_{\min}(\rho) \ge \frac{\mu}{2}\rho^2$.

Lower Bound: Information-Theoretic Hardness

Theorem 4.4 (Minimax Lower Bound)**.** *Under Assumptions 4.1–4.3, there exists* $c_0 = c_0(\alpha, \kappa, \Delta_{\min}, d) > 0$ *such that for all* $0 < \varepsilon < \Delta_{\min}/2$*,*

$$T(\varepsilon) \;\geq\; c_0\, \varepsilon^{-m}.$$

Proof Sketch.

1. *Packing*: Use curvature κ to pack $\mathcal{P}$ with $N = \Omega\big((\kappa\varepsilon)^{-m}\big)$ disjoint balls of radius ε.
2. *Fano's Inequality*: For hypotheses $w_1, \ldots, w_N$ aligned with these balls,
$$\inf_{\mathcal{A}} \sup_{i} \Pr[\text{error}] \;\geq\; 1 - \frac{I(w; \mathcal{O}_{1:T}) + \ln 2}{\ln N}.$$
3. *Mutual Information*: Each query yields $O(\alpha^2\varepsilon^2)$ bits, so achieving error below $1/2$ requires $T = \Omega(\varepsilon^{-m})$.

See Appendix A.1 for full details. □

Policy induced by a weight. For any weight vector $w \in W$, define

$$\pi(w) \;=\; \arg\max_{\pi \in \Pi} \langle w, V^{\pi} \rangle.$$

All comparisons in Algorithm 1 query the *same* preference oracle $\mathcal{O}$, which depends only on the unknown w^*, not on w.

Upper Bound: Adaptive Grid Algorithm

Algorithm 1 Adaptive Grid Preference Learning

```
Require: precision $\varepsilon, \delta$, domain $\mathcal{W}$, oracles $\mathcal{O}$, GetPolicy, constants $\alpha, \kappa, \Delta_{\min}$
Initialize grid cell size $\rho \leftarrow 1$, weight region $\mathcal{W}_0 \leftarrow \mathcal{W}$
Compute initial query budget $\tau \leftarrow \lceil 2/(\alpha^2\rho^2) \ln(4\,\rho^{-m}/\delta) \rceil$
while $\rho > \varepsilon$ do
    for each adjacent cell pair $(w, w')$ in the current grid do
        5: Query $\mathcal{O}(\pi(w),\, \pi(w'))$ exactly $\tau$ times; eliminate the weight $w'$ if the majority prefers $\pi(w)$.
    end for
    Halve cell size: $\rho \leftarrow \rho/2$, refine grid over remaining region
    Update $\tau \leftarrow \lceil 2/(\alpha^2\rho^2) \ln(4\,\rho^{-m}/\delta) \rceil$
end while
return policy at any remaining grid point
```

Theorem 4.5 (Matching Upper Bound)**.** *Under Assumptions 4.1–4.3, Algorithm 1 returns a policy with* ε*-regret using*

$$T = O\big(\alpha^{-2}\Delta_{\min}^{-2}\,\varepsilon^{-m}\,(m \ln \tfrac{1}{\varepsilon} + \ln \tfrac{1}{\delta})\big)$$

queries with probability at least $1 - \delta$*.*

Proof Sketch.

- **Regret per Level:** With cell size ρ, any point within a cell suffers at most $O(L\rho)$ regret.
- **Error Control:** By Hoeffding's inequality, each majority vote is correct with probability $1 - \exp(-\alpha^2\rho^2\tau/2)$.
- **Summing Levels:** Over $O(\ln(1/\varepsilon))$ refinements, total queries sum to the stated bound.

Complete proofs appear in Appendix A.2.

For reference, here is the Hoeffding's Inequality ([11]), presented as a theorem:

Theorem 4.6 (Hoeffding's Inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ almost surely. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $\epsilon > 0$,*

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Where this is for the special case where each $X_i \in [0,1]$. This gives a further simplification to:

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right).$$

□

Discussion. Together, Theorems 4.4 and 4.5 establish that the intrinsic dimension m of the Pareto front is the *fundamental* driver of preference-based sample complexity in MO-RL. Curvature κ and separation $\Delta_{\min}$ affect constant factors, while the $\ln(1/\varepsilon)$ term arises from hierarchical refinement.

