# QUANTIFYING DIFFERENCES IN REWARD FUNCTIONS

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

## ABSTRACT

For many tasks, the reward function is too complex to be specified procedurally, and must instead be learned from user data. Prior work has evaluated learned reward functions by examining rollouts from a policy optimized for the learned reward. However, this method cannot distinguish between the learned reward function failing to reflect user preferences, and the reinforcement learning algorithm failing to optimize the learned reward. Moreover, the rollout method is highly sensitive to details of the environment the learned reward is evaluated in, which often differ in the deployment environment. To address these problems, we introduce the *Equivalent-Policy Invariant Comparison (EPIC)* distance to quantify the difference between two reward functions directly, without training a policy. We prove EPIC is invariant on an equivalence class of reward functions that always induce the same optimal policy. Furthermore, we find EPIC can be precisely approximated and is more robust than baselines to the choice of coverage distribution. Finally, we show that EPIC distance bounds the regret of optimal policies even under different transition dynamics, and confirm empirically that it predicts policy training success. Our source code is available at — double blind: supplementary material —.

## 1 INTRODUCTION

Reinforcement learning (RL) has reached or surpassed human performance in many domains with clearly-defined reward functions, such as games [19; 14; 22] and narrowly-scoped robotic manipulation tasks [15]. Unfortunately, the reward functions for most real-world tasks are difficult or impossible to procedurally specify. Even a task as simple as peg insertion from pixels has a non-trivial reward function that must usually be learned [21, IV.A]. Most real-world tasks have far more complex reward functions than this. In particular, tasks involving human interaction depend on complex and user-dependent preferences. These challenges have inspired work on learning a reward function, whether from demonstrations [12; 16; 25; 8; 3], preferences [1; 24; 6; 17; 26] or both [10; 4].

Prior work usually evaluates the learned reward function $\hat{R}$ using the "rollout method": training a policy $\pi_{\hat{R}}$ to optimize $\hat{R}$ and then examining rollouts from $\pi_{\hat{R}}$. Unfortunately, this method is computationally expensive because it requires us to solve an RL problem. Furthermore, the rollout method produces *false negatives* when the reward $\hat{R}$ matches user preferences, but the RL algorithm fails to maximize $\hat{R}$. The rollout method also produces *false positives*: of the many reward functions inducing the desired rollout in a given environment, only a small subset align with the user's preferences. If the initial state distribution or transition dynamics change, misaligned rewards may induce undesirable policies.

Reinforcement learning is founded on the observation that it is usually easier and more robust to specify a reward function, rather than a policy maximizing that reward function. Applying this insight to reward function analysis, we develop methods to compare reward functions *directly*, without training a policy. We summarize our desiderata for reward function distances in Table 1.

We introduce the *Equivalent-Policy Invariant Comparison (EPIC)* pseudometric that meets all five desiderata. EPIC (section 4) canonicalizes the reward functions' potential-based shaping, then computes the correlation between the canonical rewards over a coverage distribution $\mathcal{D}$ of transitions. For comparison, we also propose two baselines (section 5), *Episode Return Correlation (ERC)* and *Nearest Point in Equivalence Class (NPEC)*, which partially satisfy the desiderata.

Table 1: Summary of the desiderata satisfied by each reward function distance. **Key** – the distance is: a *pseudometric* (section 3); *invariant* to potential shaping [13] and positive rescaling (section 3); a computationally *efficient* approximation achieving low error (section 6.1); *predictive* of the similarity of the trained policies (section 6.3); and *robust* to the choice of coverage distribution (section 6.2).

| Distance | Pseudometric | Invariant | Efficient | Predictive | Robust |
|---|---|---|---|---|---|
| EPIC | ✓ | ✓ | ✓ | ✓ | ✓ |
| NPEC | ✗ | ✓ | ✗ | ✓ | ✗ |
| ERC | ✓ | ✗ | ✓ | ✓ | ✗ |

EPIC works best when $\mathcal{D}$ has support on all realistic transitions. In our experiments, we achieve this by using uninformative priors, such as a uniform distribution over transitions. Moreover, we find EPIC is robust to the exact choice of distribution $\mathcal{D}$, producing similar results across a range of distributions, whereas ERC and especially NPEC are highly sensitive to the choice of $\mathcal{D}$ (section 6.2).

Reward learning algorithms are typically benchmarked on tasks with a known ground-truth reward function $R$. When using the rollout method, it is common to report the *regret*: how much less true reward $R$ is obtained by a policy $\pi_{\hat{R}}$ optimized for the learned reward $\hat{R}$ versus a policy $\pi_R$ optimized for $R$. Theorem 4.9 shows that reward functions with low EPIC distance to the true reward $R$ induce optimal policies with low regret even in unseen environments. We also confirm empirically that this result holds in practice (section 6.3). However, while reward functions with high EPIC distance tend to fail in the test environment, they sometimes still work in the training environment. EPIC therefore has a lower false positive rate than the rollout method, making it particularly attractive in safety-critical applications where reliability is a key design goal.

## 2 Related work

There exists a variety of methods to learn reward functions. One prominent family is inverse reinforcement learning (IRL; 12), which infers a reward function from demonstrations. The IRL problem is inherently underconstrained: many different reward functions can lead to the same demonstrations. Bayesian IRL [16] handles this ambiguity by inferring a posterior over reward functions. By contrast, Maximum Entropy IRL [25] selects the highest entropy reward function consistent with the demonstrations; this method has scaled to high-dimensional environments [7; 8].

An alternative approach is to learn from *preference comparisons* between two trajectories [1; 24; 6; 17]. T-REX [4] is a hybrid approach, learning from a *ranked* set of demonstrations. More directly, Cabi et al. [5] learn from "sketches" of cumulative reward over an episode.

To the best of our knowledge, there is no prior work that focuses on evaluating reward functions directly. The most closely related work is Ng et al. [13], identifying reward transformations guaranteed not to change the optimal policy. However, a variety of ad-hoc methods have been developed to evaluate reward functions. The rollout method – evaluating rollouts of a policy trained on the learned reward – is evident in the earliest work on IRL [12]. Fu et al. [8] refined the rollout method by testing on a transfer environment, inspiring our experiment in section 6.3. Recent work has compared reward functions by scatterplotting returns [10; 4], inspiring our ERC baseline (section 5.1).

## 3 Background

This section introduces material needed for the distances defined in subsequent sections. We start by defining a distance *metric*, then introduce the *Markov Decision Process (MDP)* formalism, and finally describe when reward functions induce the same optimal policy in any compatible MDP.

**Definition 3.1.** *Let $X$ be a set and $d : X \times X \to [0, \infty)$ a function. $d$ is a* premetric *if $d(x, x) = 0$ for all $x \in X$. $d$ is a* pseudometric *if, furthermore, for all $x, y, z \in X$, $d(x, y) = d(y, x)$ and $d(x, z) \leq d(x, y) + d(y, z)$. $d$ is a* metric *if, furthermore, for all $x, y \in X$, $d(x, y) = 0 \iff x = y$.*

We wish for $d(R_A, R_B) = 0$ when reward functions $R_A$ and $R_B$ are in the same equivalence class, even if $R_A \neq R_B$. This is forbidden in a metric but permitted in a pseudometric, while retaining

other guarantees such as symmetry and triangle inequality that a metric provides. Accordingly, a pseudometric is usually the best choice for a distance $d$ over reward functions.

**Definition 3.2.** *A* Markov Decision Process (MDP) $M = (\mathcal{S}, \mathcal{A}, \gamma, \mu, \mathcal{T}, R)$ *consists of a set of states $\mathcal{S}$ and a set of actions $\mathcal{A}$; a discount factor $\gamma \in [0, 1]$; an initial state distribution $\mu(s)$; a transition distribution $\mathcal{T}(s' \mid s, a)$ specifying the probability of transitioning to $s'$ from $s$ after taking action $a$; and a reward function $R(s, a, s')$ specifying the reward upon taking action $a$ in state $s$ and transitioning to state $s'$.*

A trajectory $\tau$ consists of a sequence of states and actions, $\tau = (s_0, a_0, s_1, a_1, \cdots)$, where each $s_i \in \mathcal{S}$ and $a_i \in \mathcal{A}$. The *return* on a trajectory is defined as the sum of discounted rewards, $g(\tau; R) = \sum_{t=0}^{|\tau|} \gamma^t R(s_t, a_t, s_{t+1})$, where the length of the trajectory $|\tau|$ may be infinite.

In the following, we assume a discounted ($\gamma < 1$) infinite-horizon MDP. The results can be generalized to undiscounted ($\gamma = 1$) MDPs subject to regularity conditions needed for convergence.

A *stochastic policy* $\pi(a \mid s)$ assigns probabilities to taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. The objective of an MDP is to find a policy $\pi$ that maximizes the expected return, $G(\pi) = \mathbb{E}_{\tau(\pi)}\left[g(\tau; R)\right]$, where $\tau(\pi)$ is a trajectory generated by sampling the initial state $s_0$ from $\mu$, each action $a_t$ from the policy $\pi(a_t \mid s_t)$ and successor states $s_{t+1}$ from the transition distribution $\mathcal{T}(s_{t+1} \mid s_t, a_t)$. An MDP $M$ has a set of optimal policies $\pi^*(M)$ that maximize the expected return, $\pi^*(M) = \arg\max_\pi G(\pi)$.

In this paper, we consider the setting where we only have access to an MDP\R, $M^- = (\mathcal{S}, \mathcal{A}, \gamma, \mu, \mathcal{T})$. The unknown reward function $R$ must be learned from human data. Typically, only the state space $\mathcal{S}$, action space $\mathcal{A}$ and discount $\gamma$ are known exactly, with the initial state $\mu$ and transition dynamics $\mathcal{T}$ only observable from interacting with the environment $M^-$. In the following, we describe an equivalence class whose members are guaranteed to have the same set of optimal policies in *any* MDP\R $M^-$ with fixed $\mathcal{S}$, $\mathcal{A}$ and $\gamma$ (allowing the unknown $\mathcal{T}$ and $\mu$ to take arbitrary values).

**Definition 3.3.** *A* potential shaping *reward is defined as $R(s, a, s') = \gamma\Phi(s') - \Phi(s)$, given a* potential $\Phi : \mathcal{S} \to \mathbb{R}$ *and where $\gamma$ is the MDP discount rate.*

**Definition 3.4** (Reward Equivalence)**.** *We define two bounded reward functions $R_A$ and $R_B$ to be* equivalent*, $R_A \equiv R_B$, for a fixed $(\mathcal{S}, \mathcal{A}, \gamma)$ if and only if there exists a constant $\lambda > 0$ and a bounded potential function $\Phi : \mathcal{S} \to \mathbb{R}$ such that for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$:*

$$R_B(s, a, s') = \lambda R_A(s, a, s') + \gamma\Phi(s') - \Phi(s). \tag{1}$$

Note $R_A - R_B \equiv \texttt{Zero}$ (where $\texttt{Zero}$ is the all-zero reward) if and only if $R_A \equiv R_B$ with $\lambda = 1$.

**Proposition 3.5.** *The binary relation $\equiv$ is an equivalence relation. Let $R_A, R_B, R_C : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be bounded reward functions. Then $\equiv$ is reflexive, $R_A \equiv R_A$; symmetric, $R_A \equiv R_B$ implies $R_B \equiv R_A$; and transitive, $(R_A \equiv R_B) \wedge (R_B \equiv R_C)$ implies $R_A \equiv R_C$.*

*Proof.* See section A.3.1 in supplementary material. □

The expected return of potential shaping $\gamma\Phi(s') - \Phi(s)$ on a trajectory segment $(s_0, \cdots, s_T)$ is $\gamma^T\Phi(s_T) - \Phi(s_0)$. The first term $\gamma^T\Phi(s_T) \to 0$ as $T \to \infty$, while the second term $\Phi(s_0)$ only depends on the initial state, and so potential shaping does not change the set of optimal policies [13].

Scaling a reward function by a positive factor $\lambda > 0$ scales the expected return of all trajectories by $\lambda$, leaving the set of optimal policies unchanged. The set of optimal policies is also invariant to a constant shift $c \in \mathbb{R}$ of the reward, however this can already be obtained by shifting $\Phi$ by $\frac{c}{\gamma - 1}$.*

If $R_A \equiv R_B$, for a fixed $(\mathcal{S}, \mathcal{A}, \gamma)$, then for any MDP\R $M^- = (\mathcal{S}, \mathcal{A}, \gamma, \mu, \mathcal{T})$ we have $\pi^*\left((M^-, R_A)\right) = \pi^*\left((M^-, R_B)\right)$, where $(M^-, R)$ denotes the MDP specified by $M^-$ with reward function $R$. In other words, $R_A$ and $R_B$ induce the same optimal policies for all initial state distributions $\mu$ and transition dynamics $\mathcal{T}$.

## 4 Comparing reward functions with EPIC

In this section we introduce the *Equivalent-Policy Invariant Comparison (EPIC)* pseudometric. This novel distance canonicalizes the reward functions' potential-based shaping, then compares

*Note constant shifts in the reward of an undiscounted MDP would cause the value function to diverge. Fortunately, the shaping $\gamma\Phi(s') - \Phi(s)$ is unchanged by constant shifts to $\Phi$ when $\gamma = 1$.

the canonical representatives using Pearson distance, which is invariant to scale. Together, this construction makes EPIC invariant on reward equivalence classes. See section A.3.2 for proofs.

We define the *canonically shaped reward* $C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)$ as an expectation over some arbitrary distributions $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{A}$ over states $\mathcal{S}$ and actions $\mathcal{A}$ respectively. This construction means $C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)$ only depends on $(\mathcal{S}, \mathcal{A}, \gamma)$, and not on the initial state distribution $\mu$ or transition dynamics $\mathcal{T}$. In particular, we may evaluate $R$ on transitions that are impossible in the training environment, since these may become possible in a deployment environment with a different $\mu$ or $\mathcal{T}$.

**Definition 4.1** (Canonically Shaped Reward)**.** *Let $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be a reward function. Given distributions $\mathcal{D}_\mathcal{S} \in \Delta(\mathcal{S})$ and $\mathcal{D}_\mathcal{A} \in \Delta(\mathcal{A})$ over states and actions, let $S$ and $S'$ be random variables independently sampled from $\mathcal{D}_\mathcal{S}$ and $A$ sampled from $\mathcal{D}_\mathcal{A}$. We define the* canonically shaped *$R$ to be:*

$$C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)(s,a,s') = R(s,a,s') + \mathbb{E}\left[\gamma R(s',A,S') - R(s,A,S') - \gamma R(S,A,S')\right]. \quad (2)$$

Informally, if $R'$ is shaped by potential $\Phi$, then increasing $\Phi(s)$ decreases $R'(s,a,s')$ but increases $\mathbb{E}\left[-R'(s,A,S')\right]$, canceling. Similarly, increasing $\Phi(s')$ increases $R'(s,a,s')$ but decreases $\mathbb{E}\left[\gamma R'(s',A,S')\right]$. Finally, $\mathbb{E}[R(S,A,S')]$ centers the reward, canceling constant shift.

**Proposition 4.2** (The Canonically Shaped Reward is Invariant to Shaping)**.** *Let $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be a reward function and $\Phi : \mathcal{S} \to \mathbb{R}$ a potential function. Let $\gamma \in [0,1]$ be a discount rate, and $\mathcal{D}_\mathcal{S} \in \Delta(\mathcal{S})$ and $\mathcal{D}_\mathcal{A} \in \Delta(\mathcal{A})$ be distributions over states and actions. Let $R'$ denote $R$ shaped by $\Phi$: $R'(s,a,s') = R(s,a,s') + \gamma\Phi(s') - \Phi(s)$. Then the canonically shaped $R'$ and $R$ are equal: $C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R') = C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)$.*

Proposition 4.2 holds for arbitrary distributions $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{A}$. However, in the following Proposition we show that the potential shaping introduced by the canonicalization $C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)$ is more influenced by perturbations to $R$ of higher joint probability transitions $(s,a,s')$. This suggests choosing $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{A}$ to have broad support, making $C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)$ more robust to perturbations of any given transition.

**Proposition 4.3.** *Let $\mathcal{S}$ and $\mathcal{A}$ be finite, with $|\mathcal{S}| \geq 2$. Let $\mathcal{D}_\mathcal{S} \in \Delta(\mathcal{S})$ and $\mathcal{D}_\mathcal{A} \in \Delta(\mathcal{A})$. Let $R$, $\nu : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be reward functions, with $\nu(s,a,s') = \lambda\mathbb{I}[(s,a,s') = (x,u,x')]$, $\lambda \in \mathbb{R}$, $x, x' \in \mathcal{S}$ and $u \in \mathcal{A}$. Let $\Phi_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)(s,a,s') = C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)(s,a,s') - R(s,a,s')$. Then:*

$$\left\|\Phi_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R+\nu) - \Phi_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)\right\|_\infty = \lambda\left(1 + \gamma\mathcal{D}_\mathcal{S}(x)\right)\mathcal{D}_\mathcal{A}(u)\mathcal{D}_\mathcal{S}(x'). \quad (3)$$

We have canonicalized potential shaping; next, we compare the rewards in a scale-invariant manner.

**Definition 4.4.** *The* Pearson distance *between random variables $X$ and $Y$ is defined by the expression $D_\rho(X,Y) = \frac{1}{\sqrt{2}}\sqrt{1-\rho(X,Y)}$, where $\rho(X,Y)$ is the Pearson correlation between $X$ and $Y$.*

**Lemma 4.5.** *The Pearson distance $D_\rho$ is a pseudometric. Moreover, let $a, b \in (0,\infty)$, $c, d \in \mathbb{R}$ and $X, Y$ be random variables. Then it follows that $0 \leq D_\rho(aX+c, bY+d) = D_\rho(X,Y) \leq 1$.*

We can now define EPIC in terms of the Pearson distance between canonically shaped rewards.

**Definition 4.6** (Equivalent-Policy Invariant Comparison (EPIC) pseudometric)**.** *Let $\mathcal{D}$ be some coverage distribution over transitions $s \xrightarrow{a} s'$. Let $S, A, S'$ be random variables jointly sampled from $\mathcal{D}$. Let $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{A}$ be some distributions over states $\mathcal{S}$ and $\mathcal{A}$ respectively. The* Equivalent-Policy Invariant Comparison (EPIC) *distance between reward functions $R_A$ and $R_B$ is:*

$$D_{\mathrm{EPIC}}(R_A, R_B) = D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_B)(S,A,S')\right). \quad (4)$$

**Theorem 4.7.** *The Equivalent-Policy Invariant Comparison distance is a pseudometric.*

The triangle inequality is particularly important. For example, consider an environment with an expensive to evaluate ground-truth reward $R$. Directly comparing many learned rewards $\hat{R}$ to $R$ might be prohibitively expensive. We can instead pay a one-off cost: query $R$ a finite number of times and infer a proxy reward $R_P$ with $D_{\mathrm{EPIC}}(R, R_P) \leq \epsilon$. The triangle inequality allows us to evaluate $\hat{R}$ via comparison to $R_P$, since $D_{\mathrm{EPIC}}(\hat{R}, R) \leq D_{\mathrm{EPIC}}(\hat{R}, R_P) + \epsilon$. This is particularly useful for benchmarks, which may be expensive to build but must be cheap to use.

**Theorem 4.8.** *Let $R_A$, $R'_A$, $R_B, R'_B : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be reward functions such that $R'_A \equiv R_A$ and $R'_B \equiv R_B$. Then $0 \leq D_{\mathrm{EPIC}}(R'_A, R'_B) = D_{\mathrm{EPIC}}(R_A, R_B) \leq 1$.*

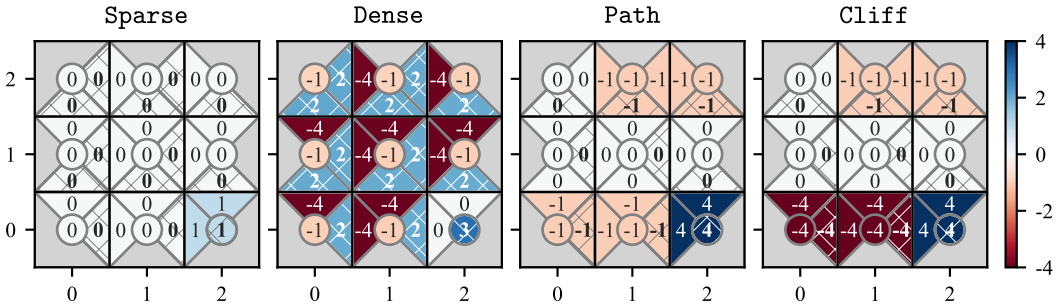



Figure 1: Heatmaps of reward functions $R(s, a, s')$ for a deterministic $3 \times 3$ gridworld. $R(s, \text{stay}, s)$ is given by the central circle in cell $s$. $R(s, a, s')$ is given by the triangular wedge in cell $s$ adjacent to cell $s'$ in direction $a$. Optimal action(s) (for infinite horizon, discount $\gamma = 0.99$) have bold labels against a hatched background. See figure A.2 for the distances between all reward pairs.

