Quantifying Differences in Reward Functions

Adam Gleave\(^{1\ast}\)  Michael Dennis\(^{2}\)  Shane Legg\(^{2}\)  Stuart Russell\(^{1}\)  Jan Leike\(^{3}\)
\(^{1}\)UC Berkeley  \(^{2}\)DeepMind  \(^{3}\)Independent
gleave@berkeley.edu

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 evaluation environment, which often differs from 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 visitation 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 code is available at [https://github.com/HumanCompatibleAI/evaluating-rewards](https://github.com/HumanCompatibleAI/evaluating-rewards).

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 [11, 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 method also produces _false positives_: many reward functions induce the desired rollout in a given environment but do not 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 \[^3\]) canonicalizes the reward functions’ potential-based shaping, then

\[^\ast\]Work partially conducted during an internship at DeepMind.

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 6.2); and *robust* to the choice of visitation distribution (section 6.3).

<table>
<thead>
<tr>
<th>Distance</th>
<th>Pseudometric</th>
<th>Invariant</th>
<th>Efficient</th>
<th>Predictive</th>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPIC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NPEC</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>ERC</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
</tbody>
</table>

computes the correlation between the canonical rewards over a visitation distribution $D$ 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.

EPIC works best when $D$ 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 $D$, producing similar results across a range of distributions, whereas ERC and especially NPEC are highly sensitive to the choice of $D$ (section 6.3).

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 6.2).

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 6.2. 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 $\mathcal{T}(s' | s, a)$ 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, $g(\tau; R) = \sum_{i=0}^{\left|\tau\right|} \gamma^i R(s_i, a_i, s_{i+1})$, 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 $\pi(a | s)$ 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, $G(\pi) = \mathbb{E}_{\pi(\tau)}[g(\tau; R)]$, where $\pi(\tau)$ is a trajectory generated by sampling the initial state $s_0$ from $\mu$, each action $a_i$ from the policy $\pi(a_i | s_i)$ and successor states $s_{i+1}$ from the transition distribution $\mathcal{T}(s_{i+1} | s_i, a_i)$. 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 $\mathcal{M}$, has a set of optimal policies $\pi$ of an MDP is to find a policy $\pi$ to undiscounted ($\gamma = 1$) MDPs subject to regularity conditions needed for convergence.

**Definition 3.3.** A potential shaping reward is defined as $R(s, a, s') = \gamma \Phi(s') - \Phi(s)$, given a potential $\Phi : \mathcal{S} \rightarrow \mathbb{R}$ 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 $\Phi : \mathcal{S} \rightarrow \mathbb{R}$ such that for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$:

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

Note $R_A - R_B \equiv \text{Zero}$ (where 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 $R_A, R_B, R_C : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ 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, $(R_A \equiv R_B) \land (R_B \equiv R_C)$ implies $R_A \equiv R_C$.

**Proof.** See section [A.3.1] in supplementary material. \qed

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 $\gamma^T \Phi(s_T) \rightarrow 0$ as $T \rightarrow \infty$, 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.

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 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, \mu, \mathcal{T})$, we have $\pi^*(M, R_A) = \pi^*(M, R_B)$, 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

\footnotetext{\footnotesize{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 $\gamma = 1$.}}
We define the canonically shaped reward $C_{\mathcal{D}_S,\mathcal{D}_A} (R)$ as an expectation over some arbitrary distributions $\mathcal{D}_S$ and $\mathcal{D}_A$ over states $S$ and actions $A$ respectively. This construction means $C_{\mathcal{D}_S,\mathcal{D}_A} (R)$ only depends on $(S, A, \gamma)$, and not on the initial state distribution $\mu$ or transition dynamics $T$. In particular, we may evaluate $\tilde{R}$ on transitions that are impossible in the training environment, since these may become possible in a deployment environment with a different $\mu$ or $T$.

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

$$C_{\mathcal{D}_S,\mathcal{D}_A} (R) (s, a, s') = R(s, a, s') + \mathbb{E} [\gamma R(s', A, S') - R(s, A, S') - \gamma R(S, A, S')] .$$

Informally, if $R'$ is shaped by potential $\Phi$, then increasing $\Phi(s)$ decreases $R'(s, a, s')$ but increases $\mathbb{E} [-R'(s, A, S')]$, canceling. Similarly, increasing $\Phi(s')$ increases $R'(s, a, s')$ but decreases $\mathbb{E} [\gamma R'(s', A, S')]$. 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 $R : S \times A \times S \rightarrow \mathbb{R}$ be a reward function and $\Phi : S \rightarrow \mathbb{R}$ a potential function. Let $\gamma \in [0, 1]$ be a discount rate, and $\mathcal{D}_S$ and $\mathcal{D}_A$ be distributions over states $S$ and $A$ respectively. Let $R'$ denote $R$ shaped by $\Phi$: $R'(s, a, s') \equiv R(s, a, s') + \gamma \Phi(s') - \Phi(s)$. Then the canonically shaped $R'$ and $R$ are equal:

$$C_{\mathcal{D}_S,\mathcal{D}_A} (R') = C_{\mathcal{D}_S,\mathcal{D}_A} (R) .$$

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

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

$$\| \Phi_{\mathcal{D}_S,\mathcal{D}_A} (R + \nu) - \Phi_{\mathcal{D}_S,\mathcal{D}_A} (R) \|_{\infty} = \lambda (1 + \gamma \mathcal{D}_S(x)) \mathcal{D}_A(u) \mathcal{D}_S(x') .$$

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 $D_P(X, Y) = \frac{1}{1/\gamma^2} \sqrt{1 - \rho(X, Y)}$, where $\rho(X, Y)$ is the Pearson correlation between $X$ and $Y$.

**Lemma 4.5.** The Pearson distance $D_P$ 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 $0 \leq D_P(aX + c, bY + d) = D_P(X, Y) \leq 1$.

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

$$D_{\text{EPIC}} (R_A, R_B) = D_P (C_{\mathcal{D}_S,\mathcal{D}_A} (R_A) (S, A, S'), C_{\mathcal{D}_S,\mathcal{D}_A} (R_B) (S, A, S')) .$$

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

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

4
We generalize the regret bound to continuous spaces in theorem A.14 via a Lipschitz assumption.

As a pedagogical example, we compute the EPIC distance between the reward functions in figure 1 which reward function you evaluate on.

Theorem 4.9. Let $M$ be a $\gamma$-discounted MDPR with discrete state and action spaces $S$ and $A$. Let $R_A, R_B: S \times A \times S \to \mathbb{R}$ be bounded rewards, and $\pi^*_A, \pi^*_B$ be respective optimal policies. Let $D_\pi(t, s, a, t+1)$ denote the distribution over transitions $S \times A \times S$ induced by policy $\pi$ at time $t$, and $D(s, a, s')$ be the visitation distribution used to compute $D_{EPIC}$. Suppose there exists $K > 0$ such that $K D(s, a, s') \geq D_\pi(t, s, a, t+1)$ for all times $t \in \mathbb{N}$, triples $(s, a, s') \in S \times A \times S$ 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 

$$G_{R_A}(\pi^*_A) - G_{R_A}(\pi^*_B) \leq 16K||R_A||_2 (1 - \gamma)^{-1} D_{EPIC}(R_A, R_B).$$

The key assumption is that the visitation distribution $D$ spends at least $1/K$th as much time at each transition as the rollouts of $\pi^*_A$ and $\pi^*_B$. In finite cases, a uniform $D$ guarantees $K \leq |S|^2 |A|$.

We generalize the regret bound to continuous spaces in theorem A.14 via a Lipschitz assumption. Importantly, the returns of $\pi^*_A$ and $\pi^*_B$ converge as $D_{EPIC}(R_A, R_B) \to 0$ in both cases, no matter which reward function you evaluate on.

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, $D_{EPIC}(Path, Cliff) = 0.27$, almost as much as $D_{EPIC}(Sparse, Cliff) = 0.37$. 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 $D_S$ and $D_A$ uniform over $S$ and $A$, and visitation distribution $D$ uniform over state-action pairs $(s, a)$, with $s'$ deterministically computed. It is important these distributions have adequate support. As an extreme example, if $D_S$ and $D$ 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.
Theorem 5.5. Let \( D \) be some distribution over trajectories. Let \( E \) be a random variable sampled from \( D \). The Episode Return Correlation distance between reward functions \( R_A \) and \( R_B \) is the Pearson distance between their episode returns on \( D \),

\[
D_{\text{ERC}}(R_A, R_B) = D_\rho(g(E; R_A), g(E; R_B)).
\]

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 \( \tau = (s_0, a_0, \ldots, s_T) \) 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 \( D \). 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 \( \rho(g(E; R_A'), g(E; R_B')) \) 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 \( D_{\text{ERC}}(R_A', R_B') \to 0 \) as \( c \to \infty \). By an analogous argument, setting \( \Phi(s_A) = c \) and \( \Phi(s_B) = -c \) gives \( D_{\text{ERC}}(R_A', R_B') \to 1 \) 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 (\( \ell^p \) distance). Let \( D \) be a visitation distribution over transitions \( s \xrightarrow{a} s' \) and let \( p \geq 1 \) be a power. The \( \ell^p \) distance between reward functions \( R_A \) and \( R_B \) is the \( \ell^p \) norm of their difference:

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

Proposition 5.3. (1) \( D_{\ell^p, D} \) is a pseudometric in \( \ell^p \) space. (2) It is a metric in \( \ell^p \) space when functions \( f \) and \( g \) are identified if \( f = g \) almost everywhere on \( D \).