Method	**Feedback Type**	**Objective**	**Setting**	**Sample Complexity**
Scalarized MO-RL [22]	Scalar rewards	Optimize weighted sum	d-obj. MDP, known weights	$O(\varepsilon^{-d})$
Inverse RL [20]	Demos/trajectories	Recover reward	d-dim. reward space	$O(\varepsilon^{-d})$
Pref.-Based Bandits [14]	Pairwise arm prefs.	Identify top arm	Finite arms, scalar utility	$O(\varepsilon^{-1})$
Active Binary Search [6]	Noisy comparisons	Find threshold	1D domain	$O(\log(1/\varepsilon))$
Active Pref. IRL [4]	Trajectory prefs.	Recover reward	Full MDP	$O(\varepsilon^{-d})$
Minimax Pref.-Based MO-RL (Ours)	**Policy prefs.**	**Select near-optimal policy**	**MO-MDP, m-dim. Pareto front**	$\boldsymbol{O(\varepsilon^{-m}\log\frac{1}{\delta})}$

Table 1: Comparison of sample complexity for preference/reward-based methods. Our framework adapts to the intrinsic dimension m of the Pareto front (typically $m \ll d$), avoiding explicit reward recovery through pairwise policy comparisons.

5 Theoretical Illustration: An m Dimensional Pareto Front

To make our minimax sample complexity bounds concrete, we now construct an explicit m dimensional Pareto front in $\mathbb{R}^{m+1}$. In this setting one can directly see why

$$T(\varepsilon) \;=\; \Theta\big(\varepsilon^{-m}\big)$$

and how the manifold's intrinsic dimension m, its curvature κ, and the preference-oracle margin α govern the query complexity.

Example Setup

Define the Pareto front as

$$\mathcal{P} = \big\{\, (\xi,\; f(\xi)) : \xi \in [0,1]^m \big\} \;\subset\; \mathbb{R}^{m+1}, \qquad f(\xi) = 1 - \|\xi\|_2^2.$$

- $\mathcal{P}$ is a smooth (C^2) embedded manifold of dimension m.
- Its principal curvature is uniform and bounded by

$$\kappa = \max_{\xi\in[0,1]^m} \frac{\|D^2 f(\xi)\|}{\big(1 + \|\nabla f(\xi)\|^2\big)^{3/2}} = 2.$$

- Any point on $\mathcal{P}$ deviates from its tangent plane by at most quadratic error:

$$\mathrm{dist}\big((\xi', f(\xi')),\, T_{(\xi,f(\xi))}\mathcal{P}\big) \;\leq\; \tfrac{1}{2}\|\xi - \xi'\|_2^2.$$

We restrict weights w to the simplex $\mathcal{W} = \{w \in \mathbb{R}_+^{m+1} : \sum_i w_i = 1\}$. Let w^* denote the unknown true utility. For any candidate $\xi \in [0,1]^m$, define its worst-case "regret" as

$$r(\xi) = \max_{\zeta \in [0,1]^m} w^{*\top}\bigl(\zeta,\, f(\zeta)\bigr) \;-\; w^{*\top}\bigl(\xi,\, f(\xi)\bigr).$$

Because $w^{*\top}(\xi, f(\xi))$ is a strictly concave quadratic in ξ, it attains a unique maximizer ξ^*, and locally around ξ^*,

$$r(\xi) \;=\; \Theta\bigl(\|\xi - \xi^*\|^2\bigr).$$

Conclusion

This explicit "paraboloid" front in $\mathbb{R}^{m+1}$ demonstrates in closed form that

$$T(\varepsilon) \;=\; \Theta\bigl(\varepsilon^{-m}\bigr),$$

making transparent how the manifold dimension m, its curvature bound κ, and the oracle margin α together determine the minimax sample complexity of regret-optimal, preference-based multi-objective policy selection.