The following is our main theoretical result, showing EPIC distance gives a bounds on the regret between policies optimizing for one of the two reward functions relative to the other. In other words, EPIC bounds the difference in returns for policies optimal for either reward function.

**Theorem 4.9.** *Let $M$ be a $\gamma$-discounted MDP\R with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$. Let $R_A, R_B : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be rewards, and $\pi_A^*, \pi_B^*$ be respective optimal policies. Let $\mathcal{D}_\pi(t, s_t, a_t, s_{t+1})$ denote the distribution over transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ induced by policy $\pi$ at time $t$, and $\mathcal{D}(s, a, s')$ be the coverage distribution used to compute $D_{\text{EPIC}}$. Suppose there exists $K > 0$ such that $K\mathcal{D}(s_t, a_t, s'_{t+1}) \geq \mathcal{D}_\pi(t, s_t, a_t, s'_{t+1})$ for all times $t \in \mathbb{N}$, triples $(s_t, a_t, s_{t+1}) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and policies $\pi \in \{\pi_A^*, \pi_B^*\}$. Then the regret under $R_A$ from executing $\pi_B^*$ instead of $\pi_A^*$ is at most $G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq 16K\|R_A\|_2 \, (1-\gamma)^{-1} \, D_{\text{EPIC}}(R_A, R_B)$.*

We generalize the regret bound to continuous spaces in theorem A.15 via a Lipschitz assumption, with Wasserstein distance replacing $K$. Importantly, the returns of $\pi_A^*$ and $\pi_B^*$ **converge** as $D_{\text{EPIC}}(R_A, R_B) \to 0$ in both cases, no matter which reward function you evaluate on.

The key assumption is that the coverage distribution $\mathcal{D}$ has adequate support for transitions occurring in rollouts of $\pi_A^*$ and $\pi_B^*$. The bound is tightest when $\mathcal{D}$ is similar to $\mathcal{D}_{\pi_A^*}$ and $\mathcal{D}_{\pi_B^*}$. However, computing $\pi_A^*$ and $\pi_B^*$ is often intractable. The MDP $M$ may be unknown, such as when making predictions about an unseen deployment environment. Even when $M$ is known, RL is computationally expensive and in non-trivial environments may fail to converge.

In finite cases, a uniform $\mathcal{D}$ satisfies the requirements with $K \leq |\mathcal{S}|^2|\mathcal{A}|$. In general, it is best to choose $\mathcal{D}$ to have broad coverage over plausible transitions. Broad coverage ensures adequate support for $\mathcal{D}_{\pi_A^*}$ and $\mathcal{D}_{\pi_B^*}$. Excluding transitions that are unlikely or impossible to occur gives tighter regret bounds due to a smaller $K$ (finite case) or Wasserstein distance (continuous case).

While EPIC upper bounds policy regret, it does not lower bound it. In fact, no reward distance can lower bound regret in arbitrary environments. For example, suppose the deployment environment transitions to a randomly chosen state independent of the action taken. In this case, all policies obtain the same expected return, so the policy regret is always zero, regardless of the reward functions.

As a pedagogical example, we compute the EPIC distance between the reward functions in figure 1 for a deterministic $3 \times 3$ gridworld. Despite assigning different rewards to each transition, `Sparse` and `Dense` are equivalent and have zero EPIC distance. By contrast, $D_{\text{EPIC}}(\texttt{Path}, \texttt{Cliff}) = 0.27$, almost as much as $D_{\text{EPIC}}(\texttt{Sparse}, \texttt{Cliff}) = 0.37$. Although `Path` and `Cliff` have identical optimal policies in deterministic settings, the rewards induce very different optimal policies under stochastic dynamics. See figure A.2 for the distances between all reward pairs.

For this example, we used state and action distributions $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{A}$ uniform over $\mathcal{S}$ and $\mathcal{A}$, and coverage distribution $\mathcal{D}$ uniform over state-action pairs $(s, a)$, with $s'$ deterministically computed. It is important these distributions have adequate support. As an extreme example, if $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}$ have no support for a particular state then the reward of that state has no effect on the distance. We can compute EPIC exactly in a tabular setting, but in general use a sample-based approximation (section A.1.1).

# 5 BASELINE APPROACHES FOR COMPARING REWARD FUNCTIONS

To the best of our knowledge, EPIC is the first method to quantitatively evaluate reward functions without training a policy. Given the lack of established methods, we develop two alternatives as baselines: Episode Return Correlation (ERC) and Nearest Point in Equivalence Class (NPEC).

## 5.1 EPISODE RETURN CORRELATION (ERC)

The goal of an MDP is to maximize expected episode return, so it is natural to compare reward functions by the returns they induce. If the return of a reward function $R_A$ is a positive affine transformation of another reward $R_B$, then $R_A$ and $R_B$ have the same set of optimal policies. This suggests using Pearson distance, which is invariant to positive affine transformations.

**Definition 5.1** (Episode Return Correlation (ERC) pseudometric)**.** *Let $\mathcal{D}$ be some distribution over trajectories. Let $E$ be a random variable sampled from $\mathcal{D}$. The* Episode Return Correlation *distance between reward functions $R_A$ and $R_B$ is the Pearson distance between their episode returns on $\mathcal{D}$,* $D_{\mathrm{ERC}}(R_A, R_B) = D_\rho(g(E; R_A), g(E; R_B))$.

Prior work has scatterplot the return of $R_A$ against $R_B$ over episodes [4, figure 3] and fixed-length segments [10, section D]. ERC is the Pearson distance of such plots, so is a natural baseline. We approximate ERC by the correlation of episode returns on a finite collection of rollouts.

Under special conditions, ERC is invariant to shaping. Let $R$ be a reward function and $\Phi$ a potential function, and define the shaped reward $R'(s, a, s') = R(s, a, s') + \gamma\Phi(s') - \Phi(s)$. The return under the shaped reward on a trajectory $\tau = (s_0, a_0, \cdots, s_T)$ is $g(\tau; R') = g(\tau; R) + \gamma^T\Phi(s_T) - \Phi(s_0)$.

If the initial state $s_0$ and terminal state $s_T$ are fixed, then $\gamma^T\Phi(s_T) - \Phi(s_0)$ is constant. Since Pearson distance is invariant to constant shifts, ERC is invariant to shaping in this case. For infinite-horizon discounted MDPs, only the initial state $s_0$ need be fixed, since $\gamma^T\Phi(s_T) \to 0$ as $T \to \infty$.

However, if the initial state $s_0$ is stochastic, the ERC distance can take on arbitrary values under shaping. Let $R_A$ and $R_B$ be two arbitrary reward functions. Suppose that there are at least two distinct initial states, $s_A$ and $s_B$, with non-zero measure in $\mathcal{D}$. Choose potential $\Phi(s) = 0$ everywhere except $\Phi(s_A) = \Phi(s_B) = c$, and let $R'_A$ and $R'_B$ denote $R_A$ and $R_B$ shaped by $\Phi$. As $c \to \infty$, the correlation $\rho\left(g(E; R'_A), g(E; R'_B)\right)$ tends to one. This is since the relative difference tends to zero, even though $g(E; R'_A)$ and $g(E; R'_B)$ continue to have the same absolute difference as $c$ varies. Consequently, the ERC pseudometric $D_{\mathrm{ERC}}(R'_A, R'_B) \to 0$ as $c \to \infty$. By an analogous argument, setting $\Phi(s_A) = c$ and $\Phi(s_B) = -c$ gives $D_{\mathrm{ERC}}(R'_A, R'_B) \to 1$ as $c \to \infty$.

## 5.2 NEAREST POINT IN EQUIVALENCE CLASS (NPEC)

NPEC takes the minimum $L^p$ distance between equivalence classes. See section A.3.3 for proofs.

**Definition 5.2** ($L^p$ distance)**.** *Let $\mathcal{D}$ be a coverage distribution over transitions $s \xrightarrow{a} s'$ and let $p \geq 1$ be a power. The $L^p$* distance *between reward functions $R_A$ and $R_B$ is the $L^p$ norm of their difference:*

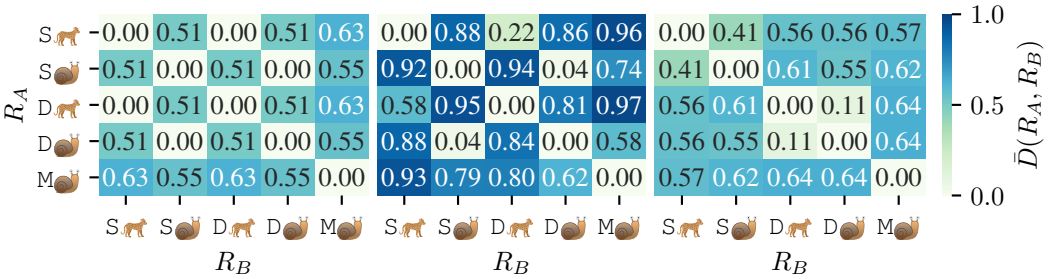
$$D_{L^p,\mathcal{D}}(R_A, R_B) = \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |R_A(s, a, s') - R_B(s, a, s')|^p \right] \right)^{1/p}. \tag{5}$$

**Proposition 5.3.** *(1) $D_{L^p,\mathcal{D}}$ is a metric in $L^p$ space, where functions $f$ and $g$ are identified if they agree almost everywhere on $\mathcal{D}$. (2) $D_{L^p,\mathcal{D}}$ is a pseudometric if functions are identified only if they agree at all points.*

The $L^p$ distance is affected by shaping and positive rescaling despite not changing the optimal policy. A natural solution is to take the distance from the *nearest point* in the equivalence class: $D^U_{\mathrm{NPEC}}(R_A, R_B) = \inf_{R'_A \equiv R_A} D_{L^p,\mathcal{D}}(R'_A, R_B)$. Unfortunately, $D^U_{\mathrm{NPEC}}$ is sensitive to $R_B$'s scale.

It is tempting to instead take the infimum over both arguments of $D_{L^p,\mathcal{D}}$. However, $\inf_{R'_A \equiv R_A, R'_B \equiv R_B} D_{L^p,\mathcal{D}}(R'_A, R_B) = 0$ since all equivalence classes come arbitrarily close to the origin in $L^p$ space. Instead, we fix this by normalizing $D^U_{\mathrm{NPEC}}$.

**Definition 5.4.** *The* Nearest Point in Equivalence Class (NPEC) premetric *is defined by:* $D_{\mathrm{NPEC}}(R_A, R_B) = \frac{D^U_{\mathrm{NPEC}}(R_A, R_B)}{D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B)}$ *when* $D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B) \neq 0$ *and* $0$ *otherwise.*




(a) EPIC (b) NPEC (c) ERC

Figure 2: Approximate distances between hand-designed reward functions in `PointMass`. The coverage distribution $\mathcal{D}$ is sampled from rollouts of a policy $\pi_{\text{uni}}$ taking actions uniformly at random. **Key**: quadratic control penalty, no control penalty. `S` is $\text{Sparse}(x) = \mathbb{1}[|x| < 0.05]$, `D` is shaped $\text{Dense}(x, x') = \text{Sparse}(x) + |x'| - |x|$, while `M` is $\text{Magnitude}(x) = -|x|$. Width of 95% confidence interval (see Figure A.3) is less than 0.02 for EPIC and ERC but as large as 0.3 for NPEC.

If $D^U_{\text{NPEC}}(\texttt{Zero}, R_B) = 0$ then $D^U_{\text{NPEC}}(R_A, R_B) = 0$ since $R_A$ can be scaled arbitrarily close to `Zero`. Since all policies are optimal for $R \equiv \texttt{Zero}$, we choose $D_{\text{NPEC}}(R_A, R_B) = 0$ in this case.

**Theorem 5.5.** *$D_{\text{NPEC}}$ is a premetric on the space of bounded reward functions. Moreover, let $R_A, R_A{}', R_B, R_B{}' : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be bounded reward functions such that $R_A \equiv R_A{}'$ and $R_B \equiv R_B{}'$. Then $0 \leq D_{\text{NPEC}}(R_A{}', R_B{}') = D_{\text{NPEC}}(R_A, R_B) \leq 1$.*

Note that $D_{\text{NPEC}}$ may not be symmetric and so is not, in general, a pseudometric: see proposition A.2.

The infimum in $D^U_{\text{NPEC}}$ can be computed exactly in a tabular setting, but in general we must approximate it using gradient descent. This gives an upper bound for $D^U_{\text{NPEC}}$, but the quotient of upper bounds $D_{\text{NPEC}}$ may be too low or too high. See section A.1.2 for details of the approximation.

## 6 Experiments

We evaluate EPIC and the baselines ERC and NPEC in a variety of continuous control tasks. First, we compute the distance between hand-designed reward functions, finding EPIC to be the most reliable distance. Although NPEC produces qualitatively similar results, it has a high degree of approximation error. Moreover, ERC sometimes suffers from pathological failures, such as assigning a high distance to equivalent rewards. Second, we find the distance of learned reward functions to a ground-truth reward predicts the return obtained by policy training, even in an unseen test environment. Finally, we show EPIC is robust to the exact choice of coverage distribution $\mathcal{D}$, whereas ERC and especially NPEC are highly sensitive to the choice of $\mathcal{D}$.

### 6.1 Comparing hand-designed reward functions

We compare procedurally specified reward functions in four tasks. Figure 2 presents results in the proof-of-concept `PointMass` task. The results for `Gridworld`, `HalfCheetah` and `Hopper`, in section A.2.4, are qualitatively similar. In `PointMass` the agent can accelerate left or right on a line. The reward functions include ( ) or exclude ( ) a quadratic penalty $\|a\|_2^2$. The sparse reward (`S`) gives a reward of 1 in the region around the goal state. The dense reward (`D`) is a shaped version of the sparse reward. The magnitude reward (`M`) is the negative distance of the agent from the goal.

We find that EPIC correctly identifies the equivalent reward pairs (`S`-`D` and `S`-`D`) with estimated distance $< 1 \times 10^{-3}$. By contrast, NPEC has substantial approximation error: $D_{\text{NPEC}}(\texttt{D}, \texttt{S}) = 0.58$. Similarly, $D_{\text{ERC}}(\texttt{D}, \texttt{S}) = 0.56$ due to ERC's erroneous handling of stochastic initial states.

Moreover, NPEC is computationally inefficient: Figure 2(b) took 31 hours to compute. By contrast, the figures for EPIC and ERC were each generated in under an hour. See section A.2.6 for more information on runtimes.

Table 2: Approximate distances of reward functions from ground-truth (GT), and the mean GT return of policies optimized from-scratch for the reward model in the *train* and *test* variants of PointMaze. We also report returns for AIRL's *generator* policy, jointly trained with the reward. Distances ($1000\times$ scale) use coverage distribution $\mathcal{D}$ from rollouts in the *train* environment of: a uniform random policy $\boldsymbol{\pi}_{\mathbf{uni}}$, an expert $\boldsymbol{\pi}^*$ and a **Mix**ture of these policies. $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{A}}$ are computed by marginalizing $\mathcal{D}$. See Table A.7 for 95% confidence intervals.

| **Reward** | $1000 \times D_{\mathrm{EPIC}}$ | | | $1000 \times D_{\mathrm{NPEC}}$ | | | $1000 \times D_{\mathrm{ERC}}$ | | | **Episode Return** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Function** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | **Gen.** | **Train** | **Test** |
| GT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | — | $-5.99$ | $-6.05$ |
| Regress | 42.2 | 36.4 | 25.9 | 0.47 | 13.5 | 0.15 | 4.80 | 40.7 | 1.40 | — | $-6.99$ | $-6.51$ |
| Pref | 50.8 | 54.4 | 32.9 | 3.08 | 217 | 2.11 | 15.1 | 179 | 8.12 | — | $-6.62$ | $-7.02$ |
| AIRL SO | 488 | 599 | 392 | 682 | 1089 | 469 | 447 | 380 | 232 | $-9.44$ | $-28.5$ | $-11.7$ |
| AIRL SA | 549 | 614 | 388 | 871 | 1041 | 443 | 507 | 465 | 208 | $-6.69$ | $-6.91$ | $-28.5$ |
| Mirage | 37.3 | 0.03 | 417 | 8.80 | 3.78 | 1580 | 32 | <0.01 | 166 | — | $-34.2$ | $-46.8$ |

### 6.2 Sensitivity of reward distance to coverage distribution

We would like the reward distances to be robust to the exact choice of coverage distribution $\mathcal{D}$. In Table 2, we report distances to the ground-truth (GT) from reward functions on the PointMaze task from Fu et al. [8], calculated under distributions induced by rollouts from three different policies. $\pi_{\mathrm{uni}}$ takes actions uniformly at random, producing broad support over transitions; $\pi^*$ is an expert policy, yielding a distribution concentrated at or near the goal; and **Mix** is a mixture of the two. In EPIC, $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{A}}$ are marginalized from $\mathcal{D}$ and so also vary between conditions.

PointMaze is a MuJoCo environment where a point mass agent must navigate around a wall to reach a goal. We evaluate four reward learning algorithms: Regression onto reward labels [*target* method from 6, section 3.3], Preference comparisons on trajectories [6], and adversarial IRL with a state-only (AIRL SO) and state-action (AIRL SA) reward model [8]. All models are trained using synthetic data from an oracle with access to the ground-truth; see section A.2.2 for details.

We find EPIC is robust to varying $\mathcal{D}$ when comparing the learned reward models: the distance varies by less than $2\times$, and the ranking between the reward models is the same across coverage distributions, except for **Mix** narrowly favoring AIRL SA over AIRL SO. By contrast, NPEC is highly sensitive to $\mathcal{D}$: the ratio of AIRL SO to Pref is over $200:1$ under $\pi_{\mathrm{uni}}$ but only $5:1$ under $\pi^*$. ERC lies somewhere in the middle: the ratio is $30:1$ under $\pi_{\mathrm{uni}}$ and $2:1$ under $\pi^*$.

We evaluate the effect of pathological choices of coverage distribution $\mathcal{D}$ in Table A.8. For example, **Ind** independently samples states and next states, giving physically impossible transitions, while **Jail** constrains rollouts to a tiny region that excludes the goal. We found that EPIC maintains the same ranking even under these extreme conditions, whilst the ranking of NPEC and ERC change.

However, the hand-designed reward function Mirage shows that, even with EPIC, care must be taken when choosing $\mathcal{D}$. Mirage assigns a larger reward when close to a "mirage" state than when at the true goal, but is identical to GT at all other points. The "mirage" state is rarely visited by random exploration $\pi_{\mathrm{uni}}$ as it is far away and on the opposite side of the wall from the agent. The expert policy $\pi^*$ is even less likely to visit it, as it is not on or close to the optimal path to the goal. As a result, the EPIC distance from Mirage to GT (Table 2, bottom row) is small under $\pi_{\mathrm{uni}}$ and $\pi^*$.

In general, any black-box method for assessing reward models – including the rollout method – only has predictive power on transitions visited during testing. Fortunately, we can achieve a broad support over states with **Mix**: it often navigates around the wall due to $\pi^*$, but strays from the goal thanks to $\pi_{\mathrm{uni}}$. As a result, EPIC under **Mix** correctly infers that Mirage is far from the ground-truth GT.

These empirical results support our theoretically inspired recommendation from section 4: "in general, it is best to choose $\mathcal{D}$ to have broad coverage over plausible transitions." Distributions such as $\pi^*$ are too narrow, assigning coverage only on a direct path from the initial state to the goal. Very broad distributions such as **Ind** waste probability mass on impossible transitions like teleporting. Distributions like **Mix** strike the right balance between these extremes.