The \( \ell^p \) 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:

\[
D^{U}_{\text{NPEC}}(R_A, R_B) = \inf_{R_A' \equiv A} D_{\ell^p, D}(R_A', R_B).
\]

Unfortunately, \( D^{U}_{\text{NPEC}} \) is sensitive to \( R_B \)'s scale.

It is tempting to instead take the infimum over both arguments of \( D_{\ell^p, D} \). However, \( \inf_{R_A' \equiv A, R_B' \equiv B} D_{\ell^p, D}(R_A', R_B') = 0 \) since all equivalence classes come arbitrarily close to the origin in \( \ell^p \) space. Instead, we fix this by normalizing \( D^{U}_{\text{NPEC}} \).

Definition 5.4. The Nearest Point in Equivalence Class (NPEC) premetric is defined by:

\[
D_{\text{NPEC}}(R_A, R_B) = \frac{D^{U}_{\text{NPEC}}(R_A, R_B)}{D^{U}_{\text{NPEC}}(\text{Zero}, R_B)} \text{ when } D^{U}_{\text{NPEC}}(\text{Zero}, R_B) \neq 0 \text{ and 0 otherwise.}
\]

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

Theorem 5.5. \( D_{\text{NPEC}} \) is a premetric. Moreover, 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 \( 0 \leq D_{\text{NPEC}}(R_A', R_B') = D_{\text{NPEC}}(R_A, R_B) \leq 1 \).

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 visitation distribution $D$, whereas ERC and especially NPEC are highly sensitive to the choice of $D$.

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 (S) or exclude (D) a quadratic control penalty $\|a\|^2$. 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: $D_{\text{NPEC}}(D, S) = 0.58$. 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. Unfortunately, $D_{\text{ERC}}(D, S) = 0.56$ due to ERC’s erroneous handling of stochastic initial states.

6.2 Predicting policy performance from reward distance

We train reward models on the PointMaze task from Fu et al. [8], and evaluate the ground-truth (GT) return of a policy optimized for the learned reward. Table 2 shows that rewards with low distance from GT achieve high returns. High distance rewards sometimes work but are sensitive to dynamics. PointMaze is a MuJoCo environment where a point mass agent must navigate around a wall to reach a goal. The train and test variants differ only in the position of the wall. 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.

Both Regression 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...
We find EPIC is robust to varying $\mathcal{D}$. We would like the reward distances to be robust to the exact choice of visitation distribution $\mathcal{D}$, without affecting the distance. In general, any black-box method for assessing reward models – including the rollout method – only has predictive power on transitions visited during testing. Nonetheless, even with EPIC some care must be taken when choosing $\mathcal{D}$. Typically, $\mathcal{D}$ is collected via rollouts of some exploration policy in an environment. This works well when the deployment environment has a similar set of reachable states to the rollout environment, even if some details of the dynamics – such as the position of the wall in PointMaze – differ. However, when the deployment environment allows a transition $(s,a,s')$ that is not physically attainable in the rollout environment, then $\mathcal{D}$ will place no support on this transition and the reward $R(s,a,s')$ can take arbitrary values without affecting the distance. In general, any black-box method for assessing reward models – including the rollout method – only has predictive power on transitions visited during testing.

Table 2: Distances of reward models from ground-truth ($\text{GT}$), and the mean $\text{GT}$ 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 visitation distribution $\mathcal{D}$ from rollouts in the train environment of: a uniform random policy $\pi_{\text{unif}}$, an expert $\pi^*$ and a mixture of these policies. $\mathcal{D}_S$ and $\mathcal{D}_A$ are computed by marginalizing $\mathcal{D}$. 95% confidence intervals (see Table A.6) are tighter than $\pm 1\%$ for EPIC and ERC but are as large as $\pm 50\%$ for NPEC due to high variance across seeds.

<table>
<thead>
<tr>
<th>Reward</th>
<th>1000 $\times$ $D_{\text{EPIC}}$</th>
<th>1000 $\times$ $D_{\text{NPEC}}$</th>
<th>1000 $\times$ $D_{\text{ERC}}$</th>
<th>Episode Return</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>Gen. Train Test</td>
</tr>
<tr>
<td>$\text{GT}$</td>
<td>0 0 0 0 0 0 0 0 0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>$\text{Reg}$</td>
<td>41.9 36.5 25.9 0.519 14.9 0.140 4.78 40.9 1.39</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>$\text{Pref}$</td>
<td>50.5 54.4 32.9 2.99 204 1.78 15.0 180 8.15</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>AIRL SO</td>
<td>488 600 395 684 3550 426 448 382 234</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>AIRL SA</td>
<td>548 614 390 823 3030 376 506 467 208</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

6.3 Sensitivity of reward distance to visitation state distribution

We would like the reward distances to be robust to the exact choice of visitation distribution $\mathcal{D}$. In Table 2, we report distances calculated under distributions induced by rollouts from three different policies. $\pi_{\text{unif}}$ 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, $\mathcal{D}_S$ and $\mathcal{D}_A$ are marginalized from $\mathcal{D}$ and so also vary between conditions.

We find EPIC is robust to varying $\mathcal{D}$: the distance varies by less than 2$\times$, and the ranking between the reward models is the same across visitation distributions, except for Mix favoring AIRL SA over AIRL SO. By contrast, NPEC is highly sensitive to the choice of $\mathcal{D}$: the distance of Pref varies by over 500$\times$ between $\pi_{\text{unif}}$ and $\pi^*$. ERC lies somewhere in the middle: the distances vary by as much as 25$\times$. Overall, EPIC is clearly the least sensitive to choice of $\mathcal{D}$ in this environment.

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


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 visitation distribution $D$, and a batch $B_M$ of $N_M$ samples from the joint state and action distributions $D_S \times D_A$. For each $(s, a, s') \in B_V$, we approximate the canonically shaped rewards (definition 4.1) by taking the mean over $B_M$:

$$C_{D_S, D_A}(R)(s, a, s') = R(s, a, s') + \mathbb{E}[\gamma R(s', A, S') - R(s, A, S') - \gamma R(S, A, S')]$$

$$\approx R(s, a, s') + \frac{1}{N_M} \sum_{(x, u) \in B_M} R(s', u, x) - \frac{1}{N_M} \sum_{(x, u) \in B_M} R(s, u, x) - c.$$

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 $c = \frac{1}{N_M} \sum_{(x, u) \in B_M} \sum_{(x', u') \in B_M} R(x, u, x')$. 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:

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

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 $J(\nu, c, w) = D(R_{\nu, c, w}(s, a, s'), R_B(s, a, s'))$ on samples $(s, a, s')$ from a visitation distribution $D$. Note the mean cost 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:

$$\arg \min_{\lambda \geq 0, c \in \mathbb{R}, s, a, s' \sim D} \mathbb{E}[(\lambda R_A(s, a, s') + c - R_B(s, a, s'))^2].$$

We use the active set method of Lawson and Hanson [11] to solve this constrained least-squares problem. These initial affine parameters minimize the $L^p$ distance $D_{L^p, D}(R_{\nu, c, 0}(s, a, s'), R_B(s, a, s'))$ under the metric $\ell(x, y) = (x - y)^2$ 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 approximation. 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 cheap, so we instead apply bootstrapping to the collected episodes, computing the ERC distance for each bootstrapped episode sample.

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 visitation
distribution of uniform random transitions \( D_{\text{unif}} \) in the simple GridWorld environment with known deterministic dynamics. In other environments, the visitation distribution is sampled from rollouts of a policy. We use a random policy \( \pi_{\text{unif}} \) for PointMass, HalfCheetah and Hopper in the hand-designed reward experiments (section 6.1). In PointMaze, we compare three visitation distributions (section 6.3) induced by rollouts of \( \pi_{\text{unif}} \), an expert policy \( \pi^* \) and a \( \text{Mix} \)ture of the two policies, sampling actions from either \( \pi_{\text{unif}} \) or \( \pi^* \) and switching between them with probability 0.05 per timestep.

### A.2.2 Training Learned Reward Models

For the experiments on learned reward functions (sections 6.2 and 6.3), 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 AIRM, we use an existing open-source implementation [23]. We developed new implementations for preference comparison and regression, available at [https://github.com/HumanCompatibleAI/evaluating-rewards](https://github.com/HumanCompatibleAI/evaluating-rewards). We also use the RL algorithm proximal policy optimization (PPO; [18]) on the ground-truth reward to train expert policies to provide demonstrations for AIRM, and on learned reward models to evaluate their performance.

For PPO and AIRM 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 timesteps 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.