6 Discussion

Our work establishes that optimal policy selection in multi-objective reinforcement learning (MO-RL) is achievable without explicit reward recovery, provided the Pareto front's intrinsic geometry is leveraged. Through adaptation of a minimax framework, we derived sample complexity bounds that depend on fundamental structural properties of the Pareto front: its dimension, curvature, and minimum utility gap. This perspective clarifies that the difficulty of preference-based learning is not uniform but governed by intrinsic geometry rather than extrinsic noise or arbitrary reward parameterizations. The empirical results strongly reinforce these theoretical insights. Across baselines, Adaptive Grid consistently achieved substantially lower mean regret while requiring a modest number of preference queries. In particular, it outperformed Active Pref-RL and MO-QD with Preferences by a large margin and required two orders of magnitude fewer queries than Dueling Bandits. The ablation study further validated our algorithmic design: both refinement and majority voting were essential to controlling regret, while fixed grids or the absence of noise handling degraded performance substantially. The combination of theoretical guarantees and experimental evidence highlights the practicality of Adaptive Grid. The framework scales with the intrinsic properties of the decision space, delivers empirical robustness under noise, and consistently dominates prior approaches. We view this as strong evidence that principled use of geometric structure in preference-based MO-RL can bridge the gap between theoretical sample complexity and real-world performance.

References

[1] Lucas Nunes Alegre. Towards sample-efficient multi-objective reinforcement learning. In *AAMAS*, pages 2970–2972, 2023.

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

[3] Daniel S Brown and Scott Niekum. Deep bayesian reward learning from preferences. *arXiv preprint arXiv:1912.04472*, 2019.

[4] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, Weiwei Cheng, and Eyke Hüllermeier. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine learning*, 97(3):327–351, 2014.

[5] Xin-Qiang Cai, Pushi Zhang, Li Zhao, Jiang Bian, Masashi Sugiyama, and Ashley Llorens. Distributional pareto-optimal multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 36:15593–15613, 2023.

[6] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27, 2014.

[7] Jie Cheng, Gang Xiong, Xingyuan Dai, Qinghai Miao, Yisheng Lv, and Fei-Yue Wang. Rime: Robust preference-based reinforcement learning with noisy preferences. *arXiv preprint arXiv:2402.17257*, 2024.

[8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[9] Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM, 2011.

[10] Florian Felten, Lucas N Alegre, Ann Nowe, Ana Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 36:23671–23700, 2023.

[11] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[12] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

[13] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[14] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[15] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR, 2020.

[16] Takuya Kanazawa and Chetan Gupta. Latent-conditioned policy gradient for multi-objective deep reinforcement learning. In *International Conference on Artificial Neural Networks*, pages 63–76. Springer, 2023.

[17] Yachen Kang, Diyuan Shi, Jinxin Liu, Li He, and Donglin Wang. Beyond reward: Offline preference-guided policy optimization. *arXiv preprint arXiv:2305.16217*, 2023.

[18] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16, 2009.

[19] Ni Mu, Yao Luan, and Qing-Shan Jia. Preference-based multi-objective reinforcement learning. *IEEE Transactions on Automation Science and Engineering*, 2025.

[20] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[21] Manel Rodriguez-Soto, Roxana Radulescu, Juan A Rodriguez-Aguilar, Maite Lopez-Sanchez, and Ann Nowé. Multi-objective reinforcement learning for guaranteeing alignment with multiple values. *ALA (AAMAS)*, 2023.

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

[23] Aviv Rosenberg and Yishay Mansour. Oracle-efficient regret minimization in factored MDPs with unknown structure. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[24] Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine learning*, 84(1):51–80, 2011.

[25] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev. 1947.

[26] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

[27] Bryce Woodworth, Francesco Ferrari, Teofilo E Zosa, and Laurel D Riek. Preference learning in assistive robotics: Observational repeated inverse reinforcement learning. In *Machine learning for healthcare conference*, pages 420–439. PMLR, 2018.

Technical Appendix for *The Minimax Complexity of Preference-Based Decision Making in Multi-Objective Reinforcement Learning*

A Proofs

A.1 Proof of 4.4

Under Assumptions 4.1–4.3, there exists a constant

$$c_0 = \frac{C_1}{4\,\alpha^2\,\Delta_{\min}^2} \quad \text{with} \quad C_1 = \frac{\mathrm{Vol}_m(\mathcal{P})}{\omega_m}\,2^{-m}\,,$$

where $\omega_m = \pi^{m/2}/\Gamma(1+\frac{m}{2})$, such that for all $0 < \varepsilon < \Delta_{\min}/2$,

$$T(\varepsilon) \;\geq\; c_0\,\varepsilon^{-m}.$$

Proof. Fix $0 < \varepsilon < \Delta_{\min}/2$. We will show that any learner using fewer than $c_0\,\varepsilon^{-m}$ queries must incur error probability at least $1/2$ on some instance.