### 6.3 PREDICTING POLICY PERFORMANCE FROM REWARD DISTANCE

`PointMaze` consists of a *train* and *test* variant, differing in the position of the wall. All four learned reward models were trained in data from the *train* variant. In Table 2, we report the ground-truth (`GT`) return of policies optimized in each variant for each reward function. We find that rewards with low distance from `GT` achieve high returns in both the *train* and *test* variant. By contrast, high distance rewards sometimes work in the *train* variant, but fail to transfer to the *test* variant.

Both `Regress` and `Pref` achieve very low distances, producing near-expert policy performance in both the *train* and *test* variants. The `AIRL SO` and `AIRL SA` models have distances an order of magnitude greater. The more expressive `AIRL SA` achieves near-expert performance in *train* but fails to transfer to *test*. The less-expressive `AIRL SO` has poor performance in both variants, although the generator policy (trained simultaneously with the reward) performs reasonably in *train*. The high distance of `Repel` and `Mirage` under **Mix** correctly predicts low policy returns.

Due to reward ambiguity, rewards such as `AIRL` $\star$ that are distant from the ground-truth `GT` can still produce a good policy. For example, a "memorized" reward function that assigns reward only to states visited by an expert will induce the expert policy in the *train* variant. Nonetheless, it will have a large distance from `GT`, even if the coverage distribution $\mathcal{D}$ only contains transitions from *train*. This is appropriate since in *test* the "expert" policy produced by this reward runs straight into the wall.

## 7 CONCLUSION

Our novel EPIC distance compares reward functions directly, without training a policy. We have proved it satisfies the axioms of a pseudometric, is bounded and invariant to equivalent rewards, and bounds the regret of optimal policies. Empirically, we find the EPIC distance between procedurally specified reward functions is more reliable than the NPEC and ERC baselines. Furthermore, we find the distance of learned reward functions to the ground-truth reward predicts the return of policies optimized for the learned reward, in both the *train* and unseen *test* environments.

## REFERENCES

[1] Riad Akrour, Marc Schoenauer, and Michele Sebag. Preference-based policy learning. In *Machine Learning and Knowledge Discovery in Databases*, 2011.

[2] Dario Amodei, Paul Christiano, and Alex Ray. Learning from human preferences, June 2017. URL https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/.

[3] Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Arian Hosseini, Pushmeet Kohli, and Edward Grefenstette. Learning to understand goal specifications by modelling reward. In *ICLR*, 2019.

[4] Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *ICML*, 2019.

[5] Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, Oleg Sushkov, David Barker, Jonathan Scholz, Misha Denil, Nando de Freitas, and Ziyu Wang. Scaling data-driven robotics with reward sketching and batch reinforcement learning. arXiv: 1909.12200v2 [cs.RO], 2019.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NIPS*, pp. 4299–4307, 2017.

[7] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *ICML*, 2016.

[8] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *ICLR*, 2018.

[9] Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, Rene Traore, Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu. Stable Baselines. `https://github.com/hill-a/stable-baselines`, 2018.

[10] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in Atari. In *NeurIPS*, pp. 8011–8023, 2018.

[11] Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*. SIAM, 1995.

[12] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *ICML*, 2000.

[13] Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: theory and application to reward shaping. In *NIPS*, 1999.

[14] OpenAI. OpenAI Five. `https://blog.openai.com/openai-five/`, 2018.

[15] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving Rubik's Cube with a robot hand. arXiv: 1910.07113v1 [cs.LG], 2019.

[16] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, 2007.

[17] Dorsa Sadigh, Anca D. Dragan, S. Shankar Sastry, and Sanjit A. Seshia. Active preference-based learning of reward functions. In *RSS*, July 2017.

[18] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv:1707.06347v2 [cs.LG], 2017.

[19] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[20] Satinder P. Singh and Richard C. Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, September 1994.

[21] Mel Vecerik, Oleg Sushkov, David Barker, Thomas Rothörl, Todd Hester, and Jon Scholz. A practical approach to insertion with variable socket position using deep reinforcement learning. In *ICRA*, 2019.

[22] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[23] Steven Wang, Adam Gleave, and Sam Toyer. imitation: implementations of inverse reinforcement learning and imitation learning algorithms. `https://github.com/humancompatibleai/imitation`, 2020.

[24] Aaron Wilson, Alan Fern, and Prasad Tadepalli. A Bayesian approach for policy learning from trajectory preference queries. In *NIPS*, 2012.

[25] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.

[26] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. arXiv: 1909.08593v2 [cs.CL], 2019.

# A SUPPLEMENTARY MATERIAL

## A.1 APPROXIMATION PROCEDURES

### A.1.1 SAMPLE-BASED APPROXIMATION FOR EPIC DISTANCE

We approximate EPIC distance (definition 4.6) by estimating Pearson distance on a set of samples, canonicalizing the reward on-demand. Specifically, we sample a batch $B_V$ of $N_V$ samples from the coverage distribution $\mathcal{D}$, and a batch $B_M$ of $N_M$ samples from the joint state and action distributions $\mathcal{D}_{\mathcal{S}} \times \mathcal{D}_{\mathcal{A}}$. For each $(s, a, s') \in B_V$, we approximate the canonically shaped rewards (definition 4.1) by taking the mean over $B_M$:

$$C_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R)(s,a,s') = R(s,a,s') + \mathbb{E}\left[\gamma R(s',A,S') - R(s,A,S') - \gamma R(S,A,S')\right] \tag{6}$$

$$\approx R(s,a,s') + \frac{\gamma}{N_M} \sum_{(x,u)\in B_M} R(s',u,x) \tag{7}$$

$$- \frac{1}{N_M} \sum_{(x,u)\in B_M} R(s,u,x) - c. \tag{8}$$

We drop the constant $c$ from the approximation since it does not affect the Pearson distance; it can also be estimated in $O(N_M^2)$ time by $c = \frac{\gamma}{N_M^2} \sum_{(x,\cdot)\in B_M} \sum_{(x',u)\in B_M} R(x,u,x')$. Finally, we compute the Pearson distance between the approximate canonically shaped rewards on the batch of samples $B_V$, yielding an $O(N_V N_M)$ time algorithm.

### A.1.2 OPTIMIZATION-BASED APPROXIMATION FOR NPEC DISTANCE

$D_{\text{NPEC}}(R_A, R_B)$ (section 5.2) is defined as the infimum of $L^p$ distance over an infinite set of equivalent reward functions $R \equiv R_A$. We approximate this using gradient descent on the reward model

$$R_{\nu,c,w}(s,a,s') = \exp(\nu) R_A(s,a,s') + c + \gamma \Phi_w(s') - \Phi_w(s), \tag{9}$$

where $\nu, c \in \mathbb{R}$ are scalar weights and $w$ is a vector of weights parameterizing a deep neural network $\Phi_w$. The constant $c \in \mathbb{R}$ is unnecessary if $\Phi_w$ has a bias term, but its inclusion simplifies the optimization problem.

We optimize $\nu, c, w$ to minimize the mean of the cost

$$J(\nu,c,w)(s,a,s') = \|R_{\nu,c,w}(s,a,s'), R_B(s,a,s')\|^p \tag{10}$$

on samples $(s, a, s')$ from a coverage distribution $\mathcal{D}$. Note

$$\mathbb{E}_{(S,A,S')\sim\mathcal{D}}\left[J(\nu,c,w)(S,A,S')\right]^{1/p} = D_{L^p,\mathcal{D}}(R_{\nu,c,w}, R_B) \tag{11}$$

upper bounds the true NPEC distance since $R_{\nu,c,w} \equiv R_A$.

We found empirically that $\nu$ and $c$ need to be initialized close to their optimal values for gradient descent to reliably converge. To resolve this problem, we initialize the affine parameters to $\nu \leftarrow \log \lambda$ and $c$ found by:

$$\underset{\lambda\geq 0, c\in\mathbb{R}}{\arg\min} \mathop{\mathbb{E}}_{s,a,s'\sim\mathcal{D}} \left(\lambda R_A(s,a,s') + c - R_B(s,a,s')\right)^2. \tag{12}$$

We use the active set method of Lawson & Hanson [11] to solve this constrained least-squares problem. These initial affine parameters minimize the $L^p$ distance $D_{L^p,\mathcal{D}}(R_{\nu,c,0}(s,a,s'), R_B(s,a,s'))$ when $p = 2$ with the potential fixed at $\Phi_0(s) = 0$.

### A.1.3 CONFIDENCE INTERVALS

We report confidence intervals to help measure the degree of error introduced by the approximations. Since approximate distances may not be normally distributed, we use bootstrapping to produce a distribution-free confidence interval. For EPIC and NPEC, we compute independent approximate distances over different seeds, and then compute a bootstrapped confidence interval on the distances for each seed. We use 30 seeds for EPIC but only 3 seeds for NPEC due to its greater computational requirements. In ERC, computing the distance is very fast, so we instead apply bootstrapping to the collected *episodes*, computing the ERC distance for each bootstrapped episode sample.

Table A.1: Summary of hyperparameters and distributions used in experiments. The uniform random coverage distribution $\mathcal{D}_{\text{unif}}$ samples states and actions uniformly at random, and samples the next state from the transition dynamics. Random policy $\pi_{\text{uni}}$ takes uniform random actions. The synthetic expert policy $\pi^*$ was trained with PPO on the ground-truth reward. **Mix**ture samples actions from either $\pi_{\text{uni}}$ or $\pi^*$, switching between them at each time step with probability $0.05$. Warmstart Size is the size of the dataset used to compute initialization parameters described in section A.1.2.

| **Parameter** | **Value** | **In experiment** |
|---|---|---|
| Coverage Distribution $\mathcal{D}$ | Random transitions $\mathcal{D}_{\text{unif}}$<br>Rollouts from $\pi_{\text{uni}}$<br>$\pi_{\text{uni}}$, $\pi^*$ and **Mix**ture | `GridWorld`<br>`PointMass`, `HalfCheetah`, `Hopper`<br>`PointMaze` |
| Bootstrap Samples | $10\,000$ | All |
| Discount $\gamma$ | $0.99$ | All |
| **EPIC** | | |
| State Distribution $\mathcal{D}_{\mathcal{S}}$ | Marginalized from $\mathcal{D}$ | All |
| Action Distribution $\mathcal{D}_{\mathcal{A}}$ | Marginalized from $\mathcal{D}$ | All |
| Seeds | 30 | All |
| Samples $N_V$ | $32\,768$ | All |
| Mean Samples $N_M$ | $32\,768$ | All |
| **NPEC** | | |
| Seeds | 3 | All |
| Total Time Steps | $1 \times 10^6$ | All |
| Optimizer | Adam | All |
| Learning Rate | $1 \times 10^{-2}$ | All |
| Batch Size | 4096 | All |
| Warmstart Size | $16\,386$ | All |
| Loss $\ell$ | $\ell(x, y) = (x - y)^2$ | All |
| **ERC** | | |
| Episodes | $131\,072$ | All |

## A.2 Experiments

### A.2.1 Hyperparameters for Approximate Distances

Table A.1 summarizes the hyperparameters and distributions used to compute the distances between reward functions. Most parameters are the same across all environments. We use a coverage distribution of uniform random transitions $\mathcal{D}_{\text{unif}}$ in the simple `GridWorld` environment with known determinstic dynamics. In other environments, the coverage distribution is sampled from rollouts of a policy. We use a random policy $\pi_{\text{uni}}$ for `PointMass`, `HalfCheetah` and `Hopper` in the hand-designed reward experiments (section 6.1). In `PointMaze`, we compare three coverage distributions (section 6.2) induced by rollouts of $\pi_{\text{uni}}$, an expert policy $\pi^*$ and a **Mix**ture of the two policies, sampling actions from either $\pi_{\text{uni}}$ or $\pi^*$ and switching between them with probability $0.05$ per time step.

### A.2.2 Training Learned Reward Models

For the experiments on learned reward functions (sections 6.3 and 6.2), we trained reward models using adversarial inverse reinforcement learning (AIRL; 8), preference comparison [6] and by regression onto the ground-truth reward [*target* method from 6, section 3.3]. For AIRL, we use an existing open-source implementation [23]. We developed new implementations for preference comparison and regression, available at — double blind: supplementary material —. We also use the RL algorithm proximal policy optimization (PPO; 18) on the ground-truth reward to train expert policies to provide demonstrations for AIRL, and on learned reward models to evaluate their performance.

Table A.2: Hyperparameters for proximal policy optimisation (PPO) [18]. We used the implementation and default hyperparameters from Hill et al. [9]. PPO was used to train expert policies on ground-truth reward and to optimize learned reward functions for evaluation.

| **Parameter** | **Value** | **In environment** |
|---|---|---|
| Total Time Steps | $1 \times 10^6$ | All |
| Batch Size | $16\,384$ | `PointMaze` |
| | $1024$ | All others |
| Discount $\gamma$ | $0.99$ | All |
| Entropy Coefficient | $0.01$ | All |
| Learning Rate | $2.5 \times 10^{-4}$ | All |
| Value Function Coefficient | $0.5$ | All |
| Gradient Clipping Threshold | $0.5$ | All |
| Ratio Clipping Thrsehold | $0.2$ | All |
| Lambda (GAE) | $0.95$ | All |
| Minibatches | $4$ | All |
| Optimization Epochs | $4$ | All |
| Parallel Environments | $8$ | All |

Table A.3: Hyperparameters for adversarial inverse reinforcement learning (AIRL) used in Wang et al. [23].

| **Parameter** | **Value** |
|---|---|
| RL Algorithm | PPO [18] |
| Total Time Steps | $102\,400$ |
| Discount $\gamma$ | $0.99$ |
| Demonstration Time Steps | $100\,000$ |
| Generator Batch Size | $2048$ |
| Discriminator Batch Size | $50$ |
| Entropy Weight | $1.0$ |
| Reward Function Architecture | MLP, two 32-unit hidden layers |
| Potential Function Architecture | MLP, two 32-unit hidden layers |

Table A.4: Hyperparameters for preference comparison used in our implementation of Christiano et al. [6].

| **Parameter** | **Value** | **Range Tested** |
|---|---|---|
| Total Time Steps | $5 \times 10^6$ | $[1, 10 \times 10^6]$ |
| Batch Size | $10\,000$ | $[500, 250\,000]$ |
| Trajectory Length | $5$ | $[1, 100]$ |
| Learning Rate | $1 \times 10^{-2}$ | $[1 \times 10^{-4}, 1 \times 10^{-1}]$ |
| Discount $\gamma$ | $0.99$ | |
| Reward Function Architecture | MLP, two 32-unit hidden layers | |
| Output L2 Regularization Weight | $1 \times 10^{-4}$ | |

Table A.5: Hyperparameters for regression used in our implementation of Christiano et al. [6, *target* method from section 3.3].

| **Parameter** | **Value** | **Range Tested** |
|---|---|---|
| Total Time Steps | $10 \times 10^6$ | $[1, 20 \times 10^6]$ |
| Batch Size | $4096$ | $[256, 16\,384]$ |
| Learning Rate | $2 \times 10^{-2}$ | $[1 \times 10^{-3}, 1 \times 10^{-1}]$ |
| Discount $\gamma$ | $0.99$ | |
| Reward Function Architecture | MLP, two 32-unit hidden layers | |

For PPO and AIRL we used the default hyperparameters in tables A.2 and A.3, finding them adequate and so performing no further tuning. For preference comparison we performed a sweep over batch size, trajectory length and learning rate to decide on the hyperparameters in table A.4. Total time steps was selected once diminishing returns were observed in loss curves. The exact value of the regularization weight was found to be unimportant, largely controlling the scale of the output at convergence. Finally, for regression we performed a sweep over batch size, learning rate and total time steps to decide on the hyperparameters in table A.5. We found batch size and learning rate to be relatively unimportant with many combinations performing well, but regression was found to converge slowly but steadily requiring a relatively large $10 \times 10^6$ time steps for good performance in our environments.

All algorithms are trained on synthetic data generated from the ground-truth reward function. AIRL is provided with a large demonstration dataset of $100\,000$ time steps from an expert policy trained on the ground-truth reward, similar in size to the total number of time steps AIRL is trained for (see table A.3). In preference comparison and regression, each batch is sampled afresh from the coverage distribution specified in table A.1 and labeled according to the ground-truth reward.

#### A.2.3 COMPUTING INFRASTRUCTURE

Experiments were conducted on a workstation (Intel i9-7920X CPU with 64 GB of RAM), and a small number of `r5.24xlarge` AWS VM instances, with 48 CPU cores on an Intel Skylake processor and 768 GB of RAM. It takes less than a week of compute on a single `r5.24xlarge` instance to run all the experiments described in this paper.

#### A.2.4 COMPARING HAND-DESIGNED REWARD FUNCTIONS

We compute distances between hand-designed reward functions in four environments: `GridWorld`, `PointMass`, `HalfCheetah` and `Hopper`. The reward functions for `GridWorld` are described in Figure A.1, and the distances are reported in Figure A.2. We report the approximate distances and confidence intervals between reward functions in the other environments in Figures A.3, A.4 and A.5.

We find the (approximate) EPIC distance closely matches our intuitions for similarity between the reward functions. NPEC often produces similar results to EPIC, but unfortunately is dogged by optimization error. This is particularly notable in higher-dimensional environments like `HalfCheetah` and `Hopper`, where the NPEC distance often exceeds the theoretical upper bound of $1.0$ and the confidence interval width is frequently larger than $0.2$.

By contrast, ERC distance generally has a tight confidence interval, but systematically fails in the presence of shaping. For example, it confidently assigns large distances between *equivalent* reward pairs in `PointMass` such as `S`🐎-`D`🐎. However, ERC produces reasonable results in `HalfCheetah` and `Hopper` where rewards are all similarly shaped. In fact, ERC picks up on a detail in `Hopper` that EPIC misses: whereas EPIC assigns a distance of around $0.71$ between all rewards of different types (running vs backflipping), ERC assigns lower distances when the rewards are in the same direction (forward or backward). Given this, ERC may be attractive in some circumstances, especially given the ease of implementation. However, we would caution against using it in isolation due to the likelihood of misleading results in the presence of shaping.

#### A.2.5 COMPARING LEARNED REWARD FUNCTIONS

Previously, we reported the mean approximate distance from a ground-truth reward of four learned reward models in `PointMaze` (Table 2). Since these distances are approximate, we report 95% lower and upper bounds computed via bootstrapping in Table A.7. We also include the relative difference of the upper and lower bounds from the mean, finding the relative difference to be fairly consistent across reward models for a given algorithm and coverage distribution pair. The relative difference is less than 1% for all EPIC and ERC distances. However, NPEC confidence intervals can be as wide as 50%: this is due to the method's high variance, and the small number of seeds we were able to run because of the method's computational expense.

| Distance Metric | Wall-Clock Time | Environment Time Steps | Reward Queries | # of Seeds | 95% CI Width Max | 95% CI Width Mean |
|---|---|---|---|---|---|---|
| EPIC Quick | 17 s | 8192 | $1.67 \times 10^7$ | 3 | 0.023 04 | 0.008 60 |
| EPIC | 6738 s | 65 536 | $1.07 \times 10^9$ | 30 | 0.005 58 | 0.002 34 |
| NPEC | 29 769 s | $7.50 \times 10^7$ | $7.50 \times 10^7$ | 3 | 0.315 91 | 0.066 20 |
| ERC | 1376 s | $6.55 \times 10^6$ | $6.55 \times 10^6$ | — | 0.015 81 | 0.005 33 |
| RL (PPO) | 14 745 s | $7.50 \times 10^7$ | $7.50 \times 10^7$ | 3 | — | — |

Table A.6: Time and resources taken by different metrics to perform 25 distance comparisons on `PointMass`, and the confidence interval widths obtained (smaller is better). Methods ***EPIC***, ***NPEC*** and ***ERC*** correspond to Figures 2(a), (b) and (c) respectively. ***EPIC Quick*** is an abbreviated version with fewer samples. ***RL (PPO)*** is estimated from the time taken using PPO to train a single policy (16m:23s) until convergence ($10^6$ time steps). EPIC samples $N_M + N_V$ time steps from the environment and performs $N_M N_V$ reward queries. In ***EPIC Quick***, $N_M = N_V = 4096$; in ***EPIC***, $N_M = N_V = 302768$. Other methods query the reward once per environment time step.