---

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>In experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visitation Distribution</td>
<td>( D_{\text{unif}} )</td>
<td>GridWorld</td>
</tr>
<tr>
<td></td>
<td>Rollouts from ( \pi_{\text{unif}} )</td>
<td>PointMass, HalfCheetah, Hopper</td>
</tr>
<tr>
<td></td>
<td>( \pi_{\text{unif}}, \pi^* ) and Mixturer</td>
<td>PointMaze</td>
</tr>
<tr>
<td>Bootstrap Samples</td>
<td>10,000</td>
<td>All</td>
</tr>
<tr>
<td>Discount ( \gamma )</td>
<td>0.99</td>
<td>All</td>
</tr>
<tr>
<td>EPIC</td>
<td>State Distribution ( D_S )</td>
<td>PointMaze</td>
</tr>
<tr>
<td></td>
<td>( N(0, 1) ) standard Gaussian</td>
<td>PointMaze</td>
</tr>
<tr>
<td></td>
<td>Marginalized from ( D )</td>
<td>PointMaze</td>
</tr>
<tr>
<td>Action Distribution ( D_A )</td>
<td>( U[-1, 1] ) continuous uniform</td>
<td>PointMaze</td>
</tr>
<tr>
<td></td>
<td>Marginalized from ( D )</td>
<td>PointMaze</td>
</tr>
<tr>
<td>Seeds</td>
<td>30</td>
<td>All</td>
</tr>
<tr>
<td>Samples ( N_V )</td>
<td>32,768</td>
<td>All</td>
</tr>
<tr>
<td>Mean Samples ( N_M )</td>
<td>32,768</td>
<td>All</td>
</tr>
<tr>
<td>NPEC</td>
<td>Seeds</td>
<td>All</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>All</td>
</tr>
<tr>
<td>Total Timesteps</td>
<td>( 1 \times 10^6 )</td>
<td>All</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
<td>All</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>( 1 \times 10^{-2} )</td>
<td>All</td>
</tr>
<tr>
<td>Batch Size</td>
<td>4096</td>
<td>All</td>
</tr>
<tr>
<td>Warmstart Size</td>
<td>16,386</td>
<td>All</td>
</tr>
<tr>
<td>Loss ( \ell )</td>
<td>( \ell(x, y) = (x - y)^2 )</td>
<td>All</td>
</tr>
<tr>
<td>ERC</td>
<td>Episodes</td>
<td>131,072</td>
</tr>
</tbody>
</table>

We also use the RL algorithm proximal policy optimization (PPO; [18]) on the ground-truth reward to train expert policies to provide demonstrations for AIRM, and on learned reward models to evaluate their performance.

For PPO and AIRM 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.
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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>In environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Timesteps</td>
<td>$1 \times 10^6$</td>
<td>All</td>
</tr>
<tr>
<td>Batch Size</td>
<td>16384</td>
<td>PointMaze</td>
</tr>
<tr>
<td>Discount $\gamma$</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Entropy Coefficient</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>Learning Rate</td>
<td>$2.5 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>Value Function Coefficient</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Gradient Clipping Threshold</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Ratio Clipping Threshold</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Lambda (GAE)</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Minibatches</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Optimization Epochs</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Parallel Environments</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL Algorithm</td>
<td>PPO [18]</td>
</tr>
<tr>
<td>Total Timesteps</td>
<td>102400</td>
</tr>
<tr>
<td>Discount $\gamma$</td>
<td>0.99</td>
</tr>
<tr>
<td>Demonstration Timesteps</td>
<td>100000</td>
</tr>
<tr>
<td>Generator Batch Size</td>
<td>2048</td>
</tr>
<tr>
<td>Discriminator Batch Size</td>
<td>50</td>
</tr>
<tr>
<td>Entropy Weight</td>
<td>1.0</td>
</tr>
<tr>
<td>Reward Function Architecture</td>
<td>MLP, two 32-unit hidden layers</td>
</tr>
<tr>
<td>Potential Function Architecture</td>
<td>MLP, two 32-unit hidden layers</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Range Tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Timesteps</td>
<td>$5 \times 10^6$</td>
<td>$[1, 10 \times 10^6]$</td>
</tr>
<tr>
<td>Batch Size</td>
<td>10000</td>
<td>$[500, 250000]$</td>
</tr>
<tr>
<td>Trajectory Length</td>
<td>5</td>
<td>$[1, 100]$</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>$1 \times 10^{-2}$</td>
<td>$[1 \times 10^{-4}, 1 \times 10^{-1}]$</td>
</tr>
<tr>
<td>Discount $\gamma$</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Reward Function Architecture</td>
<td>MLP, two 32-unit hidden layers</td>
<td></td>
</tr>
<tr>
<td>Output L2 Regularization Weight</td>
<td>$1 \times 10^{-4}$</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Range Tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Timesteps</td>
<td>$10 \times 10^6$</td>
<td>$[1, 20 \times 10^6]$</td>
</tr>
<tr>
<td>Batch Size</td>
<td>4096</td>
<td>$[256, 16384]$</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>$2 \times 10^{-2}$</td>
<td>$[1 \times 10^{-3}, 1 \times 10^{-1}]$</td>
</tr>
<tr>
<td>Discount $\gamma$</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Reward Function Architecture</td>
<td>MLP, two 32-unit hidden layers</td>
<td></td>
</tr>
</tbody>
</table>
at convergence. Finally, for regression we performed a sweep over batch size, learning rate and total timesteps 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 \(10 \times 10^6\) timesteps 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 timesteps from an expert policy trained on the ground-truth reward, similar in size to the total number of timesteps AIRL is trained for (see Table A.3). In preference comparison and regression, each batch is sampled afresh from the visitation distribution specified in Table A.1 and labeled according to the ground-truth reward.

### A.2.3 Computing infrastructure

Experiments were conducted on a small number of \texttt{n1-standard-96} Google Cloud Platform VM instances, with 48 CPU cores on an Intel Skylake processor and 360 GB of RAM. It takes less than a week of compute on a single \texttt{n1-standard-96} 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 \(S \rightarrow \text{D}\). 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.6. 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 visitation 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.
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.
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.
Figure A.3: Approximate distances between hand-designed reward functions in PointMass. The visitation distribution $D$ is sampled from rollouts of a policy $\pi_{\text{unif}}$ taking actions uniformly at random. **Key:** $\bullet$ quadratic control penalty, $\square$ no control penalty. $S$ is Sparse$(x) = 1[|x| < 0.05]$, $D$ is shaped $\text{Denise}(x, x') = \text{Sparse}(x) + |x'| - |x|$, while $M$ is $\text{Magnitude}(x) = -|x|$. **Confidence Interval (CI):** 95% CI computed by bootstrapping over 10 000 samples.

Figure A.4: Approximate distances between hand-designed reward functions in HalfCheetah. The visitation distribution $D$ is sampled from rollouts of a policy $\pi_{\text{unif}}$ taking actions uniformly at random. **Key:** $\bullet$ a reward proportional to the change in center of mass and moving forward is rewarded when $\bullet$ to the right, and moving backward is rewarded when $\bullet$ to the left, $\square$ quadratic control penalty, $\square$ no control penalty. **Confidence Interval (CI):** 95% CI computed by bootstrapping over 10 000 samples.
Figure A.5: Approximate distances between hand-designed reward functions in Hopper. The visitation distribution $\bar{D}$ is sampled from rollouts of a policy $\pi_{\text{unif}}$ taking actions uniformly at random. Key: $\uparrow$ is a reward proportional to the change in center of mass and $\downarrow$ is the backflip reward defined in Amodei et al. [2, footnote]. Moving forward is rewarded when $\uparrow$ or $\downarrow$ is to the right, and moving backward is rewarded when $\uparrow$ or $\downarrow$ is to the left. $\sqsubseteq$ quadratic control penalty. $\nabla$ no control penalty. **Confidence Interval (CI):** 95% CI computed by bootstrapping over 10,000 samples.
Table A.6: Approximate distances of learned reward models 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× scale) use visitation distribution $D$ from rollouts in the train environment of: a uniform random policy $\pi_{\text{unif}}$, an expert $\pi^*$ and a mixture of these policies. $D_S$ and $D_A$ are computed by marginalizing $D$.

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

<table>
<thead>
<tr>
<th>Reward</th>
<th>$1000 \times D_{\text{LOW}}^{\text{EPIC}}$</th>
<th>$1000 \times D_{\text{LOW}}^{\text{NPEC}}$</th>
<th>$1000 \times D_{\text{LOW}}^{\text{ERC}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
</tr>
<tr>
<td>Regress</td>
<td>41.7 36.4 25.9</td>
<td>0.506 13.3 0.126</td>
<td>4.75 40.7 1.38</td>
</tr>
<tr>
<td>Pref</td>
<td>50.2 54.3 32.7</td>
<td>2.80 159 1.76</td>
<td>15.0 179 8.11</td>
</tr>
<tr>
<td>AIRL SO</td>
<td>484 599 393</td>
<td>673 2640 417</td>
<td>446 380 232</td>
</tr>
<tr>
<td>AIRL SA</td>
<td>544 614 388</td>
<td>804 1630 370</td>
<td>505 465 206</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Reward</th>
<th>$1000 \times \bar{D}_{\text{EPIC}}$</th>
<th>$1000 \times \bar{D}_{\text{NPEC}}$</th>
<th>$1000 \times \bar{D}_{\text{ERC}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
</tr>
<tr>
<td>Regress</td>
<td>41.9 36.5 25.9</td>
<td>0.519 14.9 0.140</td>
<td>4.78 40.9 1.39</td>
</tr>
<tr>
<td>Pref</td>
<td>50.5 54.4 32.9</td>
<td>2.90 204 1.78</td>
<td>15.0 180 8.15</td>
</tr>
<tr>
<td>AIRL SO</td>
<td>488 600 395</td>
<td>684 3550 426</td>
<td>448 382 234</td>
</tr>
<tr>
<td>AIRL SA</td>
<td>548 614 390</td>
<td>823 3030 376</td>
<td>506 467 208</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Reward</th>
<th>$1000 \times D_{\text{UP}}^{\text{EPIC}}$</th>
<th>$1000 \times D_{\text{UP}}^{\text{NPEC}}$</th>
<th>$1000 \times D_{\text{UP}}^{\text{ERC}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
</tr>
<tr>
<td>Regress</td>
<td>42.2 36.5 26.0</td>
<td>0.535 16.8 0.162</td>
<td>4.80 41.1 1.40</td>
</tr>
<tr>
<td>Pref</td>
<td>50.9 54.3 33</td>
<td>3.16 240 1.80</td>
<td>15.1 181 8.19</td>
</tr>
<tr>
<td>AIRL SO</td>
<td>492 601 397</td>
<td>694 4420 436</td>
<td>450 384 235</td>
</tr>
<tr>
<td>AIRL SA</td>
<td>552 614 392</td>
<td>848 4660 385</td>
<td>508 469 209</td>
</tr>
</tbody>
</table>

(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.6b with 95% probability.

