

THE COMPARISON-BASED RL (CbRL) SETTING

WHY?

When humans evaluate multiple aspects of a problem, two or more can be in conflict (e.g., speed vs. safety), and **prevent** giving a **clear preference**

We introduce **Comparison-based Reinforcement Learning** (CbRL), where human feedback comprises four possible outcomes:

Direct Preference Option 1 is better than Option 2	Inverse Preference Option 2 is better than Option 1	Indifference Option 1 and Option 2 are equivalent	Incomparability Neither option dominates the other
--	---	---	--

FROM PbRL TO CbRL

PbRL (traditional)
Preferences only
Can only learn a **scalar** utility u



CbRL (ours)
Four types of feedback
Enables learning **vector-based** utility u

	Single Objective	Multi Objective
Rewards	RL	MORL
Human Feedback	PbRL	CbRL

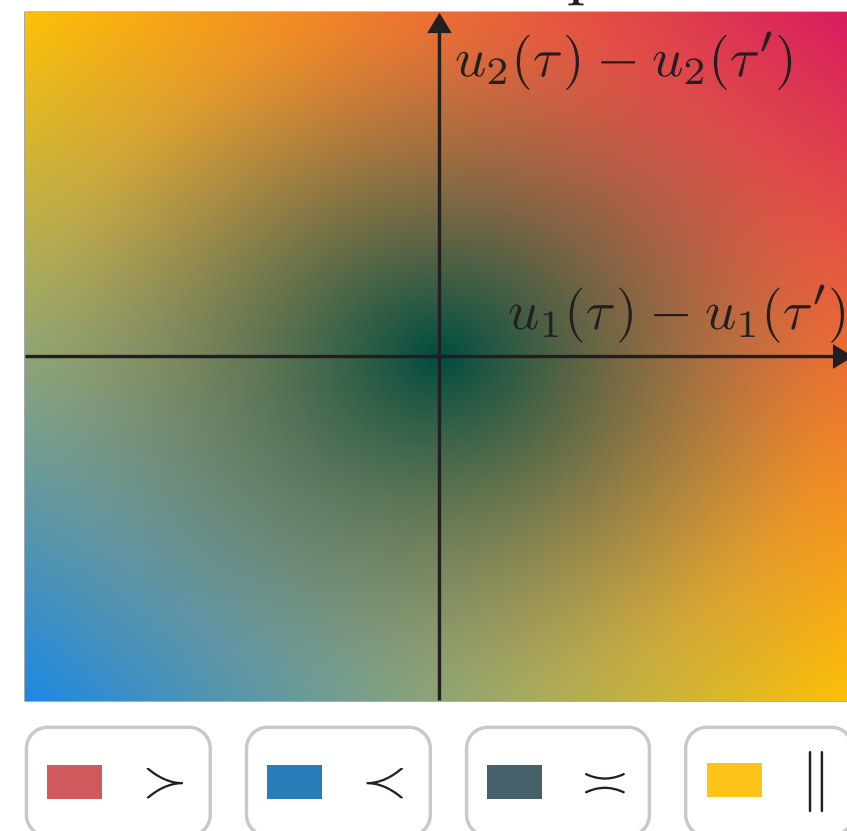
CbRL is the **multi-objective counterpart** of PbRL!

MODEL DESIDERATA

A rationality model should satisfy:

- $u(\tau) \geq u(\tau')$ (element-wise) and $u(\tau) \neq u(\tau')$ $\Rightarrow$ **prefer** τ ($>$)
- $u(\tau) \leq u(\tau')$ (element-wise) and $u(\tau) \neq u(\tau')$ $\Rightarrow$ **prefer** τ' ($<$)
- $u(\tau) = u(\tau')$ $\Rightarrow$ **indifferent** ($\asymp$)
- $u(\tau)$ is better than $u(\tau')$ w.r.t. some objectives but worse w.r.t. some others $\Rightarrow$ **incomparable** ($\parallel$)

Desiderata of a 2D problem



Key Theoretical Result (Proposition 2.1)

No model satisfying all desiderata can induce a **convex** negative log-likelihood

REFERENCES

- Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 1952.
- Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 2013.
- Christian Wirth, Riad Akrou, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 2017.