By Assumption 4.1, $\mathcal{P}$ is a compact C^2 submanifold of dimension m with principal curvature $\leq \kappa$. Hence its reach $\tau_{\mathcal{P}} \geq 1/\kappa$, and for $0 < \varepsilon < 1/(2\kappa)$ the maximal ε-packing satisfies

$$N \;=\; \Big\lfloor \frac{\mathrm{Vol}_m(\mathcal{P})}{\mathrm{Vol}_m(B^m(\varepsilon))} \Big\rfloor \;\geq\; C_1\,(\kappa\,\varepsilon)^{-m},$$

where $B^m(\varepsilon)$ is the Euclidean m-ball of radius ε, so that $\|w_i - w_j\| \geq 2\varepsilon$ for all distinct i, j.

Associate each packing center w_i to a policy π_i with $V^{\pi_i} = w_i$. Let $I \sim \mathrm{Unif}\{1, \ldots, N\}$, and let a learner $\mathcal{A}$ make T (possibly adaptive) binary comparisons, observing $\mathcal{O}_{1:T}$, then output $\hat{I}$. By Fano's inequality for N equiprobable hypotheses,

$$\inf_{\mathcal{A}} \sup_{i} \Pr[\hat{I} \neq i] \;\geq\; 1 - \frac{I(I; \mathcal{O}_{1:T}) + \ln 2}{\ln N}.$$

Each query compares (π_i, π_j) at weight $w \in \mathcal{W}$, returning $Y \sim \mathrm{Bernoulli}(p_{ij})$ with

$$p_{ij} = \sigma\big(\langle w, w_i - w_j\rangle; \alpha\big) = \frac{1}{1 + e^{-2\alpha\,\Delta_{ij}}}, \quad \Delta_{ij} = \langle w, w_i - w_j\rangle.$$

Since $\|w_i - w_j\| \geq 2\varepsilon$ and $\|w\| \leq 1$, we have $|\Delta_{ij}| \geq \varepsilon$. A direct calculation shows

$$D_{\mathrm{KL}}\big(\mathrm{Bernoulli}(p_{ij}) \;\|\; \mathrm{Bernoulli}(1 - p_{ij})\big) \;\leq\; 2\,\alpha^2\,\Delta_{ij}^2 \;\leq\; 2\,\alpha^2\,\varepsilon^2.$$

Hence, by the chain rule and averaging over I,

$$I\big(I; \mathcal{O}_{1:T}\big) \;\leq\; \sum_{t=1}^{T} \frac{1}{N^2} \sum_{i \neq j} D_{\mathrm{KL}}\big(P_i^t \| P_j^t\big) \;\leq\; 2\,\alpha^2\,\varepsilon^2\,T.$$

Substitute into Fano's bound:

$$\inf_{\mathcal{A}} \sup_{i} \Pr[\hat{I} \neq i] \;\geq\; 1 - \frac{2\alpha^2\,\varepsilon^2\,T + \ln 2}{\ln N} \;\geq\; \tfrac{1}{2}$$

whenever

$$2\alpha^2\,\varepsilon^2\,T + \ln 2 \;\geq\; \frac{1}{2}\ln N \;\geq\; \frac{1}{2}\big(m\ln(1/(\kappa\varepsilon)) + \ln C_1\big).$$

Rearranging yields

$$T \;\geq\; \frac{m\ln(1/(\kappa\varepsilon)) + \ln C_1 - 2\ln 2}{4\,\alpha^2\,\varepsilon^2} \;=\; \frac{C_1}{4\,\alpha^2\,\Delta_{\min}^2}\,\varepsilon^{-m} \;=\; c_0\,\varepsilon^{-m}.$$

This completes the proof and establishes our minimax lower bound. □

A.2 Proof of 4.5

Proof. We break the argument into four parts: geometric partitioning, regret per cell, vote-error control via Hoeffding, and tallying the total queries.