<table>
<thead>
<tr>
<th>Reward</th>
<th>$D_{\text{REL}}^{\text{EPIC}}^{%}$</th>
<th>$D_{\text{REL}}^{\text{NPEC}}^{%}$</th>
<th>$D_{\text{REL}}^{\text{ERC}}^{%}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
<td>$\pi_{\text{unif}}$ $\pi^*$ Mix</td>
</tr>
<tr>
<td>Regress</td>
<td>0.662 0.0950 0.352</td>
<td>3.04 12.9 16.0</td>
<td>0.589 0.544 0.620</td>
</tr>
<tr>
<td>Pref</td>
<td>0.683 0.158 0.411</td>
<td>6.31 21.8 1.41</td>
<td>0.499 0.538 0.481</td>
</tr>
<tr>
<td>AIRL SO</td>
<td>0.875 0.115 0.522</td>
<td>1.60 25.8 2.37</td>
<td>0.449 0.504 0.621</td>
</tr>
<tr>
<td>AIRL SA</td>
<td>0.654 0.0331 0.397</td>
<td>3.15 53.9 2.34</td>
<td>0.382 0.445 0.540</td>
</tr>
</tbody>
</table>
A.3 Proofs

A.3.1 Background

**Proposition 4.3.** Let \(\lambda \colon S \times A \times S \to \mathbb{R}\) be bounded reward functions. Then \(\lambda\) is reflexive, symmetric; \(\lambda\) implies \(R_B \equiv R_A\); and transitive, \(\lambda \equiv R_A; R_B\) implies \(R_B \equiv R_A\).

**Proof.** \(\lambda\) since \(\lambda(s,a,s') = \lambda A(s,a,s') + \gamma \Phi(s') - \Phi(s)\) for all \(s,s' \in S\) and \(a \in A\).

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

Rearranging, we have:

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

Since \(\frac{1}{\lambda} > 0\) and \(\Phi'(s) = -\frac{1}{\lambda} \Phi(s)\) is bounded potential functions, 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 \(\Phi_1, \Phi_2 : S \to \mathbb{R}\) such that for all \(s,s' \in S\) and \(a \in A\):

\[
R_B(s,a,s') = \lambda_1 R_A(s,a,s') + \gamma \Phi_1(s') - \Phi_1(s)\]
\[
R_C(s,a,s') = \lambda_2 R_B(s,a,s') + \gamma \Phi_2(s') - \Phi_2(s)\]

Substituting the expression for \(R_B\) into the expression for \(R_C\):

\[
R_C(s,a,s') = \lambda_2 (\lambda_1 R_A(s,a,s') + \gamma \Phi_1(s') - \Phi_1(s)) + \gamma \Phi_2(s') - \Phi_2(s)
\]

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 4.4.** Let \(R : S \times A \times S \to \mathbb{R}\) be a reward function and \(\Phi : S \to \mathbb{R}\) be a bounded potential function. Let \(\gamma \in [0, 1]\) be a discount rate, and \(D_S\) and \(D_A\) be distributions over states \(S\) and \(A\), respectively. Let \(R' = \Phi(\gamma R(s,a,s')) = R(s,a,s') + \gamma \Phi(s') - \Phi(s)\). Then the canonically shaped \(R'\) and \(R\) are equal:

\[
C_{D_S,D_A}(R') = C_{D_S,D_A}(R).
\]

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

\[
C_{D_S,D_A}(R') (s,a,s') = \mathbb{E} [\gamma R'(s',A',S') - R'(s,A',S') - \gamma R(S,A,S')] = \mathbb{E} [\gamma R(s',A',S') - R(s,A,S') - \gamma R(S,A,S')] = \mathbb{E} [\gamma R(s',A',S') - R(s,A,S') - \gamma R(S,A,S')] = \mathbb{E} [\gamma R(s',A',S') - R(s,A,S') - \gamma R(S,A,S')]
\]

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

**Proposition 4.5.** Let \(S\) and \(A\) be discrete state and action spaces, with \(|S| \geq 2\). Let \(R, v : S \times A \times S \to \mathbb{R}\) be reward functions, with \(v(s,a,s') = \lambda \delta[(s,a,s') = (x,u,x')]\), \(\lambda \in \mathbb{R}\), \(x, x' \in S\) and \(u \in A\). Let \(\Phi_{D_S,D_A}(R)(s,a,s') = C_{D_S,D_A}(R)(s,a,s') - R(s,a,s')\). Then:

\[
\|\Phi_{D_S,D_A}(R + \nu) - \Phi_{D_S,D_A}(R)\|_{\infty} = \lambda (1 + \gamma D_S(x)) D_A(u) D_S(x').
\]
Proof. Observe that:

\[ \Phi_{D_S, D_A}(R)(s, a, s') = \mathbb{E} [\gamma R(s', A, S') - R(s, A, S') - \gamma R(S, A, S')] , \]

where \( S \) and \( S' \) are random variables independently sampled from \( D_S \) and \( A \) sampled from \( D_A \).

Then:

\[ \Phi_{D_S, D_A}(R + \nu) - \Phi_{D_S, D_A}(R) = \Phi_{D_S, D_A}(\nu). \]

Now

\[ \| \Phi_{D_S, D_A}(R + \nu) - \Phi_{D_S, D_A}(R) \|_\infty = \max_{s, s' \in S} | \mathbb{E} [\gamma \nu(s', A, S') - \nu(s, A, S') - \gamma \nu(S, A, S')] | \]

\[ = \max_{s, s' \in S} | \lambda (\gamma I[x = s'] D_A(u) D_S(x')) - \lambda \gamma D_S(x) D_A(u) D_S(x') | \]

\[ = \max_{s, s' \in S} | |x = s| D_A(u) D_S(x') - \gamma D_S(x) D_A(u) D_S(x') | \]

\[ = \lambda (1 + \gamma D_S(x)) D_A(u) D_S(x'), \]

where the final step follows by substituting \( s = x \) and \( s' \neq x \). \( \square \)

**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 random variable \( V \), define a standardized (zero mean and variance) version:

\[ \hat{V} = \frac{V - \mathbb{E}[V]}{\sqrt{\mathbb{E}[(V - \mathbb{E}[V])^2]}} \]

The Pearson correlation coefficient on random variables \( X \) and \( Y \) is equal to the expected product of these standardized random variables:

\[ \rho(X, Y) = \mathbb{E} [\hat{X} \hat{Y}] . \]

Let \( X, Y \) and \( Z \) 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 \( A, B \):

\[ \mathbb{E} \left[ (\hat{A} - \hat{B})^2 \right] = \mathbb{E} [\hat{A}^2 - 2 \hat{A} \hat{B} + \hat{B}^2 ] \]

\[ = \mathbb{E} [\hat{A}^2 - 2 \hat{A} \hat{B} + \hat{B}^2 ] \]

\[ = \mathbb{E} [\hat{A}^2] + \mathbb{E} [\hat{B}^2] - 2 \mathbb{E} [\hat{A} \hat{B}] \]

\[ = 2 - 2 \mathbb{E} [\hat{A} \hat{B}] \]

\[ = 2 (1 - \rho(A, B)) \]

\[ = 4 D_\rho(A, B)^2 . \]

So:

\[ 4 D_\rho(X, Z) = \mathbb{E} [\hat{X} - \hat{Z}]^2 \]

\[ = \mathbb{E} [\hat{X} - \hat{Y} + \hat{Y} - \hat{Z}]^2 \]

\[ = \mathbb{E} [\hat{X} - \hat{Y}]^2 + \mathbb{E} [\hat{Y} - \hat{Z}]^2 + 2 \mathbb{E} [\hat{X} - \hat{Y}] [\hat{Y} - \hat{Z}] \]

\[ = 4 D_\rho(X, Y)^2 + 4 D_\rho(Y, Z)^2 + 8 \mathbb{E} [\hat{X} - \hat{Y}] [\hat{Y} - \hat{Z}] . \]
Theorem 4.8. Let \( D_{\rho}(X, Z)^2 \leq D_{\rho}(X, Y)^2 + D_{\rho}(Y, Z)^2 + 2D_{\rho}(X, Y)D_{\rho}(Y, Z) \)
\[ = (D_{\rho}(X, Y) + D_{\rho}(Y, Z))^2. \]
Taking the square root of both sides:
\[ D_{\rho}(X, Z) \leq D_{\rho}(X, Y) + D_{\rho}(Y, Z), \]
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 \( S \times A \times S \) to real numbers \( \mathbb{R} \).