THE MOBT MODEL

Assumption: There exists a function mapping trajectories to vector-based utilities:

$$u : \mathcal{T} \rightarrow \mathbb{R}^d$$

Comparison outcomes are generated as **softmax** over four scores:

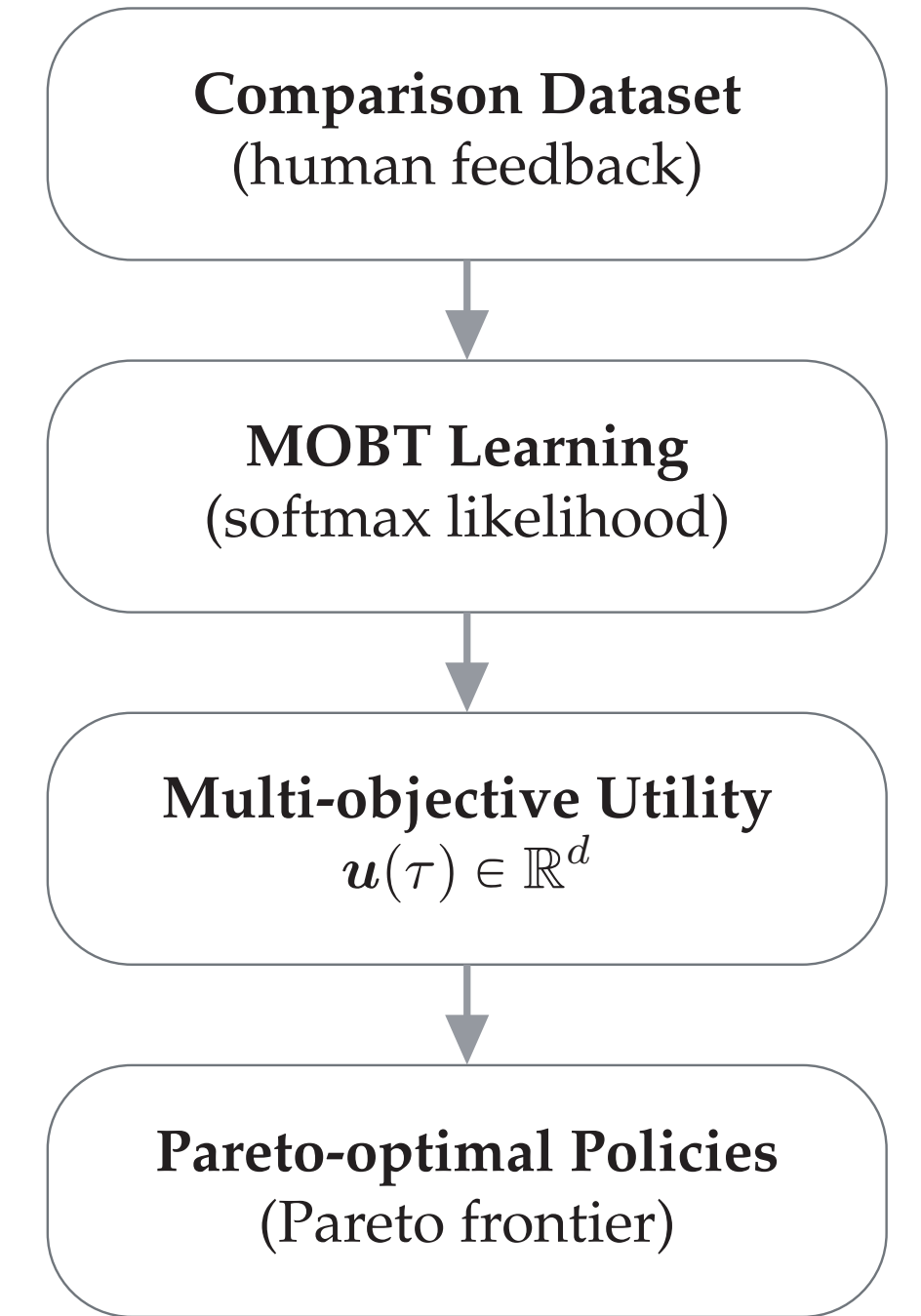
$$\begin{aligned} \text{Direct Preference } (>) & h_{>}(\tau, \tau') := \mathbf{1}_d^\top (u(\tau) - u(\tau')) / (2d) \\ \text{Inverse Preference } (<) & h_{<}(\tau, \tau') := -\mathbf{1}_d^\top (u(\tau) - u(\tau')) / (2d) \\ \text{Indifference } (\asymp) & h_{\asymp}(\tau, \tau') := \alpha \\ \text{Incomparability } (\parallel) & h_{\parallel}(\tau, \tau') := \sqrt{d} \text{std}(u(\tau) - u(\tau')) + \beta \end{aligned}$$

Probability of outcome $\circ \in \{>, <, \asymp, \parallel\}$

$$f(\circ | \tau, \tau') := \frac{\exp(h_{\circ}(\tau, \tau'))}{\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau'))}$$

Analysis of the model under the **linear utility function** $u(\tau) := W\phi(\tau)$