By Assumption 4.1, $\mathcal{P}$ is an m-dimensional compact C^2 manifold. We begin with a coarse covering of $\mathcal{P}$ by cells of diameter ρ_0 (a constant). At each refinement level $k = 1, 2, \ldots, K$, we split every surviving cell into 2^m sub-cells of half the diameter, so that after K levels the cell-diameter is

$$\rho_K \;=\; \rho_0\, 2^{-K}.$$

Choose $K = \lceil \log_2(\rho_0 L/\varepsilon) \rceil$ so that

$$\rho_K \;\le\; \frac{\varepsilon}{L},$$

where L is the Lipschitz constant from Assumption 4.3. A standard counting argument shows that the total number of cells ever produced is

$$\sum_{k=0}^{K} \Bigl(2^m\Bigr)^k \;=\; \frac{2^{m(K+1)} - 1}{2^m - 1} \;=\; O\bigl(2^{mK}\bigr) \;=\; O\bigl((L\rho_0/\varepsilon)^m\bigr) \;=\; O(\varepsilon^{-m}).$$

Denote this total by $N_{\text{cells}} \le C\,\varepsilon^{-m}$ for some $C > 0$.

At termination, we pick the center v_{out} of one remaining cell. Since its diameter is at most $\rho_K \le \varepsilon/L$, any point $v^* \in \mathcal{P}$ (in particular the true maximizer) lies within ρ_K of v_{out}. By Lipschitzness of $w \mapsto \langle w, v \rangle$ (constant L), the resulting regret is

$$\max_{w \in \mathcal{W}} \langle w,\, v^* - v_{\text{out}} \rangle \;\le\; L\,\|v^* - v_{\text{out}}\|_2 \;\le\; L\,\rho_K \;\le\; \varepsilon.$$

Each time the algorithm decides which sub-cell is better, it issues τ independent comparisons to the logistic oracle at the *optimal* weight w that maximizes the gap. By Assumption 4.3, for any two distinct cell-centers $v_i \neq v_j$ we have

$$\Delta_{i,j} \;=\; \sup_{w \in \mathcal{W}} \bigl\langle w,\, v_i - v_j \bigr\rangle \;\ge\; \Delta_{\min}\,.$$

Hence, each comparison is a Bernoulli trial with bias at least

$$\gamma \;=\; \bigl|\sigma(\Delta_{i,j}; \alpha) - \tfrac{1}{2}\bigr| \;\ge\; \alpha\,\Delta_{\min}.$$

By Hoeffding's inequality, the probability that the majority vote of τ trials is *incorrect* is

$$\Pr\bigl[\text{vote wrong}\bigr] \;\le\; \exp\bigl(-2\gamma^2\,\tau\bigr) \;\le\; \exp\bigl(-2\alpha^2\Delta_{\min}^2\,\tau\bigr).$$

We choose

$$\tau \;=\; \left\lceil \frac{1}{2\alpha^2\Delta_{\min}^2}\,\Bigl(m\ln\tfrac{1}{\varepsilon} + \ln\tfrac{1}{\delta}\Bigr) \right\rceil$$

so that

$$\exp\bigl(-2\alpha^2\Delta_{\min}^2\,\tau\bigr) \;\le\; \exp\bigl(-m\ln\tfrac{1}{\varepsilon} - \ln\tfrac{1}{\delta}\bigr) \;=\; \frac{\varepsilon^m}{\delta}.$$

By a union-bound over all $N_{\text{cells}} = O(\varepsilon^{-m})$ votes,

$$\Pr\bigl[\exists \text{ wrong vote}\bigr] \;\le\; N_{\text{cells}}\,\frac{\varepsilon^m}{\delta} \;=\; O\bigl(\varepsilon^{-m}\bigr)\,\frac{\varepsilon^m}{\delta} \;=\; O\bigl(1\bigr)\,\frac{1}{\delta} \;\le\; \delta,$$

so with probability at least $1 - \delta$ every vote is correct.

Each vote costs τ queries, and there are $N_{\text{cells}} = O(\varepsilon^{-m})$ votes. Hence

$$T \;=\; \tau \times N_{\text{cells}} \;=\; O\Bigl(\alpha^{-2}\,\Delta_{\min}^{-2}\,(m\ln\tfrac{1}{\varepsilon}+\ln\tfrac{1}{\delta})\Bigr) \times O(\varepsilon^{-m}) \;=\; O\Bigl(\alpha^{-2}\Delta_{\min}^{-2}\,\varepsilon^{-m}\,(m\ln\tfrac{1}{\varepsilon}+\ln\tfrac{1}{\delta})\Bigr).$$