Identity. Have:
\[ D_{\text{EPIC}}(R_A, R_A) = D_{\rho}(C_{D_{S},D_{A}}(R_A) (S, A, S'), C_{D_{S},D_{A}}(R_A) (S, A, S')) = 0, \]
since \( D_{\rho}(X, X) = 0. \)

Symmetry. Have:
\[ D_{\text{EPIC}}(R_A, R_B) = D_{\rho}(C_{D_{S},D_{A}}(R_A) (S, A, S'), C_{D_{S},D_{A}}(R_B) (S, A, S')) \]
\[ = D_{\rho}(C_{D_{S},D_{A}}(R_B) (S, A, S'), C_{D_{S},D_{A}}(R_A) (S, A, S')) \]
\[ = D_{\text{EPIC}}(R_B, R_A), \]
since \( D_{\rho}(X, Y) = D_{\rho}(Y, X). \)

Triangle Inequality. Have:
\[ D_{\text{EPIC}}(R_A, R_C) = D_{\rho}(C_{D_{S},D_{A}}(R_A) (S, A, S'), C_{D_{S},D_{A}}(R_C) (S, A, S')) \]
\[ \leq D_{\rho}(C_{D_{S},D_{A}}(R_A) (S, A, S'), C_{D_{S},D_{A}}(R_B) (S, A, S')) \]
\[ + D_{\rho}(C_{D_{S},D_{A}}(R_B) (S, A, S'), C_{D_{S},D_{A}}(R_C) (S, A, S')) \]
\[ = D_{\text{EPIC}}(R_A, R_B) + D_{\text{EPIC}}(R_B, R_C), \]
since \( D_{\rho}(X, Z) \leq D_{\rho}(X, Y) + D_{\rho}(Y, Z). \)

Theorem 4.8. Let \( R_A, R_A', R_B, R_B' : S \times A \times S \to \mathbb{R} \) 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 \( X \in \{ A, B \} \):
\[ R_X'(s, a, s') = R_X'(s, a, s') + \gamma \Phi_X(s') - \Phi_X(s) \]
\[ R_X(s, a, s') = \lambda_X R_X(s, a, s') \]
for some scaling factor \( \lambda_X > 0 \) and potential function \( \Phi_X : S \to \mathbb{R}. \)

By proposition 4.2
\[ C_{D_{S},D_{A}}(R_X') = C_{D_{S},D_{A}}(R_X'), \]
Moreover, since \( C_{D_{S},D_{A}}(R) \) is defined as an expectation over \( R \) and expectations are linear:
\[ C_{D_{S},D_{A}}(R_X) = \lambda_X C_{D_{S},D_{A}}(R_X). \]
Unrolling the definition of $D_{\text{EPIC}}$ and applying this result gives:

$$D_{\text{EPIC}}(R'_A, R'_B) = \lambda (C_{D_{S}} (R'_A (S, A, S'), C_{D_{S}} (R'_B (S, A, S'))) = \lambda (\Phi (R_A (S, A, S'), \Phi (R_B (S, A, S')) = \lambda (D_{\text{EPIC}} (R_A, R_B). \quad \square$$

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

**Proposition 5.3.** (1) $D_{L^p, \mathcal{D}}$ is a pseudometric in $L^p$ space. (2) It is a metric in $L^p$ space when functions $f$ and $g$ are identified if $f = g$ almost everywhere on $D$.

**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 $D_{\text{NPEC}}^U$).** Let $R_A, R_B : S \times A \times S \to \mathbb{R}$ be bounded reward functions, and $\lambda \geq 0$. Then $D_{\text{NPEC}}^U$:

- **Is invariant under $\equiv$ in source:**
  $$D_{\text{NPEC}}^U(R_A, R_B) = D_{\text{NPEC}}^U(R_B, R_B)$$
  if $R_A \equiv R_B$.

- **Invariant under scale-preserving $\equiv$ in target:**
  $$D_{\text{NPEC}}^U(R_A, R_A) = D_{\text{NPEC}}^U(R_A, R_B)$$
  if $R_A - R_B \equiv \text{Zero}$.

- **Scalable in target:**
  $$D_{\text{NPEC}}^U(R_A, \lambda R_B) = \lambda D_{\text{NPEC}}^U(R_A, R_B).$$

- **Bounded:**
  $$D_{\text{NPEC}}^U(R, R_B) \leq D_{\text{NPEC}}^U(\text{Zero}, R_B).$$

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

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

If $R_A \equiv R_B$, then:

$$D_{\text{NPEC}}^U(R_A, R_B) \triangleq \inf_{R \equiv R_A} D_{L^p, \mathcal{D}}(R, R_B) \triangleq \inf_{R \equiv R_B} D_{L^p, \mathcal{D}}(R, R_B) \triangleq D_{\text{NPEC}}^U(R_B, R_B),$$

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 $R_A - R_B \equiv \text{Zero}$, then we can write $R_A(s, a, s') - R_B(s, a, s') = \gamma \Phi(s') - \Phi(s)$ for some potential function $\Phi : S \to \mathbb{R}$. Then for any reward function $R$, since $D$ is induced by a norm:

$$D_{L^p, \mathcal{D}}(R, R_A) \triangleq \mathbb{E}_{s, a, s' \sim D} [D(R(s, a, s'), R_A(s, a, s'))] = \mathbb{E}_{s, a, s' \sim D} [\|R(s, a, s') - R_A(s, a, s')\|] = \mathbb{E}_{s, a, s' \sim D} [\|R(s, a, s') - (R_B(s, a, s') + \gamma \Phi(s') - \Phi(s))\|] = \mathbb{E}_{s, a, s' \sim D} [\|(R(s, a, s') - \gamma \Phi(s') + \Phi(s)) - R_B(s, a, s')\|] = \mathbb{E}_{s, a, s' \sim D} [D(R(s, a, s') - \gamma \Phi(s') + \Phi(s), R_B(s, a, s'))] \triangleq D_{L^p, \mathcal{D}}(f(R), R_B), \quad (3)$$
where \( f(R)(s, a, s') = R(s, a, s') - \gamma \Phi(s') + \Phi(s) \). Crucially, note \( f(R) \) is a bijection on the equivalence class \([R] \). Now, substituting this into the expression for NPEC premetric:

\[
D^U_{\text{NPEC}}(R_A, R_A) \triangleq \inf_{R \equiv R_A} D_{L^p, D}(R, R_A)
\]

\[
= \inf_{R \equiv R_A} D_{L^p, D}(f(R), R_A) \quad \text{eq. (3)}
\]

\[
= \inf_{f(R) \equiv R_A} D_{L^p, D}(f(R), R_B) \quad f \text{ bijection on } [R]
\]

\[
= \inf_{R \equiv R_A} D_{L^p, D}(R, R_B) \quad f \text{ bijection on } [R]
\]

\[
= D^U_{\text{NPEC}}(R_A, R_B).
\]

**Scalable in target** First, note that \( D_{L^p, D} \) is absolutely scalable in both arguments:

\[
D_{L^p, D}(\lambda R_A, \lambda R_B) \triangleq \mathbb E_{s,a,s' \sim D} [D(\lambda R_A(s, a, s'), \lambda R_B(s, a, s'))] \quad (4)
\]

\[
= \mathbb E_{s,a,s' \sim D} [||\lambda R_A(s, a, s') - \lambda R_B(s, a, s')||] \quad (5)
\]

\[
= \mathbb E_{s,a,s' \sim D} [||\lambda||R_A(s, a, s') - R_B(s, a, s')||] \quad ||\cdot||\text{absolutely scalable} (6)
\]

\[
= ||\lambda|| \mathbb E_{s,a,s' \sim D} [||R_A(s, a, s') - R_B(s, a, s')||] \quad (7)
\]

\[
= ||\lambda|| D_{L^p, D}(R_A, R_B). \quad (8)
\]

Now, for \( \lambda > 0 \), applying this to NPEC premetric:

\[
D^U_{\text{NPEC}}(R_A, \lambda R_B) \triangleq \inf_{R \equiv R_A} D_{L^p, D}(R, \lambda R_B)
\]

\[
= \inf_{R \equiv R_A} D_{L^p, D}(R, \lambda R_B) \quad R \equiv \lambda R
\]

\[
= \lambda \inf_{R \equiv R_A} D_{L^p, D}(R, R_B)
\]

\[
= \lambda D^U_{\text{NPEC}}(R_A, R_B).
\]

In the case \( \lambda = 0 \), then:

\[
D^U_{\text{NPEC}}(R_A, 0) \triangleq \inf_{R \equiv R_A} D_{L^p, D}(R, 0)
\]

\[
= \inf_{R \equiv R_A} D_{L^p, D}(R, 0) \quad R \equiv 1/2 R
\]

\[
= \frac{1}{2} \inf_{R \equiv R_A} D_{L^p, D}(R, 0)
\]

\[
= \frac{1}{2} D^U_{\text{NPEC}}(R_A, 0).
\]

Rearranging, we have:

\[
D^U_{\text{NPEC}}(R_A, 0) = 0.
\]

**Boundedness**

Suppose \( R_A \) is bounded by \( B \): \(|R_A(s, a, s')| \leq B \) for all \( s, s' \in S \) and \( a \in A \). Suppose the NPEC premetric \( D_{\text{NPEC}}(0, R_B) = d \). Then for any \( \epsilon > 0 \), there exists some potential function \( \Phi : S \to \mathbb R \) such that the \( L^p \) of the potential shaping \( R(s, a, s') \triangleq \gamma \Phi(s') - \Phi(s) \) from \( R_B \) satisfies:

\[
D_{L^p, D}(R, R_B) \leq d + \epsilon. \quad (9)
\]

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

\[
R'_A(s, a, s') \triangleq \lambda R_A(s, a, s') + R(s, a, s'),
\]
Theorem 5.5.

Let $D$ be a premetric. Moreover, let $R_A$ and $R_B$ be bounded reward functions such that $R_A \equiv R_A'$ and $R_B \equiv R_B'$. Then $0 \leq D_{\text{NPEC}}(R_A', R_B') = D_{\text{NPEC}}(R_A, R_B) \leq 1$.

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

Premetric

First, we will show that $D_{\text{NPEC}}$ is a premetric. Let $R_A, R_B$ be bounded reward functions on $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$.