Learning Pipeline



THEORETICAL RESULTS

Lemma 3.1

MOBT **satisfies** all CbRL desiderata

Remark 3.1

MOBT reduces to Bradley-Terry model for single-objective problems ($d = 1$)

Theorem 3.2

KL divergence between true and learned model bound:

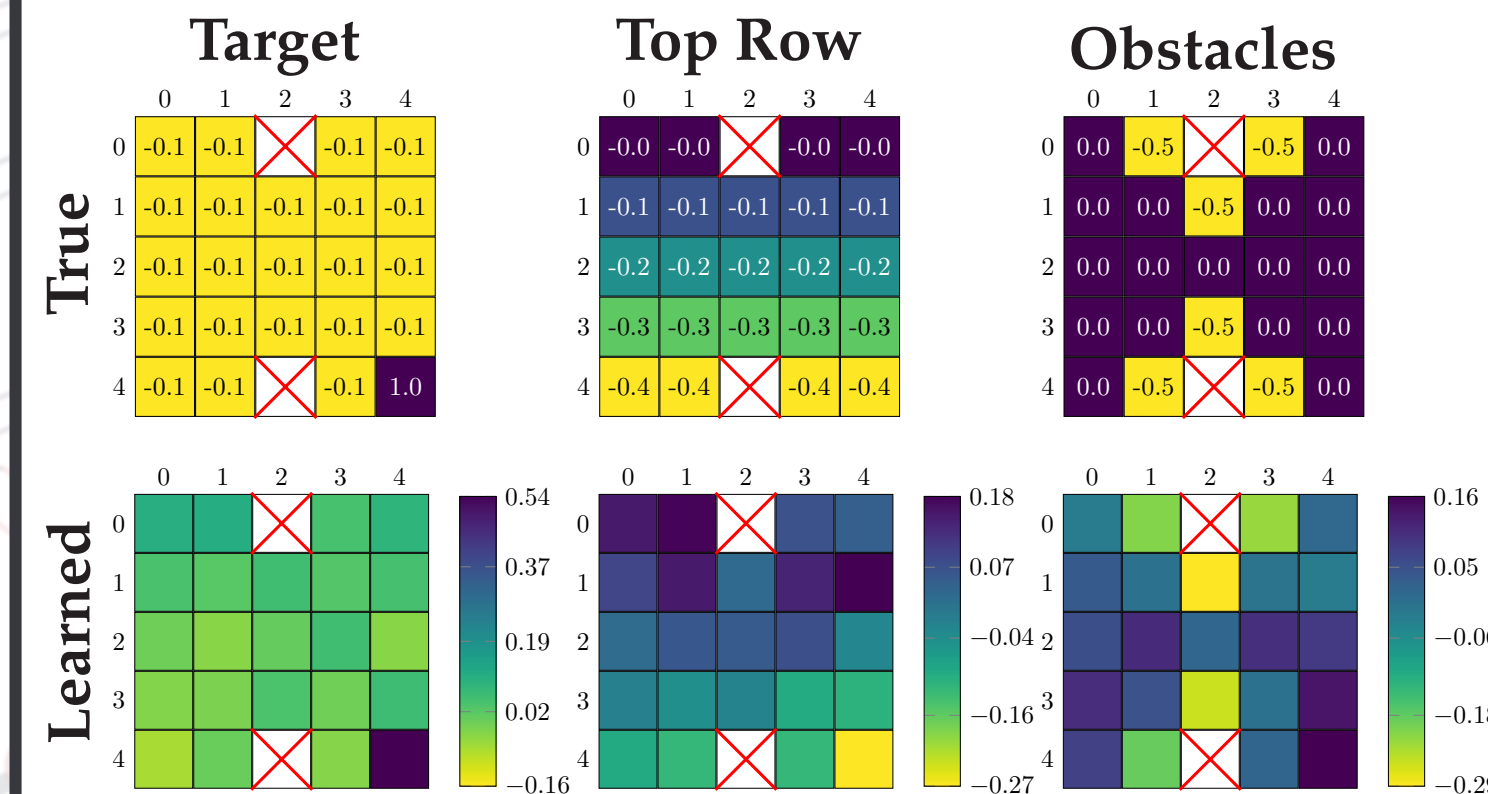
$$\tilde{O} \left(R \Delta d k \sqrt{\frac{\log(1/\delta)}{N}} \right)$$

Theorem 3.3

Generalization guarantees for **global** and **local** optima despite non-convexity