### A.2.6 Runtime of Distance Metrics

We report the empirical runtime for EPIC and baselines in Table A.6, performing 25 pairwise comparisons across 5 reward functions in `PointMass`. These comparisons were run on an unloaded machine running Ubuntu 20.04 (kernel 5.4.0-52) with an Intel i9-7920X CPU and 64 GB of RAM. We report sequential runtimes: runtimes for all methods could be decreased further by parallelizing across seeds. The algorithms were configured to use 8 parallel environments for sampling. Inference and training took place on CPU. All methods used the same TensorFlow configuration, parallelizing operations across threads both within and between operations. We found GPUs offered no performance benefit in this setting, and in some cases even increased runtime. This is due to the fixed cost of CPU-GPU communication, and the relatively small size of the observations.

We find that in just 17 seconds EPIC can provide results with a 95% confidence interval $< 0.023$, an order of magnitude tighter than NPEC running for over 8 hours. Training policies for all learned rewards in this environment using PPO is comparatively slow, taking over 4 hours even with only 3 seeds. While ERC is relatively fast, it takes a large number of samples to achieve tight confidence intervals. Moreover, since `PointMass` has stochastic initial states, ERC can take on arbitrary values under shaping, as discussed in sections 5.1 and 6.3.

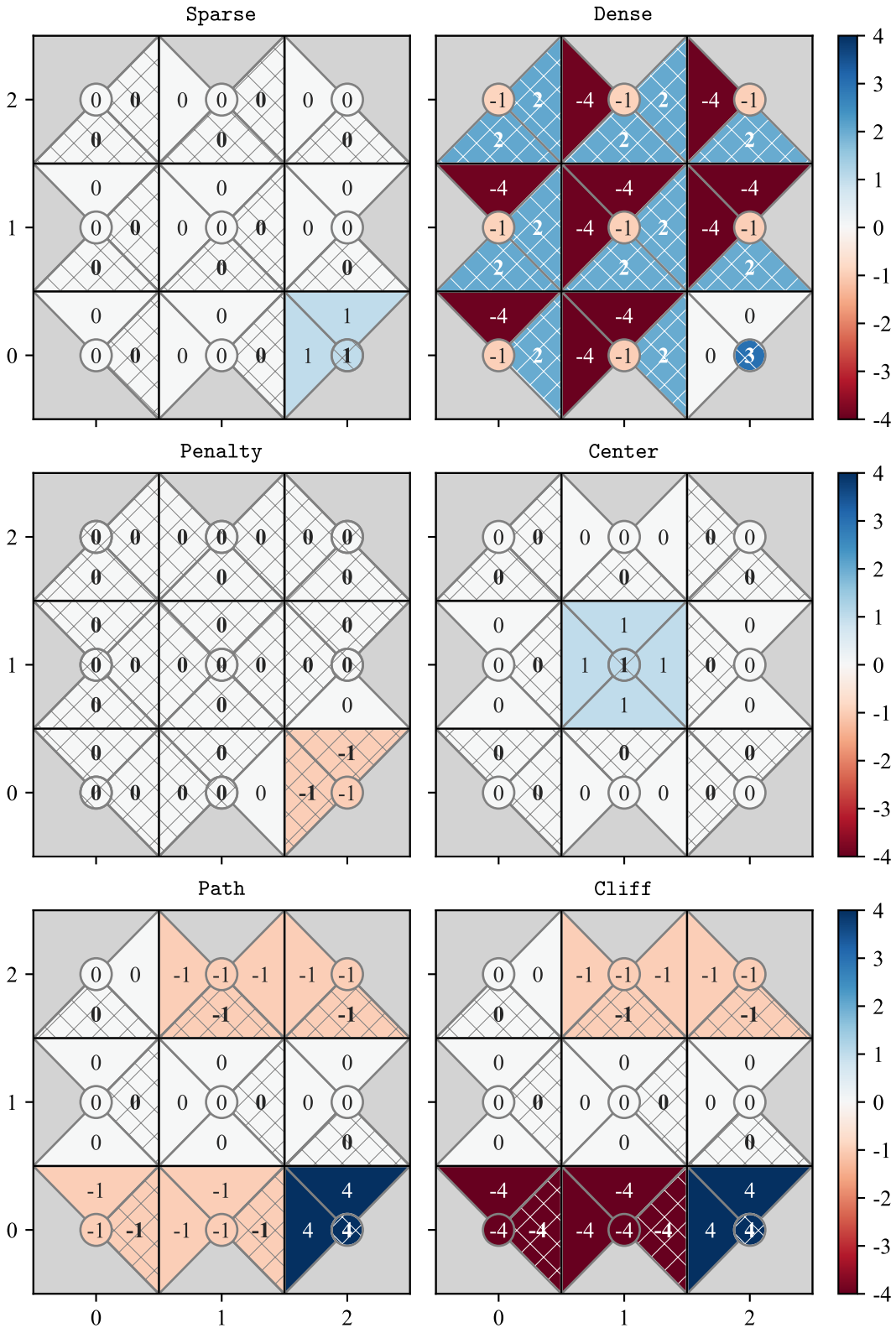
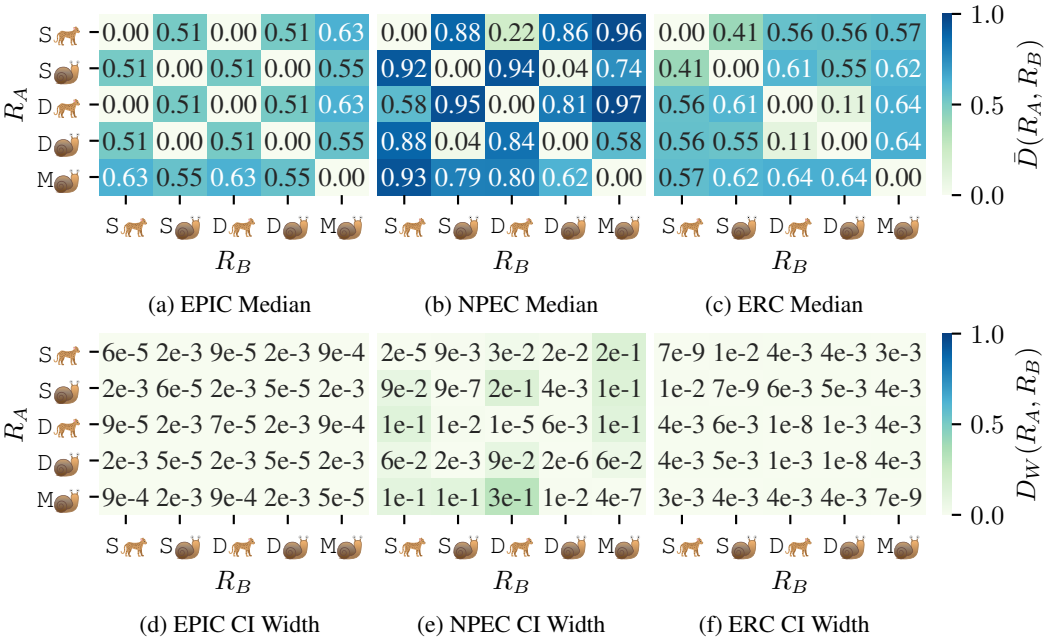



Figure A.1: Heatmaps of reward functions $R(s, a, s')$ for a $3 \times 3$ deterministic gridworld. $R(s, \text{stay}, s)$ is given by the central circle in cell $s$. $R(s, a, s')$ is given by the triangular wedge in cell $s$ adjacent to cell $s'$ in direction $a$. Optimal action(s) (for infinite horizon, discount $\gamma = 0.99$) have bold labels against a hatched background. See figure A.2 for the distance between all reward pairs.

(a) EPIC

(b) NPEC

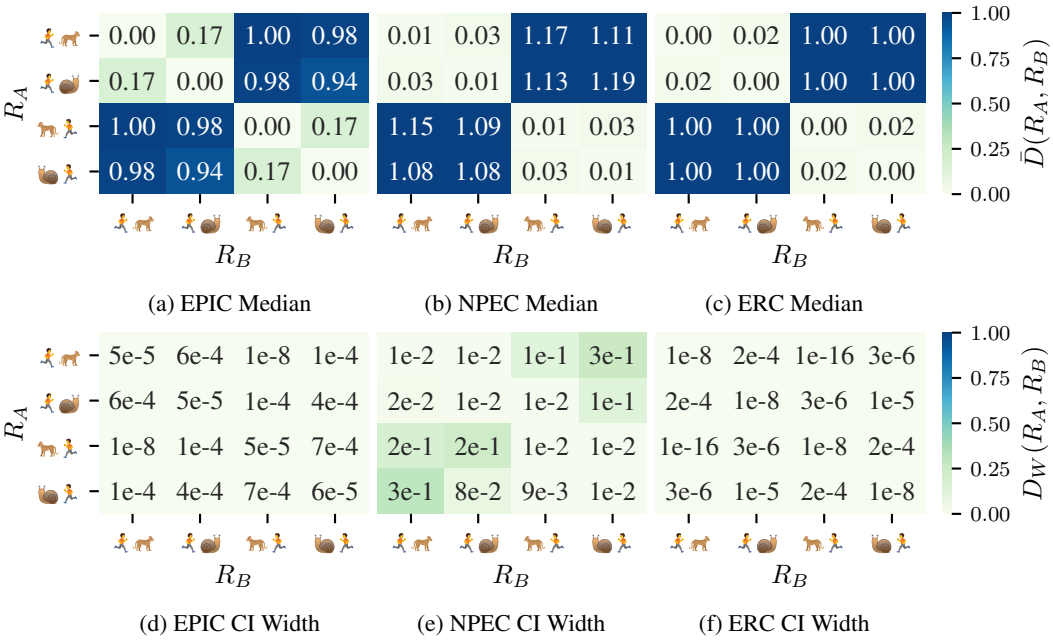
Figure A.2: Distances between hand-designed reward functions for the $3 \times 3$ deterministic `Gridworld` environment. See figure A.1 for definitions of each reward. Distances are computed using tabular algorithms. We do not report confidence intervals since these algorithms are deterministic and exact up to floating point error.




(a) EPIC Median (b) NPEC Median (c) ERC Median

(d) EPIC CI Width (e) NPEC CI Width (f) ERC CI Width

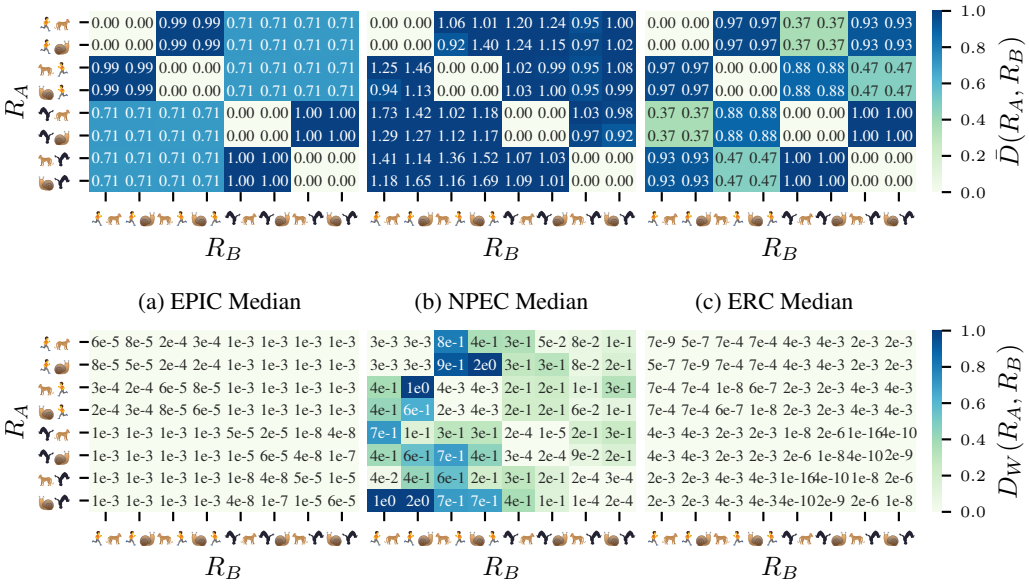
Figure A.3: Approximate distances between hand-designed reward functions in `PointMass`. The coverage distribution $\mathcal{D}$ is sampled from rollouts of a policy $\pi_{\text{uni}}$ taking actions uniformly at random. **Key**: quadratic control penalty, no control penalty. `S` is $\text{Sparse}(x) = \mathbb{1}[|x| < 0.05]$, `D` is shaped $\text{Dense}(x, x') = \text{Sparse}(x) + |x'| - |x|$, while `M` is $\text{Magnitude}(x) = -|x|$. **Confidence Interval (CI)**: 95% CI computed by bootstraping over 10 000 samples.




(a) EPIC Median (b) NPEC Median (c) ERC Median

(d) EPIC CI Width (e) NPEC CI Width (f) ERC CI Width

Figure A.4: Approximate distances between hand-designed reward functions in `HalfCheetah`. The coverage distribution $\mathcal{D}$ is sampled from rollouts of a policy $\pi_{\text{uni}}$ taking actions uniformly at random. **Key**: is a reward proportional to the change in center of mass and moving *forward* is rewarded when to the right, and moving *backward* is rewarded when to the left. quadratic control penalty, no control penalty. **Confidence Interval (CI)**: 95% CI computed by bootstraping over 10 000 samples.




(a) EPIC Median (b) NPEC Median (c) ERC Median

(d) EPIC CI Width (e) NPEC CI Width (f) ERC CI Width

Figure A.5: Approximate distances between hand-designed reward functions in `Hopper`. The coverage distribution $\mathcal{D}$ is sampled from rollouts of a policy $\pi_{\text{uni}}$ taking actions uniformly at random. **Key**: 🏃 is a reward proportional to the change in center of mass and 🦘 is the backflip reward defined in Amodei et al. [2, footnote]. Moving *forward* is rewarded when 🏃 or 🦘 is to the right, and moving *backward* is rewarded when 🏃 or 🦘 is to the left. 🐌 quadratic control penalty, 🐆 no control penalty. **Confidence Interval (CI)**: 95% CI computed by bootstraping over 10 000 samples.

Table A.7: Approximate distances of reward functions from the ground-truth (GT). We report the 95% bootstrapped lower and upper bounds, the mean, and a 95% bound on the relative error from the mean. Distances ($1000\times$ scale) use coverage distribution $\mathcal{D}$ from rollouts in the *train* environment of: a uniform random policy $\boldsymbol{\pi}_{\mathbf{uni}}$, an expert $\boldsymbol{\pi}^*$ and a **Mix**ture of these policies. $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{A}}$ are computed by marginalizing $\mathcal{D}$.

(a) 95% lower bound $D^{\mathrm{LOW}}$ of approximate distance.

| **Reward** | $\mathbf{1000 \times D^{LOW}_{EPIC}}$ | | | $\mathbf{1000 \times D^{LOW}_{NPEC}}$ | | | $\mathbf{1000 \times D^{LOW}_{ERC}}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **Function** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** |
| Regress | 41.9 | 36.4 | 25.8 | 0.46 | 13.1 | 0.13 | 4.77 | 40.4 | 1.39 |
| Pref | 50.4 | 54.3 | 32.8 | 2.95 | 210 | 1.99 | 15 | 178 | 8.09 |
| AIRL SO | 484 | 599 | 390 | 637 | 1044 | 382 | 445 | 378 | 231 |
| AIRL SA | 546 | 614 | 387 | 806 | 856 | 425 | 505 | 463 | 206 |
| Mirage | 8.96 | 0.01 | 401 | 1.42 | 1.53 | 901 | 24.1 | <0.01 | 161 |

(b) Mean approximate distance $\overline{D}$. Results are the same as Table 2.

| **Reward** | $\mathbf{1000 \times \overline{D}_{EPIC}}$ | | | $\mathbf{1000 \times \overline{D}_{NPEC}}$ | | | $\mathbf{1000 \times \overline{D}_{ERC}}$ | | | **Return** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Function** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | **Train** | **Test** |
| GT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $-5.99$ | $-6.05$ |
| Regress | 42.2 | 36.4 | 25.9 | 0.47 | 13.5 | 0.15 | 4.80 | 40.7 | 1.40 | $-6.99$ | $-6.51$ |
| Pref | 50.8 | 54.4 | 32.9 | 3.08 | 217 | 2.11 | 15.1 | 179 | 8.12 | $-6.62$ | $-7.02$ |
| AIRL SO | 488 | 599 | 392 | 682 | 1089 | 469 | 447 | 380 | 232 | $-28.5$ | $-11.7$ |
| AIRL SA | 549 | 614 | 388 | 871 | 1041 | 443 | 507 | 465 | 208 | $-6.91$ | $-28.5$ |
| Mirage | 37.3 | 0.03 | 417 | 8.80 | 3.78 | 1580 | 32 | <0.01 | 166 | $-34.2$ | $-46.8$ |

(c) 95% upper bound $D^{\mathrm{UP}}$ of approximate distance.

| **Reward** | $\mathbf{1000 \times D^{UP}_{EPIC}}$ | | | $\mathbf{1000 \times D^{UP}_{NPEC}}$ | | | $\mathbf{1000 \times D^{UP}_{ERC}}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **Function** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** |
| Regress | 42.5 | 36.5 | 26 | 0.49 | 14.2 | 0.16 | 4.82 | 40.9 | 1.41 |
| Pref | 51.1 | 54.4 | 33 | 3.15 | 224 | 2.21 | 15.2 | 180 | 8.16 |
| AIRL SO | 492 | 600 | 395 | 728 | 1125 | 549 | 449 | 382 | 234 |
| AIRL SA | 552 | 614 | 390 | 993 | 1260 | 476 | 509 | 467 | 209 |
| Mirage | 76.9 | 0.06 | 434 | 21.8 | 5.84 | 2510 | 40.7 | <0.01 | 171 |

(d) Relative 95% confidence interval $D^{\mathrm{REL}} = \max\left(\frac{\text{Upper}}{\text{Mean}} - 1, 1 - \frac{\text{Lower}}{\text{Mean}}\right)$ in percent. The population mean is contained within $\pm D^{\mathrm{REL}}\%$ of the sample mean in Table A.7b with 95% probability.

| **Reward** | $\mathbf{D^{REL}_{EPIC}\%}$ | | | $\mathbf{D^{REL}_{NPEC}\%}$ | | | $\mathbf{D^{REL}_{ERC}\%}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **Function** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** | $\pi_{\mathrm{uni}}$ | $\pi^*$ | **Mix** |
| Regress | 0.72 | 0.09 | 0.31 | 5.14 | 5.16 | 10.2 | 0.58 | 0.57 | 0.62 |
| Pref | 0.75 | 0.12 | 0.39 | 4.23 | 3.32 | 5.95 | 0.50 | 0.56 | 0.47 |
| AIRL SO | 0.80 | 0.09 | 0.66 | 6.76 | 4.17 | 18.4 | 0.45 | 0.51 | 0.64 |
| AIRL SA | 0.55 | 0.02 | 0.46 | 14 | 21.1 | 7.28 | 0.39 | 0.50 | 0.54 |
| Mirage | 106 | 66.7 | 4.08 | 148 | 59.5 | 58.8 | 27.3 | 0.60 | 2.94 |

Table A.8: Approximate distances of reward functions from the ground-truth (GT) under pathological coverage distributions. We report the 95% bootstrapped lower and upper bounds, the mean, and a 95% bound on the relative error from the mean. Distances ($1000\times$ scale) use four different coverage distributions $\mathcal{D}$. $\boldsymbol{\sigma}$ independently samples states, actions and next states from the marginal distributions of rollouts from the uniform random policy $\pi_{\text{uni}}$ in the *train* environment. **Ind** independently samples the components of states and next states from $\mathcal{N}(0,1)$, and actions from $U[-1,1]$. **Jail** consists of rollouts of $\pi_{\text{uni}}$ restricted to a small $0.09 \times 0.09$ "jail" square that excludes the goal state $0.5$ distance away. $\boldsymbol{\pi}_{\mathbf{bad}}$ are rollouts in *train* of a policy that goes to the corner opposite the goal state. $\boldsymbol{\sigma}$ and **Ind** are not supported by ERC since they do not produce complete episodes.