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

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

$$D_{\text{NPEC}}(R_A, R_A) = \frac{D_{\text{NPEC}}(R_A, R_A)}{D_{\text{NPEC}}(\text{Zero}, R_A)} = \frac{0}{D_{\text{NPEC}}(\text{Zero}, R_A)} = 0.$$
Well-defined: \( D_{\text{NPEC}}(R_A, R_B) \geq 0 \)

By proposition A.3, it follows that \( D_{L^p, \mathcal{D}}(R, R_B) \geq 0 \) for all reward functions \( R : S \times A \times S \). Thus 0 is a lower bound for \( \{ D_{L^p, \mathcal{D}}(R, R_B) \mid R : S \times A \times S \} \), 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 largest lower bound, it follows that for any reward function \( X \):

\[
D_{\text{NPEC}}(X, R_B) \triangleq \inf_{R \equiv X} D_{L^p, \mathcal{D}}(R, R_B) \geq 0.
\]

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

\[
D_{\text{NPEC}}(R_A, R_B) = \frac{D_{\text{NPEC}}(R_A, R_B)}{D_{\text{NPEC}}(\text{Zero}, R_B)} \geq 0.
\]

Invariant and Bounded

Since \( R_B' \equiv R_B \), we have \( R_B' - \lambda R_B \equiv \text{Zero} \) for some \( \lambda > 0 \). By proposition A.1, \( D_{\text{NPEC}} \) is invariant under scale-preserving \( \equiv \) in target and scalable in target. That is, for any reward \( R \):

\[
D_{\text{NPEC}}(R, R_B') = D_{\text{NPEC}}(R, \lambda R_B) = \lambda D_{\text{NPEC}}(R, R_B).
\]

By eq. (12):

\[
D_{\text{NPEC}}(R_A, R_B') = \frac{D_{\text{NPEC}}(R_A, R_B)}{D_{\text{NPEC}}(\text{Zero}, R_B)} = D_{\text{NPEC}}(R_A, R_B).
\]

By proposition A.1, \( D_{\text{NPEC}}(R_A, R_B) = D_{\text{NPEC}}(R_B', R_B) \), so:

\[
D_{\text{NPEC}}(R_A, R_B') = \frac{D_{\text{NPEC}}(R_A, R_B)}{D_{\text{NPEC}}(\text{Zero}, R_B)} = D_{\text{NPEC}}(R_A, R_B).
\]

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

\[
D_{\text{NPEC}}(R_A, R_B) = \frac{D_{\text{NPEC}}(R_A, R_B)}{D_{\text{NPEC}}(\text{Zero}, R_B)} \leq 1.
\]

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

**Proposition A.2.** \( D_{\text{NPEC}} \) is not symmetric in the undiscounted case.

**Proof.** We will provide a counterexample showing that \( D_{\text{NPEC}} \) is not symmetric.

Choose the state space \( S \) to be binary \{0, 1\} and the actions \( A \) to be the singleton \{0\}. Choose the visitation distribution \( \mathcal{D} \) to be uniform on \( s \rightarrow s \) for \( s \in S \). 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 \( D_{\text{direct}} \), 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 \( D_{L^p, \mathcal{D}}(\text{Zero}, R_A) = \frac{1}{2} (|0| + |2|) = 1 \) and \( D_{L^p, \mathcal{D}}(\text{Zero}, R_B) = \frac{1}{2} (|1| + |1|) = 1 \). Since potential shaping has no effect, \( D_{\text{NPEC}}(\text{Zero}, R) = D_{L^p, \mathcal{D}}(Zero, R) \) and so \( D(\text{Zero}, R_A) = 1 \) and \( D(\text{Zero}, R_B) = 1 \).
Now:

\[ D_{\text{UNPEC}}^{U}(R_{A}, R_{B}) = \inf_{\lambda > 0} D_{L^{p}, \mathcal{D}}(\lambda R_{A}, R_{B}) \]
\[ = \inf_{\lambda > 0} \frac{1}{2} (|1| + |2\lambda - 1|) \]
\[ = \frac{1}{2} \]

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

\[ D_{\text{UNPEC}}^{U}(R_{B}, R_{A}) = \inf_{\lambda > 0} D_{L^{p}, \mathcal{D}}(\lambda R_{B}, R_{A}) \]
\[ = \inf_{\lambda > 0} \frac{1}{2} f(\lambda) \]
\[ = \frac{1}{2} \inf_{\lambda > 0} f(\lambda), \]

where:

\[ f(\lambda) = |\lambda| + |2 - \lambda|, \quad \lambda > 0. \]

Note that:

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

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

\[ D_{\text{UNPEC}}^{U}(R_{B}, R_{A}) = 1. \]

Consequently:

\[ D_{\text{UNPEC}}(R_{A}, R_{B}) = \frac{1}{2} \neq 1 = D_{\text{UNPEC}}(R_{B}, R_{A}). \]

### A.4 Direct Distance Variant of EPIC

Previously, we used Pearson distance 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 \( R \) be a reward function mapping from transitions \( S \times A \times S \) to real numbers \( \mathbb{R} \). Let \( \| \cdot \| \) be some norm on the vector space of reward functions over the real field. Then the normalized \( R \) is:

\[ R^{N}(s, a, s') = \frac{R(s, a, s')}{\| R \|} \]

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 \( R \) be a reward function mapping from transitions \( S \times A \times S \) to real numbers \( \mathbb{R} \). Then the standardized \( R \) is:

\[ R^{S} = (C_{\mathcal{D}S, \mathcal{D}A}(R))^{N}. \]

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 visitation distribution over transitions \( s \xrightarrow{a} s' \). Let \( \mathcal{D}_{S} \) and \( \mathcal{D}_{A} \) be some distributions over states \( S \) and \( A \) 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 direct distance between their standardized versions over \( \mathcal{D} \):

\[ D_{\text{DDSR}}(R_{A}, R_{B}) = \frac{1}{2} D_{L^{p}, \mathcal{D}} \left( R^{S}_{A}(S, A, S'), R^{S}_{B}(S, A, S') \right), \]

where the norm used for direct distance is the same norm used for normalization in \( R^{N} \).
For brevity, we omit the proof that \( D_{DDSR} \) is a pseudometric, but this follows from \( D_{L_p,D} \) 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 \( S \times A \times S \) to real numbers \( \mathbb{R} \) such that \( R_A \equiv R_A' \) and \( R_B \equiv R_B' \). Then:

\[
0 \leq D_{DDSR}(R_A', R_B') = D_{DDSR}(R_A, R_B) \leq 1.
\]

**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 \( D_{L_p,D} \) being a pseudometric. The upper bound follows from the rewards being normalized to norm 1 and the triangle inequality:

\[
D_{DDSR}(R_A, R_B) = \frac{1}{2} \| R_A^S - R_B^S \| \\
\leq \frac{1}{2} (\| R_A^S \| + \| R_B^S \|) \\
= \frac{1}{2} (1 + 1) \\
= 1.
\]

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 with \( D \) chosen to be i.i.d. samples from the joint distribution \( D_S \times D_A \times D_S \), 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 \( S \times A \times S \) to real numbers \( \mathbb{R} \). Then:

\[
\mathbb{E} \left[ C_{D_S,D_A} (R) (S, A, S') \right] = 0.
\]

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

\[
\mathbb{E} \left[ C_{D_S,D_A} (R) (S, A, S') \right] = \mathbb{E} \left[ R(S, A, S') + \gamma R(S', U, X') - R(S, U, X') \right] \\
= \mathbb{E} \left[ R(S, A, S') \right] + \gamma \mathbb{E} \left[ R(S', U, X') \right] - \mathbb{E} \left[ R(S, U, X') \right] - \gamma \mathbb{E} \left[ R(X, U, X') \right] \\
= \mathbb{E} \left[ R(S, U, X') \right] + \gamma \mathbb{E} \left[ R(X, U, X') \right] - \mathbb{E} \left[ R(S, U, X') \right] - \gamma \mathbb{E} \left[ R(X, U, X') \right] \\
= 0,
\]

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

Recall from the proof of lemma 4.5 that:

\[
D_\rho(U, V) = \frac{1}{2} \sqrt{\mathbb{E} \left[ (\hat{U} - \hat{V})^2 \right]} \\
= \frac{1}{2} \| \hat{U} - \hat{V} \|_2,
\]

where \( \| \cdot \|_2 \) is the \( L^2 \) norm (treating the random variables as functions on a measure space) and \( \hat{U} \) 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 \( D_S \times D_A \times D_S \), and normalization by the \( L^2 \) norm also ensures they have unit variance. Consequently:

\[
D_{EPIC}(R_A, R_B) = D_\rho \left( C_{D_S,D_A} (R_A) (S, A, S'), C_{D_S,D_A} (R_B) (S, A, S') \right) \\
= \frac{1}{2} \left\| \left( C_{D_S,D_A} (R_A) (S, A, S') \right)^N - \left( C_{D_S,D_A} (R_B) (S, A, S') \right)^N \right\|_2 \\
= \frac{1}{2} \| R_A^S(S, A, S') - R_B^S(S, A, S') \|_2 \\
= \frac{1}{2} D_{L_p,D} (R_A^S(S, A, S'), R_B^S(S, A, S')) \\
= D_{DDSR}(R_A, R_B).
\]
A.5 Regret Bound

In this section, we derive an upper bound on the regret in terms of 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 \( D_{\text{EPIC}}(R_A, R_B) \).

First, in section A.5.1 we derive a bound for MDPs with discrete state and action spaces. In section A.6 we then present another 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 zero as \( D_{\text{EPIC}}(R_A, R_B) \to 0 \).

### A.5.1 Discrete MDPs

We start in lemma A.8 by showing that \( L^2 \) distance upper bounds \( L^1 \) distance. Next, in lemma A.9 we show regret is bounded by the \( L^1 \) distance between reward functions using an argument similar to [20]. Then in lemma A.10 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 \( D_{\text{EPIC}} \) in terms of the \( L^2 \) distance on standardized rewards, deriving a bound on regret in terms of the EPIC distance.