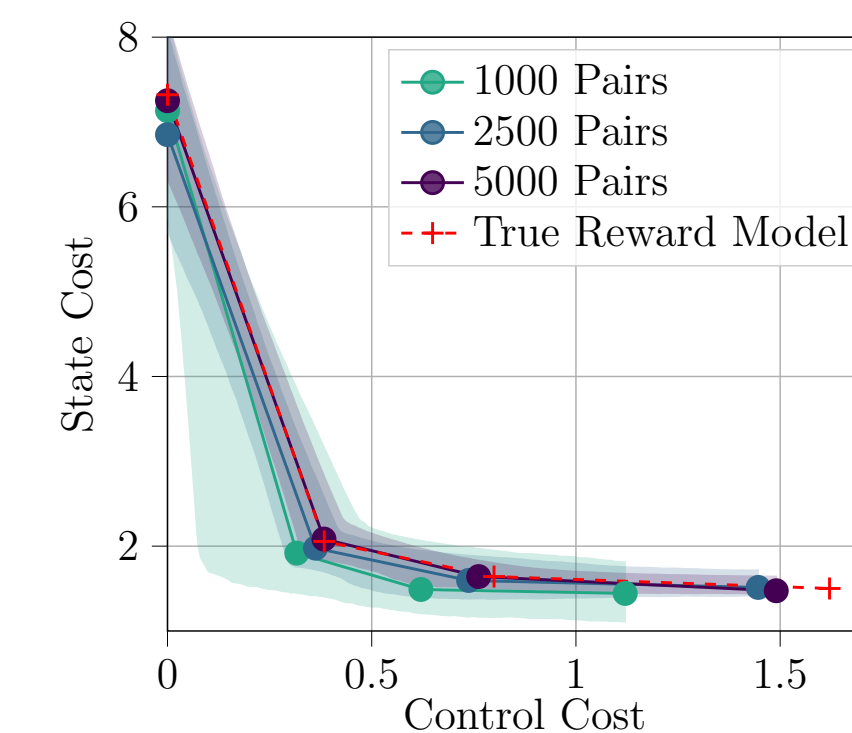
DOES IT WORK? EMPIRICAL VALIDATION

Reward Reconstruction



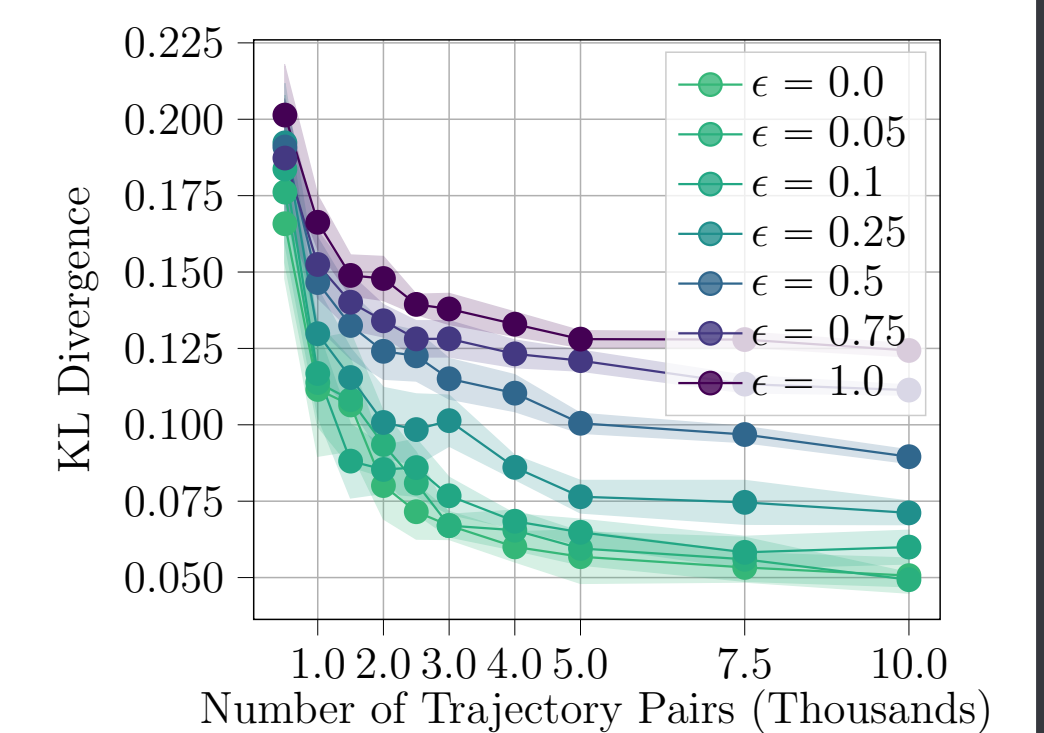
Custom MO-GridWorld environment

Pareto Frontier



MO-LQR Environment

Robustness to Irrationality



MO-Hopper environment