(a) 95% lower bound $D^{\text{LOW}}$ of approximate distance.

| **Reward** | $\mathbf{1000 \times D^{LOW}_{EPIC}}$ | | | | $\mathbf{1000 \times D^{LOW}_{NPEC}}$ | | | | $\mathbf{1000 \times D^{LOW}_{ERC}}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Function** | $\boldsymbol{\sigma}$ | **Ind** | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ | $\boldsymbol{\sigma}$ | **Ind** | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ |
| Regress | 135 | 375 | 441 | 231 | 77.9 | 553 | 2023 | 204 | 341 | 285 |
| Pref | 144 | 409 | 520 | 379 | 96.3 | 595 | 806 | 135 | 565 | 297 |
| AIRL SO | 484 | 439 | 539 | 638 | 637 | 626 | 1312 | 973 | 486 | 622 |
| AIRL SA | 555 | 450 | 858 | 774 | 657 | 645 | 1502 | 853 | 859 | 601 |
| Mirage | 34.5 | 1.13 | 0.02 | <0.01 | 0.27 | 0.12 | 12 | 23.2 | <0.01 | 0.02 |

(b) Mean approximate distance $\overline{D}$. Results are the same as Table 2.

| **Reward** | $\mathbf{1000 \times \overline{D}_{EPIC}}$ | | | | $\mathbf{1000 \times \overline{D}_{NPEC}}$ | | | | $\mathbf{1000 \times \overline{D}_{ERC}}$ | | **Return** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Function** | $\boldsymbol{\sigma}$ | **Ind** | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ | $\boldsymbol{\sigma}$ | **Ind** | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ | **Train** | **Test** |
| GT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $-5.99$ | $-6.05$ |
| Regress | 136 | 376 | 441 | 231 | 90.2 | 562 | 2161 | 228 | 343 | 286 | $-6.99$ | $-6.51$ |
| Pref | 146 | 410 | 520 | 379 | 101 | 606 | 878 | 143 | 567 | 299 | $-6.62$ | $-7.02$ |
| AIRL SO | 487 | 440 | 540 | 639 | 705 | 632 | 1406 | 1165 | 488 | 624 | $-28.5$ | $-11.7$ |
| AIRL SA | 558 | 451 | 859 | 774 | 787 | 650 | 1583 | 922 | 860 | 603 | $-6.91$ | $-28.5$ |
| Mirage | 68.7 | 2.54 | 0.05 | 0.04 | 104 | 0.25 | 18.6 | 30.4 | <0.01 | 0.02 | $-34.2$ | $-46.8$ |

(c) 95% upper bound $D^{\text{UP}}$ of approximate distance.

| **Reward** | $\mathbf{1000 \times D^{UP}_{EPIC}}$ | | | | $\mathbf{1000 \times D^{UP}_{NPEC}}$ | | | | $\mathbf{1000 \times D^{UP}_{ERC}}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Function** | $\boldsymbol{\sigma}$ | **Ind** | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ | $\boldsymbol{\sigma}$ | **Ind** | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ |
| Regress | 138 | 377 | 442 | 231 | 105 | 570 | 2261 | 248 | 344 | 288 |
| Pref | 147 | 411 | 521 | 380 | 105 | 612 | 991 | 149 | 569 | 300 |
| AIRL SO | 491 | 440 | 541 | 639 | 790 | 640 | 1467 | 1379 | 490 | 627 |
| AIRL SA | 561 | 452 | 859 | 774 | 980 | 658 | 1675 | 960 | 862 | 605 |
| Mirage | 111 | 4.45 | 0.07 | 0.07 | 309 | 0.45 | 24.9 | 35.9 | <0.01 | 0.02 |

(d) Relative 95% confidence interval $D^{\text{REL}} = \max\left(\frac{\text{Upper}}{\text{Mean}} - 1, 1 - \frac{\text{Lower}}{\text{Mean}}\right)$ in percent. The population mean is contained within $\pm D^{\text{REL}}\%$ of the sample mean in Table A.8b with 95% probability.

| **Reward** | $\mathbf{D^{REL}_{EPIC}\%}$ | | | | $\mathbf{D^{REL}_{NPEC}\%}$ | | | | $\mathbf{D^{REL}_{ERC}\%}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Function** | $\boldsymbol{\sigma}$ | **Ind** | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ | $\boldsymbol{\sigma}$ | **Ind** | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ | **Jail** | $\boldsymbol{\pi}_{\mathbf{bad}}$ |
| Regress | 1.14 | 0.25 | 0.13 | 0.06 | 15.9 | 1.73 | 6.38 | 10.8 | 0.48 | 0.50 |
| Pref | 0.96 | 0.21 | 0.14 | 0.10 | 4.52 | 1.78 | 12.8 | 6.02 | 0.37 | 0.50 |
| AIRL SO | 0.75 | 0.15 | 0.14 | 0.04 | 12 | 1.31 | 6.65 | 18.4 | 0.41 | 0.34 |
| AIRL SA | 0.56 | 0.18 | 0.05 | 0.02 | 24.5 | 1.20 | 5.78 | 7.52 | 0.14 | 0.35 |
| Mirage | 61.4 | 75 | 62.5 | 106 | 198 | 80.8 | 35.3 | 23.9 | 0.54 | 0.54 |

### A.3 PROOFS

#### A.3.1 BACKGROUND

**Proposition 3.5.** *The binary relation $\equiv$ is an equivalence relation. Let $R_A, R_B, R_C : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ be bounded reward functions. Then $\equiv$ is reflexive, $R_A \equiv R_A$; symmetric, $R_A \equiv R_B$ implies $R_B \equiv R_A$; and transitive, $(R_A \equiv R_B) \wedge (R_B \equiv R_C)$ implies $R_A \equiv R_C$.*

*Proof.* $R_A \equiv R_A$ since choosing $\lambda = 1 > 0$ and $\Phi(s) = 0$, a bounded potential function, we have $R_A(s, a, s') = \lambda R_A(s, a, s') + \gamma\Phi(s') - \Phi(s)$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$.

Suppose $R_A \equiv R_B$. Then there exists some $\lambda > 0$ and a bounded potential function $\Phi : \mathcal{S} \to \mathbb{R}$ such that $R_B(s, a, s') = \lambda R_A(s, a, s') + \gamma\Phi(s') - \Phi(s)$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Rearranging:

$$R_A(s, a, s') = \frac{1}{\lambda} R_B(s, a, s') + \gamma\left(\frac{-1}{\lambda}\Phi(s')\right) - \left(\frac{-1}{\lambda}\Phi(s)\right). \tag{13}$$

Since $\frac{1}{\lambda} > 0$ and $\Phi'(s) = \frac{-1}{\lambda}\Phi(s)$ is a bounded potential function, it follows that $R_B \equiv R_A$.

Finally, suppose $R_A \equiv R_B$ and $R_B \equiv R_C$. Then there exists some $\lambda_1, \lambda_2 > 0$ and bounded potential functions $\Phi_1, \Phi_2 : \mathcal{S} \to \mathbb{R}$ such that for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$:

$$R_B(s, a, s') = \lambda_1 R_A(s, a, s') + \gamma\Phi_1(s') - \Phi_1(s), \tag{14}$$
$$R_C(s, a, s') = \lambda_2 R_B(s, a, s') + \gamma\Phi_2(s') - \Phi_2(s). \tag{15}$$

Substituting the expression for $R_B$ into the expression for $R_C$:

$$\begin{aligned} R_C(s, a, s') &= \lambda_2\left(\lambda_1 R_A(s, a, s') + \gamma\Phi_1(s') - \Phi_1(s)\right) + \gamma\Phi_2(s') - \Phi_2(s) && (16)\\ &= \lambda_1\lambda_2 R_A(s, a, s') + \gamma\left(\lambda_2\Phi_1(s') + \Phi_2(s')\right) - \left(\lambda_2\Phi_1(s) + \Phi_2(s)\right) && (17)\\ &= \lambda R_A(s, a, s') + \gamma\Phi(s') - \Phi(s), && (18)\end{aligned}$$

where $\lambda = \lambda_1\lambda_2 > 0$ and $\Phi(s) = \lambda_2\Phi_1(s) + \Phi_2(s)$ is bounded. Thus $R_A \equiv R_C$. □

#### A.3.2 EQUIVALENT-POLICY INVARIANT COMPARISON (EPIC) PSEUDOMETRIC

**Proposition 4.2** (The Canonically Shaped Reward is Invariant to Shaping)**.** *Let $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be a reward function and $\Phi : \mathcal{S} \to \mathbb{R}$ a potential function. Let $\gamma \in [0, 1]$ be a discount rate, and $\mathcal{D}_\mathcal{S} \in \Delta(\mathcal{S})$ and $\mathcal{D}_\mathcal{A} \in \Delta(\mathcal{A})$ be distributions over states and actions. Let $R'$ denote $R$ shaped by $\Phi$: $R'(s, a, s') = R(s, a, s') + \gamma\Phi(s') - \Phi(s)$. Then the canonically shaped $R'$ and $R$ are equal: $C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R') = C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)$.*

*Proof.* Let $s, a, s' \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Then by substituting in the definition of $R'$ and using linearity of expectations:

$$\begin{aligned} C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R')(s, a, s') &\triangleq R'(s, a, s') + \mathbb{E}\left[\gamma R'(s', A, S') - R'(s, A, S') - \gamma R'(S, A, S')\right] && (19)\\ &= \left(R(s, a, s') + \gamma\Phi(s') - \Phi(s)\right) && (20)\\ &\quad + \mathbb{E}\left[\gamma R(s', a, S') + \gamma^2\Phi(S') - \gamma\Phi(s')\right]\\ &\quad - \mathbb{E}\left[R(s, A, S') + \gamma\Phi(S') - \Phi(s)\right]\\ &\quad - \mathbb{E}\left[\gamma R(S, A, S') + \gamma^2\Phi(S') - \gamma\Phi(S)\right]\\ &= R(s, a, s') + \mathbb{E}\left[\gamma R(s', a, S') - R(s, A, S') - \gamma R(S, A, S')\right] && (21)\\ &\quad + \left(\gamma\Phi(s') - \Phi(s)\right) - \mathbb{E}\left[\gamma\Phi(s') - \Phi(s)\right]\\ &\quad + \mathbb{E}\left[\gamma^2\Phi(S') - \gamma\Phi(S')\right] - \mathbb{E}\left[\gamma^2\Phi(S') - \gamma\Phi(S)\right]\\ &= R(s, a, s') + \mathbb{E}\left[\gamma R(s', a, S') - R(s, A, S') - \gamma R(S, A, S')\right] && (22)\\ &\triangleq C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)(s, a, s'), && (23)\end{aligned}$$

where the penultimate step uses $\mathbb{E}[\Phi(S')] = \mathbb{E}[\Phi(S)]$ since $S$ and $S'$ are identically distributed. □

**Proposition 4.3.** *Let $\mathcal{S}$ and $\mathcal{A}$ be finite, with $|\mathcal{S}| \geq 2$. Let $\mathcal{D}_{\mathcal{S}} \in \Delta(\mathcal{S})$ and $\mathcal{D}_{\mathcal{A}} \in \Delta(\mathcal{A})$. Let $R,\ \nu : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be reward functions, with $\nu(s,a,s') = \lambda \mathbb{I}[(s,a,s') = (x,u,x')]$, $\lambda \in \mathbb{R}$, $x, x' \in \mathcal{S}$ and $u \in \mathcal{A}$. Let $\Phi_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R)(s,a,s') = C_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R)(s,a,s') - R(s,a,s')$. Then:*

$$\left\|\Phi_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R+\nu) - \Phi_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R)\right\|_{\infty} = \lambda\left(1+\gamma\mathcal{D}_{\mathcal{S}}(x)\right)\mathcal{D}_{\mathcal{A}}(u)\mathcal{D}_{\mathcal{S}}(x'). \tag{3}$$

*Proof.* Observe that:

$$\Phi_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R)(s,a,s') = \mathbb{E}\left[\gamma R(s',A,S') - R(s,A,S') - \gamma R(S,A,S')\right], \tag{24}$$

where $S$ and $S'$ are random variables independently sampled from $\mathcal{D}_{\mathcal{S}}$, and $A$ independently sampled from $\mathcal{D}_{\mathcal{A}}$.

Then:

$$\Phi_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R+\nu) - \Phi_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R) = \Phi_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(\nu). \tag{25}$$

Now:

$$\begin{aligned}
\text{LHS} &\triangleq \left\|\Phi_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R+\nu) - \Phi_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R)\right\|_{\infty} && (26)\\
&= \max_{s,s'\in\mathcal{S}} \left|\mathbb{E}\left[\gamma\nu(s',A,S') - \nu(s,A,S') - \gamma\nu(S,A,S')\right]\right| && (27)\\
&= \max_{s,s'\in\mathcal{S}} \left|\lambda\left(\gamma\mathbb{I}[x=s']\mathcal{D}_{\mathcal{A}}(u)\mathcal{D}_{\mathcal{S}}(x')\right.\right. && (28)\\
&\quad - \left.\left.\mathbb{I}[x=s]\mathcal{D}_{\mathcal{A}}(u)\mathcal{D}_{\mathcal{S}}(x') - \gamma\mathcal{D}_{\mathcal{S}}(x)\mathcal{D}_{\mathcal{A}}(u)\mathcal{D}_{\mathcal{S}}(x')\right)\right| && (29)\\
&= \max_{s,s'\in\mathcal{S}} \left|\lambda\mathcal{D}_{\mathcal{A}}(u)\mathcal{D}_{\mathcal{S}}(x')\left(\gamma\mathbb{I}[x=s'] - \mathbb{I}[x=s] - \gamma\mathcal{D}_{\mathcal{S}}(x)\right)\right| && (30)\\
&= \lambda\left(1+\gamma\mathcal{D}_{\mathcal{S}}(x)\right)\mathcal{D}_{\mathcal{A}}(u)\mathcal{D}_{\mathcal{S}}(x'), && (31)
\end{aligned}$$

where the final step follows by substituting $s = x$ and $s' \neq x$ (using $|\mathcal{S}| \geq 2$). □

**Lemma 4.5.** *The Pearson distance $D_\rho$ is a pseudometric. Moreover, let $a, b \in (0,\infty)$, $c, d \in \mathbb{R}$ and $X, Y$ be random variables. Then it follows that $0 \leq D_\rho(aX+c, bY+d) = D_\rho(X,Y) \leq 1$.*

*Proof.* For a non-constant random variable $V$, define a standardized (zero mean and unit variance) version:

$$Z(V) = \frac{V - \mathbb{E}[V]}{\sqrt{\mathbb{E}\left[(V-\mathbb{E}[V])^2\right]}}. \tag{32}$$

The Pearson correlation coefficient on random variables $A$ and $B$ is equal to the expected product of these standardized random variables:

$$\rho(A,B) = \mathbb{E}\left[Z(A)Z(B)\right]. \tag{33}$$

Let $W$, $X$, $Y$ be random variables.

**Identity**. Have $\rho(X,X) = 1$, so $D_\rho(X,X) = 0$.

**Symmetry**. Have $\rho(X,Y) = \rho(Y,X)$ by commutativity of multiplication, so $D_\rho(X,Y) = D_\rho(Y,X)$.

**Triangle Inequality**. For any random variables $A, B$:

$$\begin{aligned}
\mathbb{E}\left[(Z(A)-Z(B))^2\right] &= \mathbb{E}\left[Z(A)^2 - 2Z(A)Z(B) + Z(B)^2\right] && (34)\\
&= \mathbb{E}\left[Z(A)^2\right] + \mathbb{E}\left[Z(B)^2\right] - 2\mathbb{E}\left[Z(A)Z(B)\right] && (35)\\
&= 2 - 2\mathbb{E}\left[Z(A)Z(B)\right] && (36)\\
&= 2\left(1-\rho(A,B)\right) && (37)\\
&= 4D_\rho(A,B)^2. && (38)
\end{aligned}$$

So:

$$4D_\rho(W,Y)^2 = \mathbb{E}\left[(Z(W)-Z(Y))^2\right] \tag{39}$$
$$= \mathbb{E}\left[(Z(W)-Z(X)+Z(X)-Z(Y))^2\right] \tag{40}$$
$$= \mathbb{E}\left[(Z(W)-Z(X))^2\right] + \mathbb{E}\left[(Z(X)-Z(Y))^2\right] \tag{41}$$
$$+ 2\mathbb{E}\left[(Z(W)-Z(X))(Z(X)-Z(Y))\right]$$
$$= 4D_\rho(W,X)^2 + 4D_\rho(X,Y)^2 + 8\mathbb{E}\left[(Z(W)-Z(X))(Z(X)-Z(Y))\right]. \tag{42}$$

Since $\langle A, B\rangle = \mathbb{E}[AB]$ is an inner product over $\mathbb{R}$, it follows by the Cauchy-Schwarz inequality that $\mathbb{E}[AB] \leq \sqrt{\mathbb{E}[A^2]\mathbb{E}[B^2]}$. So:

$$D_\rho(W,Y)^2 \leq D_\rho(W,X)^2 + D_\rho(X,Y)^2 + 2D_\rho(W,X)D_\rho(X,Y) \tag{43}$$
$$= (D_\rho(W,X) + D_\rho(X,Y))^2. \tag{44}$$

Taking the square root of both sides:

$$D_\rho(W,Y) \leq D_\rho(W,X) + D_\rho(X,Y), \tag{45}$$

as required.

**Positive Affine Invariant and Bounded** $D_\rho(aX+c, bY+d) = D_\rho(X,Y)$ is immediate from $\rho(X,Y)$ invariant to positive affine transformations. Have $-1 \leq \rho(X,Y) \leq 1$, so $0 \leq 1-\rho(X,Y) \leq 2$ thus $0 \leq D_\rho(X,Y) \leq 1$. □

**Theorem 4.7.** *The Equivalent-Policy Invariant Comparison distance is a pseudometric.*

*Proof.* The result follows from $D_\rho$ being a pseudometric. Let $R_A$, $R_B$ and $R_C$ be reward functions mapping from transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ to real numbers $\mathbb{R}$.

**Identity**. Have:

$$D_{\text{EPIC}}(R_A, R_A) = D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)(S,A,S')\right) = 0, \tag{46}$$

since $D_\rho(X,X) = 0$.

**Symmetry**. Have:

$$D_{\text{EPIC}}(R_A, R_B) = D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_B)(S,A,S')\right) \tag{47}$$
$$= D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_B)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)(S,A,S')\right) \tag{48}$$
$$= D_{\text{EPIC}}(R_B, R_A), \tag{49}$$

since $D_\rho(X,Y) = D_\rho(Y,X)$.

**Triangle Inequality**. Have:

$$D_{\text{EPIC}}(R_A, R_C) = D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_C)(S,A,S')\right) \tag{50}$$
$$\leq D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_B)(S,A,S')\right) \tag{51}$$
$$+ D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_B)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_C)(S,A,S')\right) \tag{52}$$
$$= D_{\text{EPIC}}(R_A, R_B) + D_{\text{EPIC}}(R_B, R_C), \tag{53}$$

since $D_\rho(X,Z) \leq D_\rho(X,Y) + D_\rho(Y,Z)$. □

**Theorem 4.8.** *Let $R_A$, $R'_A$, $R_B$, $R'_B$ : $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be reward functions such that $R'_A \equiv R_A$ and $R'_B \equiv R_B$. Then $0 \leq D_{\text{EPIC}}(R'_A, R'_B) = D_{\text{EPIC}}(R_A, R_B) \leq 1$.*

*Proof.* Since $D_{\text{EPIC}}$ is defined in terms of $D_\rho$, the bounds $0 \leq D_{\text{EPIC}}(R'_A, R'_B)$ and $D_{\text{EPIC}}(R_A, R_B) \leq 1$ are immediate from the bounds in lemma 4.5.

Since $R'_A \equiv R_A$ and $R'_B \equiv R_B$, we can write for $X \in \{A, B\}$:

$$R^\lambda_X(s,a,s') = \lambda_X R_X(s,a,s'), \tag{54}$$
$$R'_X(s,a,s') = R^\lambda_X(s,a,s') + \gamma\Phi_X(s') - \Phi_X(s), \tag{55}$$

for some scaling factor $\lambda_X > 0$ and potential function $\Phi_X : \mathcal{S} \to \mathbb{R}$.