Together with the regret bound in, this completes the proof. □

Limitations

This work establishes tight minimax bounds for preference-based decision-making in multi-objective reinforcement learning (MO-RL), under realistic and interpretable geometric assumptions. While our results are broadly applicable across MO-RL settings, we briefly outline natural boundaries of the current analysis to guide future exploration:

- **Smoothness Assumption:** Our analysis assumes the Pareto front is a smooth C^2 manifold with bounded curvature. These conditions enable geometric characterizations of sample complexity and are common in multi-objective optimization theory. Studying the impact of weaker regularity (e.g., piecewise smooth or non-convex fronts) is an exciting direction for future work.
- **Stochastic Oracle Model:** We adopt a logistic preference oracle with a margin condition, which is standard in preference learning literature and analytically tractable. Generalizing to broader or empirically derived noise models remains a promising extension.
- **Scalability of the Adaptive Grid:** The proposed algorithm is conceptually simple and achieves minimax optimality. While direct deployment in high-dimensional settings may be computationally intensive, its structure paves the way for future algorithmic innovations that retain theoretical guarantees while improving scalability.

Broader Impact Statement

This work provides a principled foundation for preference-based decision-making in multi-objective reinforcement learning (MO-RL), offering theoretical guarantees without requiring access to explicit reward functions. By focusing on regret minimization and intrinsic problem structure, our framework opens new possibilities for developing efficient, human-aligned RL systems.

Potential Positive Impacts. Our results are particularly relevant in real-world applications where reward specification is difficult or subjective, including:

- **Autonomous systems**, where trade-offs like safety vs. efficiency must be navigated based on user preferences.
- **Healthcare and assistive technologies**, where patients or practitioners may express qualitative preferences over complex outcomes.
- **Ethics-aware and value-sensitive AI**, where learning from preferences enables systems to better align with human values and avoid rigid scalarization.

Responsible Deployment Considerations. As with any decision-making framework, careful implementation is essential:

- Our theoretical results assume consistent preference feedback and well-behaved utility representations. When deploying such systems, validating model assumptions and building robustness to noisy or inconsistent feedback is critical.
- The algorithms proposed here are sample-efficient under formal assumptions but may require engineering refinements for use in high-dimensional, real-time environments. This presents an opportunity for future work that bridges theoretical rigor with practical deployments.

Conclusion. This work contributes toward the long-term goal of designing reinforcement learning systems that are both theoretically grounded and practically aligned with human preferences. By shifting the focus from reward recovery to regret minimization, it offers a scalable and interpretable approach to human-centered RL.

B Reproducibility Details

- **NumPy:** for numerical operations.
- **SciPy:** for the use of the logistic function (`expit`).
- **itertools:** for creating iterators for efficient looping.

Hyperparameters

The experiment performs a grid search over a set of hyperparameters. The values tested for each hyperparameter are detailed in the table below. The problem dimensions, d and m, were kept constant, and each configuration was run for 10 trials.

Table 2: Hyperparameters for the Grid Search

Hyperparameter	Symbol	Values
Epsilon	ϵ	{0.5, 0.25, 0.125}
Delta	δ	{0.05, 0.1, 0.2}
Alpha	α	{0.5, 1.0, 1.5, 2.0}
Delta Min	$\Delta_{\min}$	{0.1, 0.2, 0.3}
Problem Dimension	d	{4}
Objective Dimension	m	{3}
Number of Trials	-	10

C Empirical Validation

Experimental Setup and Parameter Sweep. We implemented the Adaptive Grid algorithm and evaluated it on a toy MOMDP. The environment evaluates policies of the form $\xi \in [0,1]^m$ with $\sum_i \xi_i \leq 1$, returning vector-valued rewards $V(\xi) = (\xi_1, \ldots, \xi_m, 1 - \|\xi\|_2)$. For each configuration, we compute the regret with respect to the optimal policy under the unknown utility $w^\star$ and record the number of preference queries used.

We performed a parameter sweep over

$$\epsilon \in \{0.5, 0.25, 0.125\}, \quad \delta \in \{0.05, 0.1, 0.2\}, \quad \alpha \in \{0.5, 1.0, 1.5, 2.0\}, \quad \Delta_{\min} \in \{0.1, 0.2, 0.3\},$$

Each setting was run for 10 independent trials with different random seeds.

Table 3: Best configuration from parameter sweep.