**Lemma A.8.** Let \( (\Omega, \mathcal{F}, p) \) be a probability space and \( f : \Omega \to \mathbb{R} \) a measurable function whose absolute value raised to the \( p \)-th power for \( p \in \{1, 2\} \) has a finite expectation. Then the \( L^1 \) norm of \( f \) is bounded above by the \( L^2 \) norm:

\[
\|f\|_1 \leq \|f\|_2. \tag{13}
\]

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

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

\[
= \mathbb{E}[|f(X)|^2 - \mathbb{E}[|f(X)|]^2] \geq 0.
\]

Rearranging terms, we have

\[
\|f\|_2^2 = \mathbb{E}[|f(X)|^2] \geq \mathbb{E}[|f(X)|]^2 = \|f\|_1^2.
\]

Taking the square root of both sides gives:

\[
\|f\|_1 \leq \|f\|_2. \tag*{\square}
\]

**Lemma A.9.** Let \( M \) be an MDP with discrete state and action spaces \( S \) and \( A \). Let \( R_A, R_B : S \times A \times S \to \mathbb{R} \) be rewards. Let \( \pi_A^* \) and \( \pi_B^* \) be policies optimal for rewards \( R_A \) and \( R_B \) in \( M \). Let \( D_\pi(t, s_1, a, s_1) \) denote the distribution over trajectories that policy \( \pi \) induces in \( M \) at time step \( t \). Let \( D(t, s, a, s') \) be the (stationary) visitation distribution over transitions \( S \times A \times S \) used to compute \( D_{\text{EPIC}} \). Suppose that there exists some \( K > 0 \) such that \( KD(t, s_1, a_1, s_1') \geq D\pi(t, s_1, a_1, s_1') \) for all time steps \( t \in \mathbb{N} \), triples \( s_1, a_1, s_1' \in S \times A \times S \), 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:

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

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

\[
G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) = |G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*)|
\]

\[
\leq |G_{R_A}(\pi_A^*) - G_{R_B}(\pi_B^*)| + |G_{R_B}(\pi_B^*) - G_{R_A}(\pi_A^*)| 
\]

\[
\leq |G_{R_A}(\pi_A^*) - G_{R_B}(\pi_B^*)| + |G_{R_B}(\pi_B^*) - G_{R_A}(\pi_B^*)|. \tag{14}
\]

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 \( \pi \in \{\pi_A^*, \pi_B^*\} \):

\[
|G_{R_A}(\pi) - G_{R_B}(\pi)| \leq \frac{K}{1-\gamma} D_{L^1, \mathcal{D}}(R_A, R_B).
\]
Writing $R(\tau) = \sum_{t=0}^{T} \gamma^t R(s_t, a_t, s'_t)$ for convenience, we have that for $\pi \in \{\pi^*_A, \pi^*_B\}$:

$$|G_{R_A}(\pi) - G_{R_B}(\pi)| = \left| \mathbb{E}_{\tau \sim D_\pi} \left[ \sum_{t=0}^{\infty} \gamma^t (R_A(s_t, a_t, s'_t) - R_B(s_t, a_t, s'_t)) \right] \right|$$

$$\leq \mathbb{E}_{\tau \sim D_\pi} \left[ \sum_{t=0}^{\infty} \gamma^t |R_A(s_t, a_t, s'_t) - R_B(s_t, a_t, s'_t)| \right]$$

$$= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s_t, a_t, s_{t+1} \sim D_\pi} [|R_A(s_t, a_t, s'_t) - R_B(s_t, a_t, s'_t)|]$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_{s_t, a_t, s_{t+1} \in S \times A \times S} D_\pi(t, s_t, a_t, s_{t+1}) |R_A(s_t, a_t, s'_t) - R_B(s_t, a_t, s'_t)|$$

$$\leq K \sum_{t=0}^{\infty} \gamma^t \sum_{s_t, a_t, s_{t+1} \in S \times A \times S} D(s_t, a_t, s_{t+1}) |R_A(s_t, a_t, s'_t) - R_B(s_t, a_t, s'_t)|$$

$$= K \sum_{t=0}^{\infty} \gamma^t D_{L^1,\mathcal{D}}(R_A, R_B)$$

$$= \frac{K}{1 - \gamma} D_{L^1,\mathcal{D}}(R_A, R_B),$$

as required.

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

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

Rearranging gives:

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

So certainly:

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

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

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

It follows that

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

Substituting inequalities $15$ and $16$ into eq. $14$ yields the required result. \qed

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

**Lemma A.10.** Let $M$ be an MDP with state and action spaces $S$ and $A$. Let $R_A, R_B : S \times A \times S \rightarrow \mathbb{R}$ 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 $R^*_A$ from executing $\pi^*_B$ instead of $\pi^*_A$ is upper bounded by some $U \in \mathbb{R}$:

$$G_{R^*_A}(\pi^*_A) - G_{R^*_A}(\pi^*_B) \leq U. \quad (17)$$

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

$$G_{R_A}(\pi^*_A) - G_{R_A}(\pi_B) \leq 4U\|R_A\|_2.$$
Theorem 4.9. Let $R_A, R_B : S \times A \times S \to \mathbb{R}$ be bounded rewards, and $\pi^*_A, \pi^*_B$ be respective optimal policies. Let $D_t(s, a, s')$ denote the visitation distribution over transitions $S \times A \times S$ induced by policy $\pi$ at time $t$, and $D(s, a, s')$ be the visitation distribution used to compute $D_{\text{EPIC}}$. Suppose there exists $K > 0$ such that $K D_t(s, a, s') \geq D_t(s, a, s')$ for all times $t \in \mathbb{N}$, triples $(s_t, a_t, s_{t+1}) \in S \times A \times S$ 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 start by defining a relaxation of the Wasserstein distance. Then, we bound the expected value under distribution $\nu$. Applying lemma A.10 yields the required result:

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

Applying lemma A.8 we obtain:

$$D_{L^1, D}(R_A^S, R_B^S) = \left\| R_A^S(S, A, S') - R_B^S(S, A, S') \right\|_1 \leq 2D_{\text{EPIC}}(R_A, R_B). \tag{20}$$

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.9 and eq. 20 gives

$$G_{R^S_A}(\pi_A^*) - G_{R^S_B}(\pi_B^*) \leq \frac{2K}{1 - \gamma} D_{L^1, D}(R_A^S, R_B^S) \leq \frac{4K}{1 - \gamma} D_{\text{EPIC}}(R_A, R_B). \tag{21}$$

Applying lemma A.10 yields the required result:

$$G_{R_A}(\pi_A^*) - G_{R_B}(\pi_B^*) \leq \frac{16K}{1 - \gamma} D_{\text{EPIC}}(R_A, R_B). \tag{22}$$

### 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 $\nu$ may be discrete. However, the expectation over a continuous distribution $D$ is unaffected by the reward at any finite subset of points. Accordingly, the reward can be varied arbitrarily on transitions $T$ – 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.11. In lemma A.12 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.13 we bound the regret in terms of the $L^1$ distance between the rewards plus $W_\alpha$; this is analogous to lemma A.9 in the discrete case. Finally, in theorem A.14 we use the previous results to bound the regret in terms of the EPIC distance plus $W_\alpha$.

**Definition A.11.** Let $S$ be some set and let $\mu, \nu \in \Delta(S)$ be probability distributions on $S$. We define the relaxed Wasserstein distance between $\mu$ and $\nu$ by:

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

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

$$\int_S p(x, y) \, dy = \mu(x), \tag{22}$$

$$\int_S p(x, x) \, dx \leq \alpha \nu(y). \tag{23}$$

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

**Lemma A.12.** Let $S$ be some set and let $\mu, \nu \in \Delta(S)$ be probability distributions on $S$. Let $f : S \to \mathbb{R}$ be an $L$-Lipschitz function on norm $\| \cdot \|$. Then, for any $\alpha \geq 1$:

$$\mathbb{E}_{X \sim \mu} [\| f(X) \|] \leq \alpha \mathbb{E}_{Y \sim \nu} [\| f(Y) \|] + L W_\alpha(\mu, \nu).$$
Proof. Let $p \in \Gamma_\alpha(\mu, \nu)$. Then:

$$\mathbb{E}_{X \sim \mu} [||f(X)||] \triangleq \int |f(x)| d\mu(x)$$

definition of $\mathbb{E}$

$$= \int |f(x)| d\mu(x, y)$$

$\mu$ is a marginal of $p$

$$\leq \int |f(y)| + L \|x - y\| d\nu(x, y)$$

$f$ L-Lipschitz

$$= \int |f(y)| d\nu(x, y) + L \int \|x - y\| d\nu(x, y)$$

$$= \int |f(y)| \alpha \nu(y) dy + L \int \|x - y\| d\nu(x, y)$$

eq 23

$$= \alpha \mathbb{E}_{Y \sim \nu} [||f(Y)||] + L \int \|x - y\| d\nu(x, y)$$

definition of $\mathbb{E}$.