By proposition 4.2:

$$C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R'_X\right) = C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R^\lambda_X\right). \tag{56}$$

Moreover, since $C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)$ is defined as an expectation over $R$ and expectations are linear:

$$C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R^\lambda_X\right) = \lambda_X C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R_X\right). \tag{57}$$

Unrolling the definition of $D_{\text{EPIC}}$ and applying this result gives:

$$\begin{aligned}
D_{\text{EPIC}}(R'_A, R'_B) &= D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R'_A\right)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R'_B\right)(S,A,S')\right) && (58)\\
&= D_\rho\left(\lambda_A C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R_A\right)(S,A,S'), \lambda_B C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R_B\right)(S,A,S')\right) && \text{eqs. 56 and 57}\\
&= D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R_A\right)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R_B\right)(S,A,S')\right) && \text{lemma 4.5}\\
&= D_{\text{EPIC}}(R_A, R_B). && \square
\end{aligned}$$

### A.3.3 Nearest Point in Equivalence Class (NPEC) premetric

**Proposition 5.3.** *(1) $D_{L^p,\mathcal{D}}$ is a metric in $L^p$ space, where functions $f$ and $g$ are identified if they agree almost everywhere on $\mathcal{D}$. (2) $D_{L^p,\mathcal{D}}$ is a pseudometric if functions are identified only if they agree at all points.*

*Proof.* (1) $D_{L^p,\mathcal{D}}$ is a metric in the $L^p$ space since $L^p$ is a norm in the $L^p$ space, and $d(x,y) = \|x-y\|$ is always a metric. (2) As $f = g$ at all points implies $f = g$ almost everywhere, certainly $D_{L^p,\mathcal{D}}(R,R) = 0$. Symmetry and triangle inequality do not depend on identity so still hold. □

**Proposition A.1** (Properties of $D^U_{\text{NPEC}}$)**.** *Let $R_A, R_B, R_C : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be bounded reward functions, and $\lambda \geq 0$. Then $D^U_{\text{NPEC}}$:*

- ***Is invariant under $\equiv$ in source**:*
  $D^U_{\text{NPEC}}(R_A, R_C) = D^U_{\text{NPEC}}(R_B, R_C)$ *if* $R_A \equiv R_B$.
- ***Invariant under scale-preserving $\equiv$ in target**:*
  $D^U_{\text{NPEC}}(R_A, R_B) = D^U_{\text{NPEC}}(R_A, R_C)$ *if* $R_B - R_C \equiv \texttt{Zero}$.
- ***Scalable in target**:*
  $D^U_{\text{NPEC}}(R_A, \lambda R_B) = \lambda D^U_{\text{NPEC}}(R_A, R_B)$.
- ***Bounded**:*
  $D^U_{\text{NPEC}}(R_A, R_B) \leq D^U_{\text{NPEC}}(\texttt{Zero}, R_B)$.

*Proof.* We will show each case in turn.

**Invariance under $\equiv$ in source**

If $R_A \equiv R_B$, then:

$$\begin{aligned}
D^U_{\text{NPEC}}(R_A, R_C) &\triangleq \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}(R, R_C) && (59)\\
&= \inf_{R \equiv R_B} D_{L^p,\mathcal{D}}(R, R_C) && (60)\\
&\triangleq D^U_{\text{NPEC}}(R_B, R_C), && (61)\\
& && (62)
\end{aligned}$$

since $R \equiv R_A$ if and only if $R \equiv R_B$ as $\equiv$ is an equivalence relation.

**Invariance under scale-preserving $\equiv$ in target**

If $R_B - R_C \equiv \texttt{Zero}$, then we can write $R_B(s,a,s') - R_C(s,a,s') = \gamma\Phi(s') - \Phi(s)$ for some potential function $\Phi : \mathcal{S} \to \mathbb{R}$. Define $f(R)(s,a,s') = R(s,a,s') - \gamma\Phi(s') + \Phi(s)$. Then for any

reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$:

$$\begin{aligned}
D_{L^p,\mathcal{D}}(R, R_B) &\triangleq \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |R(s,a,s') - R_B(s,a,s')|^p \right] \right)^{1/p} \\
&= \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |R(s,a,s') - (R_C(s,a,s') + \gamma\Phi(s') - \Phi(s))|^p \right] \right)^{1/p} \\
&= \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |(R(s,a,s') - \gamma\Phi(s') + \Phi(s)) - R_C(s,a,s')|^p \right] \right)^{1/p} \\
&= \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |f(R)(s,a,s') - R_C(s,a,s')|^p \right] \right)^{1/p} \\
&\triangleq D_{L^p,\mathcal{D}}(f(R), R_C), \qquad (63)
\end{aligned}$$

Crucially, note $f(R)$ is a bijection on the equivalence class $[R]$. Now, substituting this into the expression for the NPEC premetric:

$$\begin{aligned}
D^U_{\text{NPEC}}(R_A, R_B) &\triangleq \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}(R, R_B) \\
&= \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}(f(R), R_C) && \text{eq. 63} \\
&= \inf_{f(R) \equiv R_A} D_{L^p,\mathcal{D}}(f(R), R_C) && f \text{ bijection on } [R] \\
&= \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}(R, R_C) && f \text{ bijection on } [R] \\
&\triangleq D^U_{\text{NPEC}}(R_A, R_C). \qquad (64)
\end{aligned}$$

**Scalable in target**

First, note that $D_{L^p,\mathcal{D}}$ is absolutely scalable in both arguments:

$$\begin{aligned}
D_{L^p,\mathcal{D}}(\lambda R_A, \lambda R_B) &\triangleq \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |\lambda R_A(s,a,s') - \lambda R_B(s,a,s')|^p \right] \right)^{1/p} \\
&= \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |\lambda|^p \, |R_A(s,a,s') - R_B(s,a,s')|^p \right] \right)^{1/p} && |\cdot| \text{ absolutely scalable} \\
&= \left( |\lambda|^p \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |R_A(s,a,s') - R_B(s,a,s')|^p \right] \right)^{1/p} && \mathbb{E} \text{ linear} \\
&= |\lambda| \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |R_A(s,a,s') - R_B(s,a,s')|^p \right] \right)^{1/p} \\
&\triangleq |\lambda| \, D_{L^p,\mathcal{D}}(R_A, R_B). \qquad (65)
\end{aligned}$$

Now, for $\lambda > 0$, applying this to $D^U_{\text{NPEC}}$:

$$\begin{aligned}
D^U_{\text{NPEC}}(R_A, \lambda R_B) &\triangleq \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}(R, \lambda R_B) && (66) \\
&= \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}(\lambda R, \lambda R_B) && R \equiv \lambda R \qquad (67) \\
&= \inf_{R \equiv R_A} \lambda D_{L^p,\mathcal{D}}(R, R_B) && \text{eq. 65} \qquad (68) \\
&= \lambda \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}(R, R_B) && (69) \\
&\triangleq \lambda D^U_{\text{NPEC}}(R_A, R_B). && (70)
\end{aligned}$$

In the case $\lambda = 0$, then:

$$D^U_{\text{NPEC}}(R_A, \texttt{Zero}) \triangleq \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}(R, \texttt{Zero}) \tag{71}$$

$$= \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}\left(\frac{1}{2}R, \texttt{Zero}\right) \qquad R \equiv \frac{1}{2}R \tag{72}$$

$$= \inf_{R \equiv R_A} \frac{1}{2} D_{L^p,\mathcal{D}}(R, \texttt{Zero}) \tag{73}$$

$$= \frac{1}{2} \inf_{R \equiv R_A} D_{L^p,\mathcal{D}}(R, \texttt{Zero}) \tag{74}$$

$$= \frac{1}{2} D^U_{\text{NPEC}}(R_A, \texttt{Zero}). \tag{75}$$

Rearranging, we have:

$$D^U_{\text{NPEC}}(R_A, \texttt{Zero}) = 0. \tag{76}$$

**Bounded**

Let $d \triangleq D_{\text{NPEC}}(\texttt{Zero}, R_B)$. Then for any $\epsilon > 0$, there exists some potential function $\Phi : \mathcal{S} \to \mathbb{R}$ such that the $L^p$ distance of the potential shaping $R(s, a, s') \triangleq \gamma\Phi(s) - \Phi(s)$ from $R_B$ satisfies:

$$D_{L^p,\mathcal{D}}(R, R_B) \leq d + \epsilon. \tag{77}$$

Let $\lambda \in [0, 1]$. Define:

$$R'_\lambda(s, a, s') \triangleq \lambda R_A(s, a, s') + R(s, a, s'). \tag{78}$$

Now:

$$D_{L^p,\mathcal{D}}(R'_\lambda, R) \triangleq \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |R'_\lambda(s, a, s') - R(s, a, s')|^p \right] \right)^{1/p} \tag{79}$$

$$= \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |\lambda R_A(s, a, s')|^p \right] \right)^{1/p} \tag{80}$$

$$= \left( |\lambda|^p \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |R_A(s, a, s')|^p \right] \right)^{1/p} \tag{81}$$

$$= |\lambda| \left( \mathop{\mathbb{E}}_{s,a,s' \sim \mathcal{D}} \left[ |R_A(s, a, s')|^p \right] \right)^{1/p} \tag{82}$$

$$= |\lambda| D_{L^p,\mathcal{D}}(R_A, \texttt{Zero}). \tag{83}$$

Since $R_A$ is bounded, $D_{L^p,\mathcal{D}}(R_A, \texttt{Zero})$ must be finite, so:

$$\lim_{\lambda \to 0^+} D_{L^p,\mathcal{D}}(R'_\lambda, R) = 0. \tag{84}$$

It follows that for any $\epsilon > 0$ there exists some $\lambda > 0$ such that:

$$D_{L^p,\mathcal{D}}(R'_\lambda, R) \leq \epsilon. \tag{85}$$

Note that $R_A \equiv R'_\lambda$ for all $\lambda > 0$. So:

$$D_{\text{NPEC}}(R_A, R_B) \leq D_{L^p,\mathcal{D}}(R'_\lambda, R_B) \tag{86}$$

$$\leq D_{L^p,\mathcal{D}}(R'_\lambda, R) + D_{L^p,\mathcal{D}}(R, R_B) \qquad \text{prop. 5.3} \tag{87}$$

$$\leq \epsilon + (d + \epsilon) \qquad \text{eq. 77 and eq. 85} \tag{88}$$

$$= d + 2\epsilon. \tag{89}$$

Since $\epsilon > 0$ can be made arbitrarily small, it follows:

$$D_{\text{NPEC}}(R_A, R_B) \leq d \triangleq D_{\text{NPEC}}(\texttt{Zero}, R_B), \tag{90}$$

completing the proof. □

**Theorem 5.5.** *$D_{\mathrm{NPEC}}$ is a premetric on the space of bounded reward functions. Moreover, let $R_A, R_A{}', R_B, R_B{}' : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be bounded reward functions such that $R_A \equiv R_A{}'$ and $R_B \equiv R_B{}'$. Then $0 \leq D_{\mathrm{NPEC}}(R_A{}', R_B{}') = D_{\mathrm{NPEC}}(R_A, R_B) \leq 1$.*

*Proof.* We will first prove $D_{\mathrm{NPEC}}$ is a premetric, and then prove it is invariant and bounded.

**Premetric**

First, we will show that $D_{\mathrm{NPEC}}$ is a premetric.

*Respects identity:* $D_{\mathrm{NPEC}}(R_A, R_A) = 0$

If $D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_A) = 0$ then $D_{\mathrm{NPEC}}(R_A, R_A) = 0$ as required. Suppose from now on that $D^U_{\mathrm{NPEC}}(R_A, R_A) \neq 0$. It follows from prop 5.3 that $D_{L^p,\mathcal{D}}(R_A, R_A) = 0$. Since $X \equiv X$, 0 is an upper bound for $D^U_{\mathrm{NPEC}}(R_A, R_A)$. By prop 5.3 $D_{L^p,\mathcal{D}}$ is non-negative, so this is also a lower bound for $D^U_{\mathrm{NPEC}}(R_A, R_A)$. So $D^U_{\mathrm{NPEC}}(R_A, R_A) = 0$ and:

$$D_{\mathrm{NPEC}}(R_A, R_A) = \frac{D^U_{\mathrm{NPEC}}(R_A, R_A)}{D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_A)} = \frac{0}{D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_A)} = 0. \tag{91}$$

*Well-defined:* $D_{\mathrm{NPEC}}(R_A, R_B) \geq 0$

By prop 5.3, it follows that $D_{L^p,\mathcal{D}}(R, R_B) \geq 0$ for all reward functions $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Thus 0 is a lower bound for $\{D_{L^p,\mathcal{D}}(R, R_B) \mid R : \mathcal{S} \times \mathcal{A} \times \mathcal{S}\}$, and thus certainly a lower bound for $\{D_{L^p,\mathcal{D}}(R, Y) \mid R \equiv X\}$ for any reward function $X$. Since the infimum is the *greatest* lower bound, it follows that for any reward function $X$:

$$D^U_{\mathrm{NPEC}}(X, R_B) \triangleq \inf_{R \equiv X} D_{L^p,\mathcal{D}}(R, R_B) \geq 0. \tag{92}$$

In the case that $D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B) = 0$, then $D_{\mathrm{NPEC}}(R_A, R_B) = 0$ which is non-negative. From now on, suppose that $D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B) \neq 0$. The quotient of a non-negative value with a positive value is non-negative, so:

$$D_{\mathrm{NPEC}}(R_A, R_B) = \frac{D^U_{\mathrm{NPEC}}(R_A, R_B)}{D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B)} \geq 0. \tag{93}$$

**Invariant and Bounded**

Since $R'_B \equiv R_B$, we have $R'_B - \lambda R_B \equiv \texttt{Zero}$ for some $\lambda > 0$. By proposition A.1, $D^U_{\mathrm{NPEC}}$ is invariant under scale-preserving $\equiv$ in target and scalable in target. That is, for any reward $R$:

$$D^U_{\mathrm{NPEC}}(R, R'_B) = D^U_{\mathrm{NPEC}}(R, \lambda R_B) = \lambda D^U_{\mathrm{NPEC}}(R, R_B). \tag{94}$$

In particular, $D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B{}') = \lambda D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B)$. As $\lambda > 0$, it follows that $D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B{}') = 0 \iff D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B) = 0$.

Suppose $D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B) = 0$. Then $D_{\mathrm{NPEC}}(R, R_B) = 0 = D_{\mathrm{NPEC}}(R, R_B{}')$ for any reward $R$, so the result trivially holds. From now on, suppose $D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B) \neq 0$.

By proposition A.1, $D^U_{\mathrm{NPEC}}$ is invariant to $\equiv$ in source. That is, $D^U_{\mathrm{NPEC}}(R_A, R_B) = D^U_{\mathrm{NPEC}}(R'_A, R_B)$, so:

$$D_{\mathrm{NPEC}}(R'_A, R_B) = \frac{D^U_{\mathrm{NPEC}}(R'_A, R_B)}{D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B)} = \frac{D^U_{\mathrm{NPEC}}(R_A, R_B)}{D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B)} = D_{\mathrm{NPEC}}(R_A, R_B). \tag{95}$$

By eq. (94):

$$D_{\mathrm{NPEC}}(R_A, R'_B) = \frac{\lambda D^U_{\mathrm{NPEC}}(R_A, R_B)}{\lambda D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B)} = \frac{D^U_{\mathrm{NPEC}}(R_A, R_B)}{D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B)} = D_{\mathrm{NPEC}}(R_A, R_B). \tag{96}$$

Since $D_{\mathrm{NPEC}}$ is a premetric it is non-negative. By the boundedness property of proposition A.1, $D^U_{\mathrm{NPEC}}(R, R_B) \leq D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B)$, so:

$$D_{\mathrm{NPEC}}(R_A, R_B) = \frac{D^U_{\mathrm{NPEC}}(R_A, R_B)}{D^U_{\mathrm{NPEC}}(\texttt{Zero}, R_B)} \leq 1, \tag{97}$$

which completes the proof. □

Note when $D_{L^p,\mathcal{D}}$ is a metric, then $D_{\mathrm{NPEC}}(X, Y) = 0$ if and only if $X = Y$.

**Proposition A.2.** $D_{\mathrm{NPEC}}$ *is* not *symmetric in the undiscounted case.*

*Proof.* We will provide a counterexample showing that $D_{\mathrm{NPEC}}$ is not symmetric.

Choose the state space $\mathcal{S}$ to be binary $\{0, 1\}$ and the actions $\mathcal{A}$ to be the singleton $\{0\}$. Choose the coverage distribution $\mathcal{D}$ to be uniform on $s \xrightarrow{0} s$ for $s \in \mathcal{S}$. Take $\gamma = 1$, i.e. undiscounted. Note that as the successor state is always the same as the start state, potential shaping has no effect on $D_{L^p,\mathcal{D}}$, so WLOG we will assume potential shaping is always zero.

Now, take $R_A(s) = 2s$ and $R_B(s) = 1$. Take $p = 1$ for the $L^p$ distance. Observe that $D_{L^p,\mathcal{D}}(\texttt{Zero}, R_A) = \frac{1}{2}(|0| + |2|) = 1$ and $D_{L^p,\mathcal{D}}(\texttt{Zero}, R_B) = \frac{1}{2}(|1| + |1|) = 1$. Since potential shaping has no effect, $D^U_{\mathrm{NPEC}}(\texttt{Zero}, R) = D_{L^p,\mathcal{D}}(\texttt{Zero}, R)$ and so $D_{\mathrm{NPEC}}(\texttt{Zero}, R_A) = 1$ and $D_{\mathrm{NPEC}}(\texttt{Zero}, R_B) = 1$.

Now:

$$D^U_{\mathrm{NPEC}}(R_A, R_B) = \inf_{\lambda>0} D_{L^p,\mathcal{D}}(\lambda R_A, R_B) \tag{98}$$

$$= \inf_{\lambda>0} \frac{1}{2}(|1| + |2\lambda - 1|) \tag{99}$$

$$= \frac{1}{2}, \tag{100}$$

with the infimum attained at $\lambda = \frac{1}{2}$. But:

$$D^U_{\mathrm{NPEC}}(R_B, R_A) = \inf_{\lambda>0} D_{L^p,\mathcal{D}}(\lambda R_B, R_A) \tag{101}$$

$$= \inf_{\lambda>0} \frac{1}{2} f(\lambda) \tag{102}$$

$$= \frac{1}{2} \inf_{\lambda>0} f(\lambda), \tag{103}$$

where:

$$f(\lambda) = |\lambda| + |2 - \lambda|, \qquad \lambda > 0. \tag{104}$$

Note that:

$$f(\lambda) = \begin{cases} 2 & \lambda \in (0, 2], \\ 2\lambda - 2 & \lambda \in (2, \infty). \end{cases} \tag{105}$$

So $f(\lambda) \geq 2$ on all of its domain, thus:

$$D^U_{\mathrm{NPEC}}(R_B, R_A) = 1. \tag{106}$$

Consequently:

$$D_{\mathrm{NPEC}}(R_A, R_B) = \frac{1}{2} \neq 1 = D_{\mathrm{NPEC}}(R_B, R_A), \tag{107}$$

so $D_{\mathrm{NPEC}}$ is not symmetric. □

### A.4 Full Normalization Variant of EPIC

Previously, we used the Pearson distance $D_\rho$ to compare the canonicalized rewards. Pearson distance is naturally invariant to scaling. An alternative is to explicitly normalize the canonicalized rewards, and then compare them using any metric over functions.

**Definition A.3** (Normalized Reward)**.** *Let $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be a bounded reward function. Let $\|\cdot\|$ be a norm on the vector space of reward functions over the real field. Then the normalized $R$ is:*

$$R^N(s, a, s') = \frac{R(s, a, s')}{\|R\|} \tag{108}$$

Note that $(\lambda R)^N = R^N$ for any $\lambda > 0$ as norms are absolutely homogeneous.

We say a reward is *standardized* if it has been canonicalized and then normalized.