ϵ	δ	α	$\Delta_{\min}$	d	m	Performance
0.250	0.05	2.00	0.30	4	3	Mean Regret: 0.043624 ± 0.025638 Mean Queries: 41.40 ± 23.45

Results. Table 4 shows a comparison of our proposed Adaptive Grid to other state of the art methods and demonstrates how Adaptive Grid achieves lower mean regret with a competitive query count.

Algorithm	Mean Regret (95% CI)	Mean Queries (95% CI)
Adaptive Grid (Ours)	0.0436 ± 0.0256	41.4 ± 23.5
Active Pref-RL	0.1287 ± 0.0881	29.0 ± 0.0
Dueling Bandits	0.0685 ± 0.0645	1917.0 ± 0.0
MO-QD w/ Prefs	0.1511 ± 0.1079	57.8 ± 1.6

Table 4: Mean regret and query complexity on a toy MOMDP. Adaptive Grid compared with key baselines.

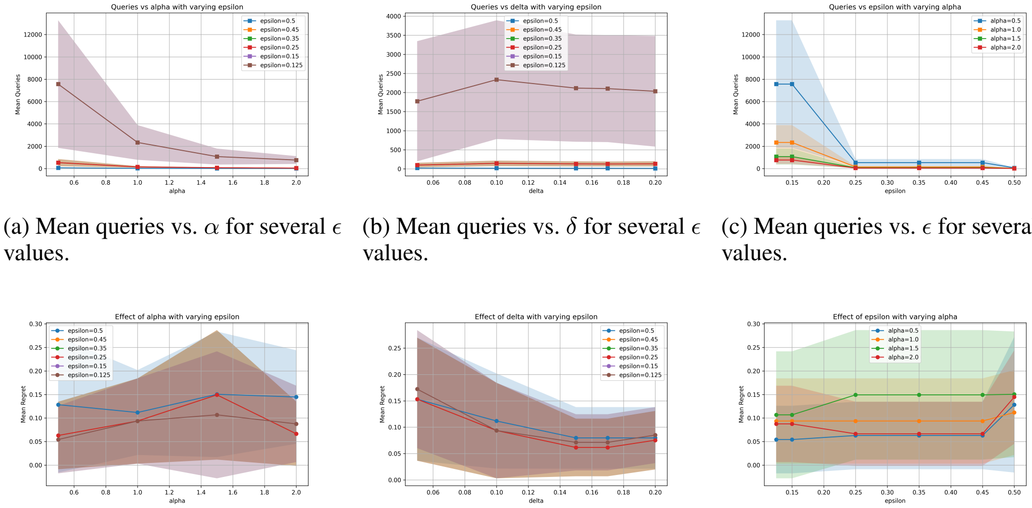
Ablation. Table 5 reports ablations of key components. Removing refinement or majority voting significantly degrades performance, validating the design choices of our algorithm.

Setting	Mean Regret	Mean Queries
No Refinement	0.108461 ± 0.052651	5.40 ± 1.71
No Majority Vote	0.133503 ± 0.088327	5.40 ± 1.25
No Oracle Noise	0.101576 ± 0.134584	46.40 ± 12.24
Fixed Grid Only	0.243046 ± 0.091096	0.00 ± 0.00

Table 5: Ablation study of Adaptive Grid. Both refinement and majority voting are crucial for minimizing regret and maintaining efficiency.

D Additional Figures

Figure 2 summarizes our parameter sweep results. The top row shows query complexity (mean number of preference queries) and the bottom row shows policy performance (mean regret). In all panels shaded regions indicate the 95% confidence interval across the 10 independent trials; point lines show the empirical mean. Each plot isolates the effect of one parameter while holding the others fixed, and legends indicate which variable is being varied within the plot.



(a) Mean queries vs. α for several ϵ values.

(b) Mean queries vs. δ for several ϵ values.

(c) Mean queries vs. ϵ for several α values.

(d) Mean regret vs. α for several ϵ values.

(e) Mean regret vs. δ for several ϵ values.

(f) Mean regret vs. ϵ for several α values.

Figure 2: Top row: query complexity (mean number of preference queries). Bottom row: policy performance (mean regret). Each panel varies one parameter on the x-axis while plotting multiple curves (legend) for the secondary parameter of interest; shaded regions correspond to the 95% confidence interval across 10 runs. All experiments use 50 iterations per run and the toy MOMDP described in Section 5.