**Definition A.4** (Standardized Reward)**.** *Let $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be a bounded reward function. Then the standardized $R$ is:*

$$R^S = \left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R\right)\right)^N. \tag{109}$$

Now, we can define a pseudometric based on the direct distance between the standardized rewards.

**Definition A.5** (Direct Distance Standardized Reward)**.** *Let $\mathcal{D}$ be some coverage distribution over transitions $s \xrightarrow{a} s'$. Let $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{A}$ be some distributions over states $\mathcal{S}$ and $\mathcal{A}$ respectively. Let $S, A, S'$ be random variables jointly sampled from $\mathcal{D}$. The* Direct Distance Standardized Reward pseudometric *between two reward functions $R_A$ and $R_B$ is the $L^p$ distance between their standardized versions over $\mathcal{D}$:*

$$D_{\mathrm{DDSR}}(R_A, R_B) = \frac{1}{2} D_{L^p,\mathcal{D}}\left(R_A^S(S,A,S'), R_B^S(S,A,S')\right), \tag{110}$$

*where the normalization step, $R^N$, uses the $L^p$ norm.*

For brevity, we omit the proof that $D_{\mathrm{DDSR}}$ is a pseudometric, but this follows from $D_{L^p,\mathcal{D}}$ being a pseudometric in a similar fashion to theorem 4.7. Note it additionally is invariant to equivalence classes, similarly to EPIC.

**Theorem A.6.** *Let $R_A$, $R_A{}'$, $R_B$ and $R_B{}'$ be reward functions mapping from transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ to real numbers $\mathbb{R}$ such that $R_A \equiv R_A{}'$ and $R_B \equiv R_B{}'$. Then:*

$$0 \leq D_{\mathrm{DDSR}}(R_A', R_B') = D_{\mathrm{DDSR}}(R_A, R_B) \leq 1. \tag{111}$$

*Proof.* The invariance under the equivalence class follows from $R^S$ being invariant to potential shaping and scale in $R$. The non-negativity follows from $D_{L^p,\mathcal{D}}$ being a pseudometric. The upper bound follows from the rewards being normalized to norm 1 and the triangle inequality:

$$\begin{aligned}
D_{\mathrm{DDSR}}(R_A, R_B) &= \frac{1}{2}\|R_A^S - R_B^S\| && (112)\\
&\leq \frac{1}{2}\left(\|R_A^S\| + \|R_B^S\|\right) && (113)\\
&= \frac{1}{2}\left(1+1\right) && (114)\\
&= 1. && \square
\end{aligned}$$

Since both DDSR and EPIC are pseudometrics and invariant on equivalent rewards, it is interesting to consider the connection between them. In fact, under the $L^2$ norm, then DDSR recovers EPIC. First, we will show that canonical shaping centers the reward functions.

**Lemma A.7** (The Canonically Shaped Reward is Mean Zero)**.** *Let $R$ be a reward function mapping from transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ to real numbers $\mathbb{R}$. Then:*

$$\mathbb{E}\left[C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R\right)(S,A,S')\right] = 0. \tag{115}$$

*Proof.* Let $X$, $U$ and $X'$ be random variables that are independent of $S$, $A$ and $S'$ but identically distributed.

$$\begin{aligned}
\mathrm{LHS} &\triangleq \mathbb{E}\left[C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R\right)(S,A,S')\right] && (116)\\
&= \mathbb{E}\left[R(S,A,S') + \gamma R(S',U,X') - R(S,U,X') - \gamma R(X,U,X')\right] && (117)\\
&= \mathbb{E}\left[R(S,A,S')\right] + \gamma\mathbb{E}\left[R(S',U,X')\right] - \mathbb{E}\left[R(S,U,X')\right] - \gamma\mathbb{E}\left[R(X,U,X')\right] && (118)\\
&= \mathbb{E}\left[R(S,U,X')\right] + \gamma\mathbb{E}\left[R(X,U,X')\right] - \mathbb{E}\left[R(S,U,X')\right] - \gamma\mathbb{E}\left[R(X,U,X')\right] && (119)\\
&= 0, && (120)
\end{aligned}$$

where the penultimate step follows since $A$ is identically distributed to $U$, and $S'$ is identically distributed to $X'$ and therefore to $X$. $\square$

**Theorem A.8.** *$D_{\mathrm{DDSR}}$ with $p = 2$ is equivalent to $D_{\mathrm{EPIC}}$. Let $R_A$ and $R_B$ be reward functions mapping from transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ to real numbers $\mathbb{R}$. Then:*

$$D_{\mathrm{DDSR}}(R_A, R_B) = D_{\mathrm{EPIC}}(R_A, R_B). \tag{121}$$

*Proof.* Recall from the proof of lemma 4.5 that:

$$D_\rho(U,V) = \frac{1}{2}\sqrt{\mathbb{E}\left[(Z(U)-Z(V))^2\right]} \tag{122}$$

$$= \frac{1}{2}\left\|Z(U)-Z(V)\right\|_2, \tag{123}$$

where $\|\cdot\|_2$ is the $L^2$ norm (treating the random variables as functions on a measure space) and $Z(U)$ is a centered (zero-mean) and rescaled (unit variance) random variable. By lemma A.7, the canonically shaped reward functions are already centered under the joint distribution $\mathcal{D}_\mathcal{S} \times \mathcal{D}_\mathcal{A} \times \mathcal{D}_\mathcal{S}$, and normalization by the $L^2$ norm also ensures they have unit variance. Consequently:

$$D_{\mathrm{EPIC}}(R_A, R_B) = D_\rho\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R_A\right)(S,A,S'), C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R_B\right)(S,A,S')\right) \tag{124}$$

$$= \frac{1}{2}\left\|\left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R_A\right)(S,A,S')\right)^N - \left(C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}\left(R_B\right)(S,A,S')\right)^N\right\|_2 \tag{125}$$

$$= \frac{1}{2}\left\|R_A^S(S,A,S') - R_B^S(S,A,S')\right\|_2 \tag{126}$$

$$= \frac{1}{2}D_{L^p,\mathcal{D}}\left(R_A^S(S,A,S'), R_B^S(S,A,S')\right) \tag{127}$$

$$= D_{\mathrm{DDSR}}(R_A, R_B), \tag{128}$$

completing the proof. □

### A.5 Regret Bound

In this section, we derive an upper bound on the regret in terms of the EPIC distance. Specifically, given two reward functions $R_A$ and $R_B$ with optimal policies $\pi_A^*$ and $\pi_B^*$, we show that the regret (under reward $R_A$) of using policy $\pi_B^*$ instead of a policy $\pi_A^*$ is bounded by a function of $D_{\mathrm{EPIC}}(R_A, R_B)$. First, in section A.5.1 we derive a bound for MDPs with finite state and action spaces. In section A.6 we then present an alternative bound for MDPs with arbitrary state and action spaces and Lipschitz reward functions. Finally, in section A.7 we show that in both cases the regret tends to 0 as $D_{\mathrm{EPIC}}(R_A, R_B) \to 0$.

#### A.5.1 Discrete MDPs

We start in lemma A.9 by showing that $L^2$ distance upper bounds $L^1$ distance. Next, in lemma A.10 we show regret is bounded by the $L^1$ distance between reward functions using an argument similar to [20]. Then in lemma A.11 we relate regret bounds for standardized rewards $R^S$ to the original reward $R$. Finally, in theorem 4.9 we use section A.4 to express $D_{\mathrm{EPIC}}$ in terms of the $L^2$ distance on standardized rewards, deriving a bound on regret in terms of the EPIC distance.

**Lemma A.9.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $f : \Omega \to \mathbb{R}$ a measurable function whose absolute value raised to the $n$-th power for $n \in \{1, 2\}$ has a finite expectation. Then the $L^1$ norm of $f$ is bounded above by the $L^2$ norm:*

$$\|f\|_1 \leq \|f\|_2. \tag{129}$$

*Proof.* Let $X$ be a random variable sampled from $P$, and consider the variance of $f(X)$:

$$\mathbb{E}\left[\left(|f(X)| - \mathbb{E}\left[|f(X)|\right]\right)^2\right] = \mathbb{E}\left[|f(X)|^2 - 2|f(X)|\mathbb{E}\left[|f(X)|\right] + \mathbb{E}\left[|f(X)|\right]^2\right] \tag{130}$$

$$= \mathbb{E}\left[|f(X)|^2\right] - 2\mathbb{E}\left[|f(X)|\right]\mathbb{E}\left[|f(X)|\right] + \mathbb{E}\left[|f(X)|\right]^2 \tag{131}$$

$$= \mathbb{E}\left[|f(X)|^2\right] - \mathbb{E}\left[|f(X)|\right]^2 \tag{132}$$

$$\geq 0. \tag{133}$$

Rearranging terms, we have

$$\|f\|_2^2 = \mathbb{E}\left[|f(X)|^2\right] \geq \mathbb{E}\left[|f(X)|\right]^2 = \|f\|_1^2. \tag{134}$$

Taking the square root of both sides gives:

$$\|f\|_1 \leq \|f\|_2, \tag{135}$$

as required. □

**Lemma A.10.** *Let $M$ be an MDP\R with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$. Let $R_A, R_B : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be rewards. Let $\pi_A^*$ and $\pi_B^*$ be policies optimal for rewards $R_A$ and $R_B$ in $M$. Let $\mathcal{D}_\pi(t, s_t, a_t, s_{t+1})$ denote the distribution over trajectories that policy $\pi$ induces in $M$ at time step $t$. Let $\mathcal{D}(s, a, s')$ be the (stationary) coverage distribution over transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ used to compute $D_{\text{EPIC}}$. Suppose that there exists some $K > 0$ such that $K\mathcal{D}(s_t, a_t, s'_{t+1}) \geq \mathcal{D}_\pi(t, s_t, a_t, s'_{t+1})$ for all time steps $t \in \mathbb{N}$, triples $s_t, a_t, s_{t+1} \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and policies $\pi \in \{\pi_A^*, \pi_B^*\}$. Then the regret under $R_A$ from executing $\pi_B^*$ optimal for $R_B$ instead of $\pi_A^*$ is at most:*

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq \frac{2K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B). \tag{136}$$

*Proof.* Noting $G_{R_A}(\pi)$ is maximized when $\pi = \pi_A^*$, it is immediate that

$$\begin{aligned} G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) &= |G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*)| && (137)\\ &= |(G_{R_A}(\pi_A^*) - G_{R_B}(\pi_B^*)) + (G_{R_B}(\pi_B^*) - G_{R_A}(\pi_B^*))| && (138)\\ &\leq |G_{R_A}(\pi_A^*) - G_{R_B}(\pi_B^*)| + |G_{R_B}(\pi_B^*) - G_{R_A}(\pi_B^*)|\,. && (139) \end{aligned}$$

We will show that both these terms are bounded above by $\frac{K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B)$, from which the result follows.

First, we will show that for policy $\pi \in \{\pi_A^*, \pi_B^*\}$:

$$|G_{R_A}(\pi) - G_{R_B}(\pi)| \leq \frac{K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B). \tag{140}$$

Let $T$ be the horizon of $M$. This may be infinite ($T = \infty$) when $\gamma < 1$; note since $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ is bounded, so are $R_A, R_B$ so the discounted infinite returns $G_{R_A}(\pi), G_{R_B}(\pi)$ converge (as do their differences). Writing $\tau = (s_0, a_0, s_1, a_1, \cdots)$, we have for any policy $\pi$:

$$\begin{aligned} \Delta &\triangleq |G_{R_A}(\pi) - G_{R_B}(\pi)| && (141)\\ &= \left| \mathbb{E}_{\tau \sim \mathcal{D}_\pi} \left[ \sum_{t=0}^{T} \gamma^t \left( R_A(s_t, a_t, s_{t+1}) - R_B(s_t, a_t, s_{t+1}) \right) \right] \right| && (142)\\ &\leq \mathop{\mathbb{E}}_{\tau \sim \mathcal{D}_\pi} \left[ \sum_{t=0}^{T} \gamma^t \left| R_A(s_t, a_t, s_{t+1}) - R_B(s_t, a_t, s_{t+1}) \right| \right] && (143)\\ &= \sum_{t=0}^{T} \gamma^t \mathop{\mathbb{E}}_{s_t, a_t, s_{t+1} \sim \mathcal{D}_\pi} \left[ \left| R_A(s_t, a_t, s_{t+1}) - R_B(s_t, a_t, s_{t+1}) \right| \right] && (144)\\ &= \sum_{t=0}^{T} \gamma^t \sum_{s_t, a_t, s_{t+1} \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \mathcal{D}_\pi(t, s_t, a_t, s_{t+1}) \left| R_A(s_t, a_t, s_{t+1}) - R_B(s_t, a_t, s_{t+1}) \right|. && (145) \end{aligned}$$

Let $\pi \in \{\pi_A^*, \pi_B^*\}$. By assumption, $\mathcal{D}_\pi(t, s_t, a_t, s'_{t+1}) \leq K\mathcal{D}(s_t, a_t, s'_{t+1})$, so:

$$\begin{aligned} \Delta &\leq K \sum_{t=0}^{T} \gamma^t \sum_{s_t, a_t, s_{t+1} \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \mathcal{D}(s_t, a_t, s_{t+1}) \left| R_A(s_t, a_t, s_{t+1}) - R_B(s_t, a_t, s_{t+1}) \right| && (146)\\ &= K \sum_{t=0}^{T} \gamma^t D_{L^1,\mathcal{D}}(R_A, R_B) && (147)\\ &= \frac{K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B), && (148) \end{aligned}$$

as required.

In particular, substituting $\pi = \pi_B^*$ gives:

$$|G_{R_B}(\pi_B^*) - G_{R_A}(\pi_B^*)| = |G_{R_A}(\pi_B^*) - G_{R_B}(\pi_B^*)| \leq \frac{K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B). \tag{149}$$

Rearranging gives:

$$G_{R_A}(\pi_B^*) \geq G_{R_B}(\pi_B^*) - \frac{K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B). \tag{150}$$

So certainly:

$$G_{R_A}(\pi_A^*) = \max_{\pi} G_{R_A}(\pi) \geq G_{R_B}(\pi_B^*) - \frac{K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B). \tag{151}$$

By a symmetric argument, substituting $\pi = \pi_A^*$ gives:

$$G_{R_B}(\pi_B^*) = \max_{\pi} G_{R_B}(\pi) \geq G_{R_A}(\pi_A^*) - \frac{K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B). \tag{152}$$

Eqs. 151 and 152 respectively give $G_{R_B}(\pi_B^*) - G_{R_A}(\pi_A^*) \leq \frac{K}{1-\gamma}$ and $G_{R_A}(\pi_A^*) - G_{R_B}(\pi_B^*) \leq \frac{K}{1-\gamma}$. Combining these gives:

$$|G_{R_A}(\pi_A^*) - G_{R_B}(\pi_B^*)| \leq \frac{K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B). \tag{153}$$

Substituting inequalities 149 and 153 into eq. 139 yields the required result. □

Note that if $\mathcal{D} = \mathcal{D}_{\text{unif}}$, uniform over $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$, then $K \leq |\mathcal{S}|^2|\mathcal{A}|$.

**Lemma A.11.** *Let $M$ be an MDP\R with state and action spaces $\mathcal{S}$ and $\mathcal{A}$. Let $R_A, R_B : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ be bounded rewards. Let $\pi_A^*$ and $\pi_B^*$ be policies optimal for rewards $R_A$ and $R_B$ in $M$. Suppose the regret under the standardized reward $R_A^S$ from executing $\pi_B^*$ instead of $\pi_A^*$ is upper bounded by some $U \in \mathbb{R}$:*

$$G_{R_A^S}(\pi_A^*) - G_{R_A^S}(\pi_B^*) \leq U. \tag{154}$$

*Then the regret under the original reward $R_A$ is bounded by:*

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B) \leq 4U\|R_A\|_2. \tag{155}$$

*Proof.* Recall that

$$R^S = \frac{C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)}{\|C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)\|_2}, \tag{156}$$

where $C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)$ is simply $R$ shaped with some (bounded) potential $\Phi$. It follows that:

$$G_{R^S}(\pi) = \frac{1}{\|C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)\|_2} G_{C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)}(\pi) \tag{157}$$

$$= \frac{1}{\|C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)\|_2} \left(G_R(\pi) - \mathbb{E}_{s_0\sim\mu}\left[\Phi(s_0)\right]\right), \tag{158}$$

where $s_0$ depends only on the initial state distribution $\mu$. [†] Since $s_0$ does not depend on $\pi$, the terms cancel when taking the difference in returns:

$$G_{R_A^S}(\pi_A^*) - G_{R_A^S}(\pi_B^*) = \frac{1}{\|C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)\|_2} \left(G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*)\right). \tag{159}$$

Combining this with eq 154 gives

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq U\|C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)\|_2. \tag{160}$$

Finally, we will bound $\|C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R_A)\|_2$ in terms of $\|R_A\|_2$, completing the proof. Recall:

$$C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)(s,a,s') = R(s,a,s') + \mathbb{E}\left[\gamma R(s',A,S') - R(s,A,S') - \gamma R(S,A,S')\right], \tag{161}$$

where $S$ and $S'$ are random variables independently sampled from $\mathcal{D}_\mathcal{S}$ and $A$ sampled from $\mathcal{D}_\mathcal{A}$. By the triangle inequality on the $L^2$ norm and linearity of expectations, we have:

$$\|C_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{A}}(R)\|_2 \leq \|R\|_2 + \gamma\|f\|_2 + \|g\|_2 + \gamma|c|, \tag{162}$$

[†] In the finite-horizon case, there is also a term $\gamma^T\Phi(s_T)$, where $s_T$ is the fixed terminal state. Since $s_T$ is fixed, it also cancels in eq. 159. This term can be neglected in the discounted infinite-horizon case as $\gamma^T\Phi(s_T) \to 0$ as $T \to \infty$ for any bounded $\Phi$.

where $f(s,a,s') = \mathbb{E}\left[R(s',A,S')\right]$, $g(s,a,s') = \mathbb{E}\left[R(s,A,S')\right]$ and $c = \mathbb{E}\left[R(S,A,S')\right]$. Letting $X'$ be a random variable sampled from $\mathcal{D}_{\mathcal{S}}$ independently from $S$ and $S'$, have

$$\|f\|_2^2 = \mathbb{E}_{X'}\left[\mathbb{E}\left[R(X',A,S')\right]^2\right] \tag{163}$$
$$\leq \mathbb{E}_{X'}\left[\mathbb{E}\left[R(X',A,S')^2\right]\right] \tag{164}$$
$$= \mathbb{E}\left[R(X',A,S')^2\right] \tag{165}$$
$$= \|R\|_2^2. \tag{166}$$

So $\|f\|_2 \leq \|R\|_2$ and, by an analogous argument, $\|g\|_2 \leq \|R\|_2$. Similarly

$$|c| = |\mathbb{E}\left[R(S,A,S')\right]| \tag{167}$$
$$\leq \mathbb{E}\left[|R(S,A,S')|\right] \tag{168}$$
$$= \|R\|_1 \tag{169}$$
$$\leq \|R\|_2 \qquad \text{lemma A.9.} \tag{170}$$

Combining these results, we have

$$\|C_{\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{A}}}(R)\|_2 \leq 4\|R\|_2. \tag{171}$$

Substituting eq. 171 into eq. 160 yields:

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq 4U\|R_A\|_2, \tag{172}$$

as required. □

**Theorem 4.9.** *Let $M$ be a $\gamma$-discounted MDP\R with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$. Let $R_A, R_B : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be rewards, and $\pi_A^*, \pi_B^*$ be respective optimal policies. Let $\mathcal{D}_\pi(t, s_t, a_t, s_{t+1})$ denote the distribution over transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ induced by policy $\pi$ at time $t$, and $\mathcal{D}(s,a,s')$ be the coverage distribution used to compute $D_{\text{EPIC}}$. Suppose there exists $K > 0$ such that $K\mathcal{D}(s_t, a_t, s'_{t+1}) \geq \mathcal{D}_\pi(t, s_t, a_t, s'_{t+1})$ for all times $t \in \mathbb{N}$, triples $(s_t, a_t, s_{t+1}) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and policies $\pi \in \{\pi_A^*, \pi_B^*\}$. Then the regret under $R_A$ from executing $\pi_B^*$ instead of $\pi_A^*$ is at most $G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq 16K\|R_A\|_2 (1-\gamma)^{-1} D_{\text{EPIC}}(R_A, R_B)$.*

*Proof.* Recall from section A.4 that:

$$D_{\text{EPIC}}(R_A, R_B) = \frac{1}{2}\left\|R_A^S(S,A,S') - R_B^S(S,A,S')\right\|_2. \tag{173}$$

Applying lemma A.9 we obtain:

$$D_{L^1,\mathcal{D}}(R_A^S, R_B^S) = \left\|R_A^S(S,A,S') - R_B^S(S,A,S')\right\|_1 \leq 2D_{\text{EPIC}}(R_A, R_B). \tag{174}$$

Note that $\pi_A^*$ is optimal for $R_A^S$ and $\pi_B^*$ is optimal for $R_B^S$ since the set of optimal policies for $R^S$ is the same as for $R$. Applying lemma A.10 and eq. 174 gives

$$G_{R_A^S}(\pi_A^*) - G_{R_A^S}(\pi_B^*) \leq \frac{2K}{1-\gamma} D_{L^1,\mathcal{D}}(R_A^S, R_B^S) \leq \frac{4K}{1-\gamma} D_{\text{EPIC}}(R_A, R_B). \tag{175}$$

Since $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ is bounded, $R_A$ and $R_B$ must be bounded, so we can apply lemma A.11, giving:

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq \frac{16K\|R_A\|_2}{1-\gamma} D_{\text{EPIC}}(R_A, R_B), \tag{176}$$

completing the proof. □

### A.6 Lipschitz Reward Functions

In this section, we generalize the previous results to MDPs with continuous state and action spaces. The challenge is that even though the spaces may be continuous, the distribution $\mathcal{D}_{\pi^*}$ induced by an optimal policy $\pi^*$ may only have support on some measure zero set of transitions $B$. However, the expectation over a continuous distribution $\mathcal{D}$ is unaffected by the reward at any measure zero subset of points. Accordingly, the reward can be varied *arbitrarily* on transitions $B$ – causing arbitrarily

small or large regret – while leaving the EPIC distance fixed. To rule out this pathological case, we assume the rewards are Lipschitz smooth. This guarantees that if the expected difference between rewards is small on a given region, then all points in this region have bounded reward difference.

We start by defining a relaxation of the Wasserstein distance $W_\alpha$ in definition A.12. In lemma A.13 we then bound the expected value under distribution $\mu$ in terms of the expected value under alternative distribution $\nu$ plus $W_\alpha(\mu, \nu)$. Next, in lemma A.14 we bound the regret in terms of the $L^1$ distance between the rewards plus $W_\alpha$; this is analogous to lemma A.10 in the finite case. Finally, in theorem A.15 we use the previous results to bound the regret in terms of the EPIC distance plus $W_\alpha$.

**Definition A.12.** *Let $S$ be some set and let $\mu, \nu$ be probability measures on $S$ with finite first moment. We define the* relaxed Wasserstein distance *between $\mu$ and $\nu$ by:*

$$W_\alpha(\mu, \nu) \triangleq \inf_{p \in \Gamma_\alpha(\mu,\nu)} \int \|x - y\| \, dp(x, y), \tag{177}$$

*where $\Gamma_\alpha(\mu, \nu)$ is the set of probability measures on $S \times S$ satisfying for all $x, y \in S$:*

$$\int_S p(x, y) dy = \mu(x), \tag{178}$$

$$\int_S p(x, y) dx \leq \alpha \nu(y). \tag{179}$$

Note that $W_1$ is equal to the (unrelaxed) Wasserstein distance (in the $\ell_1$ norm).

**Lemma A.13.** *Let $S$ be some set and let $\mu, \nu$ be probability measures on $S$. Let $f : S \to \mathbb{R}$ be an $L$-Lipschitz function on the $\ell_1$ norm $\|\cdot\|_1$. Then, for any $\alpha \geq 1$:*

$$\mathbb{E}_{X\sim\mu}\left[|f(X)|\right] \leq \alpha \mathbb{E}_{Y\sim\nu}\left[|f(Y)|\right] + L W_\alpha(\mu, \nu). \tag{180}$$

*Proof.* Let $p \in \Gamma_\alpha(\mu, \nu)$. Then:

$$\begin{aligned}
\mathbb{E}_{X\sim\mu}\left[|f(X)|\right] &\triangleq \int |f(x)| d\mu(x) && \text{definition of } \mathbb{E} && (181)\\
&= \int |f(x)| dp(x, y) && \mu \text{ is a marginal of } p && (182)\\
&\leq \int |f(y)| + L \|x - y\| \, dp(x, y) && f\ L\text{-Lipschitz} && (183)\\
&= \int |f(y)| dp(x, y) + L \int \|x - y\| \, dp(x, y) && && (184)\\
&= \int |f(y)| \int p(x, y) dx dy + L \int \|x - y\| \, dp(x, y) && && (185)\\
&\leq \int |f(y)| \alpha \nu(y) dy + L \int \|x - y\| \, dp(x, y) && \text{eq. 179} && (186)\\
&= \alpha \mathbb{E}_{Y\sim\nu}\left[|f(Y)|\right] + L \int \|x - y\| \, dp(x, y) && \text{definition of } \mathbb{E}. && (187)
\end{aligned}$$

Since this holds for all choices of $p$, we can take the infimum of both sides, giving:

$$\begin{aligned}
\mathbb{E}_{X\sim\mu}\left[|f(X)|\right] &\leq \alpha \mathbb{E}_{Y\sim\nu}\left[|f(Y)|\right] + L \inf_{p \in \Gamma_\alpha(\mu,\nu)} \int \|x - y\| \, dp(x, y) && (188)\\
&= \alpha \mathbb{E}_{Y\sim\nu}\left[|f(Y)|\right] + L W_\alpha(\mu, \nu). && \square
\end{aligned}$$

**Lemma A.14.** *Let $M$ be an MDP\R with state and action spaces $\mathcal{S}$ and $\mathcal{A}$. Let $R_A, R_B : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be $L$-Lipschitz, bounded rewards on the $\ell_1$ norm $\|\cdot\|_1$. Let $\pi_A^*$ and $\pi_B^*$ be policies optimal for rewards $R_A$ and $R_B$ in $M$. Let $\mathcal{D}_{\pi,t}(s_t, a_t, s_{t+1})$ denote the distribution over trajectories that policy $\pi$ induces in $M$ at time step $t$. Let $\mathcal{D}(s, a, s')$ be the (stationary) coverage distribution over transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ used to compute $D_{\text{EPIC}}$. Let $\alpha \geq 1$, and let $B_\alpha(t) = \max_{\pi \in \{\pi_A^*, \pi_B^*\}} W_\alpha\left(\mathcal{D}_{\pi,t}, \mathcal{D}\right)$. Then the regret under $R_A$ from executing $\pi_B^*$ optimal for $R_B$ instead of $\pi_A^*$ is at most:*

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq \frac{2\alpha}{1 - \gamma} D_{L^1,\mathcal{D}}(R_A, R_B) + 4L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t). \tag{189}$$

*Proof.* By the same argument as lemma A.10 up to eq. 145, we have for any policy $\pi$:

$$|G_{R_A}(\pi) - G_{R_B}(\pi)| \leq \sum_{t=0}^{\infty} \gamma^t D_{L^1,\mathcal{D}_{\pi,t}}(R_A, R_B). \tag{190}$$

Let $f(s, a, s') = R_A(s, a, s') - R_B(s, a, s')$, and note $f$ is $2L$-Lipschitz and bounded since $R_A$ and $R_B$ are both $L$-Lipschitz and bounded. Now, by lemma A.13, letting $\mu = \mathcal{D}_{\pi,t}$ and $\nu = \mathcal{D}$, we have:

$$D_{L^1,\mathcal{D}_{\pi,t}}(R_A, R_B) \leq \alpha D_{L^1,\mathcal{D}}(R_A, R_B) + 2LW_\alpha(\mathcal{D}_{\pi,t}, \mathcal{D}). \tag{191}$$

So, for $\pi \in \{\pi_A^*, \pi_B^*\}$, it follows that

$$|G_{R_A}(\pi) - G_{R_B}(\pi)| \leq \frac{\alpha}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B) + 2L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t). \tag{192}$$

By the same argument as for eq. 149 to 153 in lemma A.10, it follows that

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq \frac{2\alpha}{1-\gamma} D_{L^1,\mathcal{D}}(R_A, R_B) + 4L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t), \tag{193}$$

completing the proof. □

**Theorem A.15.** *Let $M$ be an MDP\R with state and action spaces $\mathcal{S}$ and $\mathcal{A}$. Let $R_A, R_B : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be bounded, $L$-Lipschitz rewards on the $\ell_1$ norm $\|\cdot\|_1$. Let $\pi_A^*$ and $\pi_B^*$ be policies optimal for rewards $R_A$ and $R_B$ in $M$. Let $\mathcal{D}_\pi(t, s_t, a_t, s_{t+1})$ denote the distribution over trajectories that policy $\pi$ induces in $M$ at time step $t$. Let $\mathcal{D}(s, a, s')$ be the (stationary) coverage distribution over transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ used to compute $D_{\text{EPIC}}$. Let $\alpha \geq 1$, and let $B_\alpha(t) = \max_{\pi \in \pi_A^*, \pi_B^*} W_\alpha(\mathcal{D}_{\pi,t}, \mathcal{D})$. Then the regret under $R_A$ from executing $\pi_B^*$ optimal for $R_B$ instead of $\pi_A^*$ is at most:*

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq 16 \|R_A\|_2 \left( \frac{\alpha}{1-\gamma} D_{\text{EPIC}}(R_A, R_B) + L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t) \right). \tag{194}$$

*Proof.* The proof for theorem 4.9 holds in the general setting up to eq. 174. Applying lemma A.14 to eq. 174 gives

$$G_{R_A^S}(\pi_A^*) - G_{R_A^S}(\pi_B^*) \leq \frac{2\alpha}{1-\gamma} D_{L^1,\mathcal{D}}(R_A^S, R_B^S) + 4L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t) \tag{195}$$

$$\leq \frac{4\alpha}{1-\gamma} D_{\text{EPIC}}(R_A, R_B) + 4L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t). \tag{196}$$

Applying lemma A.11 yields

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq 16 \|R_A\|_2 \left( \frac{\alpha}{1-\gamma} D_{\text{EPIC}}(R_A, R_B) + L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t) \right), \tag{197}$$

as required. □

### A.7 Limiting Behavior of Regret

The regret bound for finite MDPs, Theorem 4.9, directly implies that, as EPIC distance tends to 0, the regret also tends to 0. By contrast, our regret bound in theorem A.15 for (possibly continuous) MDPs with Lipschitz reward functions includes the relaxed Wasserstein distance $W_\alpha$ as an additive term. At first glance, it might therefore appear possible for the regret to be positive even with a zero EPIC distance. However, in this section we will show that in fact the regret tends to 0 as $D_{\text{EPIC}}(R_A, R_B) \to 0$ in the Lipschitz case as well as the finite case.

We show in lemma A.16 that if the expectation of a non-negative function over a totally bounded measurable metric space $M$ tends to zero under one distribution with adequate support, then it also tends to zero under all other distributions. For example, taking $M$ to be a hypercube in Euclidean space with the Lebesque measure satisfies these assumptions. We conclude in theorem A.17 by showing the regret tends to 0 as the EPIC distance tends to 0.

**Lemma A.16.** *Let $M = (S, d)$ be a totally bounded metric space, where $d(x, y) = \|x - y\|$. Let $(S, A, \mu)$ be a measure space on $S$ with the Borel $\sigma$-algebra $A$ and measure $\mu$. Let $p, q \in \Delta(S)$ be probability density functions on $S$. Let $\delta > 0$ such that $p(s) \geq \delta$ for all $s \in S$. Let $f_n : S \to \mathbb{R}$ be a sequence of L-Lipschitz functions on norm $\|\cdot\|$. Suppose $\lim_{n\to\infty} \mathbb{E}_{X\sim p}[|f_n(X)|] = 0$. Then $\lim_{n\to\infty} \mathbb{E}_{Y\sim q}[|f_n(Y)|] = 0$.*

*Proof.* Since $M$ is totally bounded, for each $r > 0$ there exists a finite collection of open balls in $S$ of radius $r$ whose union contains $M$. Let $B_r(c) = \{s \in S \mid \|s - c\| < r\}$, the open ball of radius $r$ centered at $c$. Let $C(r)$ denote some finite collection of $Q(r)$ open balls:

$$C(r) = \{B_r(c_{r,n}) \mid n \in \{1, \cdots, Q(r)\}\}, \tag{198}$$

such that $\bigcup_{B\in C(r)} B = S$.

It is possible for some balls $B_r(c_n)$ to have measure zero, $\mu(B_r(c_n)) = 0$, such as if $S$ contains an isolated point $c_n$. Define $P(r)$ to be the subset of $C(r)$ with positive measure:

$$P(r) = \{B \in C(r) \mid \mu(B) > 0\}, \tag{199}$$

and let $p_{r,1}, \cdots, p_{r,Q'(r)}$ denote the centers of the balls in $P$. Since $P(r)$ is a finite collection, it must have a minimum measure:

$$\alpha(r) = \min_{B\in P} \mu(B). \tag{200}$$

Moreover, by construction of $P$, $\alpha(r) > 0$.

Let $S'(r)$ be the union only over balls of positive measure:

$$S'(r) = \bigcup_{B\in P(r)} B. \tag{201}$$

Now, let $D(r) = S \setminus S'(r)$, comprising the (finite number of) measure zero balls in $C(r)$. Since measures are countably additive, it follows that $D(r)$ is itself measure zero: $\mu(D(r)) = 0$. Consequently:

$$\int_S g(s)d\mu = \int_{S'(r)} g(s)d\mu, \tag{202}$$

for any measurable function $g : S \to \mathbb{R}$.

Since $\lim_{n\to\infty} \mathbb{E}_{X\sim p}[|f_n(X)|] = 0$, for all $r > 0$ there exists some $N_r \in \mathbb{N}$ such that for all $n \geq N_r$:

$$\mathbb{E}_{X\sim p}[|f_n(X)|] < \delta L r \alpha(r). \tag{203}$$

By Lipschitz continuity, for any $s, s' \in S$:

$$|f_n(s')| \geq |f_n(s)| - L\|s' - s\|. \tag{204}$$

In particular, since any point $s \in S'(r)$ is at most $r$ distance from some ball center $p_{r,i}$, then $|f_n(p_{r,i})| \geq |f_n(s)| - Lr$. So if there exists $s \in S'(r)$ such that $|f_n(s)| \geq 3Lr$, then there must exist a ball center $p_{r,i}$ with $|f_n(p_{r,i})| \geq 2Lr$. Then for any point $s' \in B_r(p_{r,i})$:

$$|f_n(s')| \geq |f_n(p_{r,i})| - Lr \geq Lr. \tag{205}$$

Now, we have:

$$\begin{aligned}
\mathbb{E}_{X\sim p}[|f_n(X)|] &\triangleq \int_{s\in S} |f_n(s)|p(s)d\mu(s) && (206)\\
&= \int_{s\in S'(r)} |f_n(s)|p(s)d\mu(s) && \text{eq. 202} \quad (207)\\
&\geq \int_{s\in B_r(p_{r,i})} |f_n(s)|p(s)d\mu(s) && \text{non-negativity of } |f_n(s)| \quad (208)\\
&\geq \delta \int_{s\in B_r(p_{r,i})} |f_n(s)|d\mu(s) && p(s) \geq \delta \quad (209)\\
&\geq \delta \cdot Lr \int_{s\in B_r(p_{r,i})} 1 d\mu(s) && \text{eq. 205} \quad (210)\\
&= \delta L r \mu(B_r(p_{r,i})) && \text{integrating w.r.t. } \mu \quad (211)\\
&\geq \delta L r \alpha(r) && \alpha(r) \text{ minimum of } \mu(B_r(p_{r,i})). \quad (212)
\end{aligned}$$

But this contradicts eq. 203, and so can only hold if $n < N_r$. It follows that for all $n \geq N_r$ and $s \in S'(r)$, we have $|f_n(s)| < 3Lr$, and so in particular:

$$\mathbb{E}_{Y\sim q}\left[|f_n(Y)|\right] < 3Lr. \tag{213}$$

Let $\epsilon > 0$. Choose $r = \frac{\epsilon}{3L}$. Then for all $n \geq N_r$, $\mathbb{E}_{Y\sim q}\left[|f_n(Y)|\right] < \epsilon$. It follows that:

$$\lim_{n\to\infty} \mathbb{E}_{Y\sim q}\left[|f_n(Y)|\right] = 0, \tag{214}$$

completing the proof. □

**Theorem A.17.** *Let $M$ be an MDP\R with state and action spaces $\mathcal{S}$ and $\mathcal{A}$. Let $R_A, R_B : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be bounded rewards on some norm $\|\cdot\|$ on $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Let $\pi_A^*$ and $\pi_B^*$ be policies optimal for rewards $R_A$ and $R_B$ in $M$. Let $\mathcal{D}_\pi(t, s_t, a_t, s_{t+1})$ denote the distribution over trajectories that policy $\pi$ induces in $M$ at time step $t$. Let $\mathcal{D}(s, a, s')$ be the (stationary) coverage distribution over transitions $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ used to compute $D_{\text{EPIC}}$.*

*Suppose that either:*

1. Discrete*: $\mathcal{S}$ and $\mathcal{A}$ are* discrete*. Moreover, suppose that there exists some $K > 0$ such that $K\mathcal{D}(s_t, a_t, s'_{t+1}) \geq \mathcal{D}_\pi(t, s_t, a_t, s'_{t+1})$ for all time steps $t \in \mathbb{N}$, triples $s_t, a_t, s_{t+1} \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and policies $\pi \in \{\pi_A^*, \pi_B^*\}$.*
2. Lipschitz*: $(\mathcal{S} \times \mathcal{A} \times \mathcal{S}, d)$ is a totally bounded measurable metric space where $d(x, y) = \|x - y\|$. Moreover, $R_A$ and $R_B$ are $L$-*Lipschitz *on $\|\cdot\|$. Furthermore, suppose there exists some $\delta > 0$ such that $\mathcal{D}(s, a, s') \geq \delta$ for all $s, a, s' \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, and that $\mathcal{D}_\pi(t, s_t, a_t, s_{t+1})$ is a non-degenerate probability density function (i.e. no single point has positive measure).*

*Then as $D_{\text{EPIC}}(R_A, R_B) \to 0$, $G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \to 0$.*

*Proof.* In case *(1) Discrete*, by theorem 4.9:

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq \frac{16K\|R_A\|_2}{1-\gamma} D_{\text{EPIC}}(R_A, R_B). \tag{215}$$

Moreover, by optimality of $\pi_A^*$ we have $0 \leq G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*)$. So by the squeeze theorem, as $D_{\text{EPIC}}(R_A, R_B) \to 0$, $G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \to 0$.

From now on, suppose we are in case *(2) Lipschitz*. By the same argument as lemma A.10 up to eq. 145, we have for any policy $\pi$:

$$\left|G_{R_A^S}(\pi) - G_{R_B^S}(\pi)\right| \leq \sum_{t=0}^{\infty} \gamma^t D_{L^1,\mathcal{D}_{\pi,t}}(R_A^S, R_B^S). \tag{216}$$

Applying lemma A.11 we have:

$$|G_{R_A}(\pi) - G_{R_B}(\pi)| \leq 4\,\|R_A\|_2 \sum_{t=0}^{\infty} \gamma^t D_{L^1,\mathcal{D}_{\pi,t}}(R_A^S, R_B^S). \tag{217}$$

By equation 174, we know that $D_{L^1,\mathcal{D}}(R_A^S, R_B^S) \to 0$ as $D_{\text{EPIC}}(R_A, R_B) \to 0$. By lemma A.16, we know that $D_{L^1,\mathcal{D}_{\pi,t}}(R_A^S, R_B^S) \to 0$ as $D_{L^1,\mathcal{D}}(R_A^S, R_B^S) \to 0$. So we can conclude that as $D_{\text{EPIC}}(R_A, R_B) \to 0$:

$$|G_{R_A}(\pi) - G_{R_B}(\pi)| \to 0, \tag{218}$$

as required. □