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

$$\mathbb{E}_{X \sim \mu} [||f(X)||] \leq \alpha \mathbb{E}_{Y \sim \nu} [||f(Y)||] + L \inf_{p \in \Gamma_\alpha(\mu, \nu)} \int \|x - y\| d\nu(x, y)$$

$$= \alpha \mathbb{E}_{Y \sim \nu} [||f(Y)||] + LW_\alpha(\mu, \nu).$$

\[\square\]

Lemma A.13. Let $M$ be an MDP with state and action spaces $S$ and $A$. Let $R_A, R_B : S \times A \times S \to \mathbb{R}$ be L-Lipschitz rewards on some norm $||\cdot||$ on $S \times A \times S$. Let $\pi_A^*$ and $\pi_B^*$ be policies optimal for rewards $R_A$ and $R_B$ in $M$. Let $D_{\pi, t}(s_t, a_t, s_{t+1})$ denote the distribution over trajectories that policy $\pi$ induces in $M$ at time step $t$. Let $D(s, a, s')$ be the (stationary) visitation distribution over transitions $S \times A \times S$ used to compute $D_{\text{EPIC}}$. Let $\alpha \geq 1$, and let $B_\alpha(t) = \max_{\pi \in \{\pi_A^*, \pi_B^*\}} W_\alpha(D_{\pi, t}, D)$. Then the regret under $R_A$ from executing $\pi_B^*$ optimal for $R_B$ instead of $\pi_A^*$ is at most:

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq \frac{2\alpha}{1 - \gamma} D_{L_1, D}(R_A, R_B) + 4L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t).$$

Proof. By the same argument as lemma A.9, we have for any policy $\pi$:

$$|G_{R_A}(\pi) - G_{R_B}(\pi)| \leq \sum_{t=0}^{\infty} \gamma^t D_{L_1, D}(R_A, R_B).$$

Let $f(s, a, s') = R_A(s, a, s') - R_B(s, a, s')$, and note $f$ is 2L-Lipschitz since $R_A$ and $R_B$ are both L-Lipschitz. Now, by lemma A.12, letting $\mu = D_{\pi, t}$ and $\nu = D$, we have:

$$D_{L_1, D}(R_A, R_B) \leq \alpha D_{L_1, D}(R_A, R_B) + 2LW_\alpha(D_{\pi, t}, D).$$

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

$$|G_{R_A}(\pi) - G_{R_B}(\pi)| \leq \frac{\alpha}{1 - \gamma} D_{L_1, D}(R_A, R_B) + 2L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t).$$

By the same argument as for lemma A.9, it follows that:

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq \frac{2\alpha}{1 - \gamma} D_{L_1, D}(R_A, R_B) + 4L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t).$$

\[\square\]

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

$$G_{R_A}(\pi_A^*) - G_{R_A}(\pi_B^*) \leq 16 \|R_A\|_2 \left( \frac{\alpha}{1 - \gamma} D_{\text{EPIC}}(R_A, R_B) + L \sum_{t=0}^{\infty} \gamma^t B_\alpha(t) \right).$$
A key insight is that the relaxed Wasserstein distance. Applying lemma A.10 yields the required result:

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

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

Proof. Suppose it follows that for any $\alpha$ not contain point mass probabilities. Let

Let $0 < \delta > 0$ and suppose $\nu(y) \geq \delta$ for all $y \in S$; moreover, assume $\nu$ is not degenerate (i.e. does not contain point mass probabilities). Let $f : S \to \mathbb{R}$ be an $L$-Lipschitz function on norm $\| \cdot \|$. Then as $E_{Y \sim \nu} [\| f(Y) \|] \to 0$, $E_{X \sim \mu} [\| f(X) \|] \to 0$ for arbitrary $\mu \in \Delta(S)$.

Proof. Suppose $E_{Y \sim \nu} [\| f(Y) \|] \leq \epsilon^2$ for some $\epsilon > 0$. Choose $\alpha = 1/\epsilon$, then by lemma A.12

$$E_{X \sim \mu} [\| f(X) \|] \leq \epsilon + LW_{1/\epsilon}(\mu, \nu).$$

It remains to show that $W_\alpha(\mu, \nu) \to 0$ as $\alpha \to \infty$. Define $A(r, z) \triangleq \{ s \in S \mid \| s - z \| \leq r \}$ and

$$V(r, x) \triangleq \int_{A(r, x)} \nu(y) dy.$$ 

Note $V(r, x)$ is continuous in $r$ since $\nu$ is not a degenerate distribution. Moreover, $V(0, x) = 0$ while $V(R, x) = 1$ for sufficiently large $R$ (noting that $S$ is bounded). By the intermediate value theorem, it follows that for any $\alpha \geq 1$ there always exists an $r \in [0, R]$ for which $V(r, x) = \frac{1}{\alpha}$, although this $r$ need not be unique. We let $r^*(x)$ denote some mapping such that $V(r^*(x), x) = \frac{1}{\alpha}$. Now, define a joint distribution

$$p(x, y) = \begin{cases} \alpha \mu(x) \nu(y), & \| x - y \| \leq r^*(y) \\ 0, & \text{otherwise}, \end{cases}$$

Note that $p \in \Gamma_\alpha(\mu, \nu)$ since, for all $x, y \in S$:

$$\int_S p(x, y) dy = \alpha \mu(x) V(r^*(x), x) = \mu(x),$$

$$\int_S p(x, y) dx = \alpha \nu(y) \int_{A(r^*(y), y)} \mu(x) dx \leq \alpha \nu(y) \int_S \mu(x) dx = \alpha \nu(y).$$

### A.7 Limiting Behavior of Regret

Theorem 4.9 - the regret bound for discrete MDPs - directly implies that, as EPIC distance tends to 0, the regret also tends to 0. By contrast, our regret bound in theorem A.14 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 $D_{EPIC}(R_A, R_B) \to 0$ in the Lipschitz case as well as the discrete case.

A key insight is that the relaxed Wasserstein distance $W_\alpha$ tends to 0 as $\alpha \to \infty$. However, the first term becomes arbitrarily large as $\alpha \to \infty$. In lemma A.15 we choose $\alpha$ to balance these competing demands. We conclude in theorem A.16 by showing the regret tends to 0 as EPIC distance tends to 0.

**Lemma A.15.** Let $S$ be a set bounded on norm $\| \cdot \|$. Let $\mu, \nu \in \Delta(S)$ be probability distributions on $S$. Let $\delta > 0$ and suppose $\nu(y) \geq \delta$ for all $y \in S$; moreover, assume $\nu$ is not degenerate (i.e. does not contain point mass probabilities). Let $f : S \to \mathbb{R}$ be an $L$-Lipschitz function on norm $\| \cdot \|$. Then as $E_{Y \sim \nu} [\| f(Y) \|] \to 0$, $E_{X \sim \mu} [\| f(X) \|] \to 0$ for arbitrary $\mu \in \Delta(S)$.

**Proof.** Suppose $E_{Y \sim \nu} [\| f(Y) \|] \leq \epsilon^2$ for some $\epsilon > 0$. Choose $\alpha = 1/\epsilon$, then by lemma A.12

$$E_{X \sim \mu} [\| f(X) \|] \leq \epsilon + LW_{1/\epsilon}(\mu, \nu).$$

It remains to show that $W_\alpha(\mu, \nu) \to 0$ as $\alpha \to \infty$. Define $A(r, z) \triangleq \{ s \in S \mid \| s - z \| \leq r \}$ and

$$V(r, x) \triangleq \int_{A(r, x)} \nu(y) dy.$$ 

Note $V(r, x)$ is continuous in $r$ since $\nu$ is not a degenerate distribution. Moreover, $V(0, x) = 0$ while $V(R, x) = 1$ for sufficiently large $R$ (noting that $S$ is bounded). By the intermediate value theorem, it follows that for any $\alpha \geq 1$ there always exists an $r \in [0, R]$ for which $V(r, x) = \frac{1}{\alpha}$, although this $r$ need not be unique. We let $r^*(x)$ denote some mapping such that $V(r^*(x), x) = \frac{1}{\alpha}$. Now, define a joint distribution

$$p(x, y) = \begin{cases} \alpha \mu(x) \nu(y), & \| x - y \| \leq r^*(y) \\ 0, & \text{otherwise}, \end{cases}$$

Note that $p \in \Gamma_\alpha(\mu, \nu)$ since, for all $x, y \in S$:

$$\int_S p(x, y) dy = \alpha \mu(x) V(r^*(x), x) = \mu(x),$$

$$\int_S p(x, y) dx = \alpha \nu(y) \int_{A(r^*(y), y)} \mu(x) dx \leq \alpha \nu(y) \int_S \mu(x) dx = \alpha \nu(y).$$

34
Then as
Applying lemma A.10 we have:
\[ \text{Theorem A.16.} \]

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

Suppose that either:

1. Discrete: \( S \) and \( A \) are discrete. Moreover, suppose that there exists some \( K > 0 \) such that \( KD(s_t, a_t, s_{t+1}) \geq D_\pi(t, s_t, a_t, s_{t+1}^\pi) \) for all time steps \( t \in \mathbb{N} \), triples \( s_t, a_t, s_{t+1} \in S \times A \times S \) and policies \( \pi \in \{ \pi_A^*, \pi_B^* \} \).

2. Lipschitz: \( R_A \) and \( R_B \) are \( L \)-Lipschitz. Moreover, suppose there exists some \( \delta > 0 \) such that \( D(s, a, s') \geq \delta \) for all \( s, a, s' \in S \times A \times S \) and that \( D \) is not degenerate (i.e. does not contain point mass probabilities).

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

\[ \text{Proof.} \] In case (1) Discrete, by theorem 4.9

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

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.9, we have for any policy \( \pi \):

\[ |G_{R_A}(\pi) - G_{R_B}(\pi)| \leq \sum_{t=0}^{\infty} \gamma^t D_{L^1, D_{\pi}}(R_A, R_B). \]

Applying lemma A.10 we have:

\[ |G_{R_A}(\pi) - G_{R_B}(\pi)| \leq 4\|R_A\| \sum_{t=0}^{\infty} \gamma^t D_{L^1, D_{\pi}}(R_A^S, R_B^S). \]

By equation 20 we know that \( D_{L^1, D_{\pi}}(R_A^S, R_B^S) \to 0 \) as \( D_{\text{EPIC}}(R_A, R_B) \to 0 \). By lemma A.15 we know that \( D_{L^1, D_{\pi}}(R_A^S, R_B^S) \to 0 \) as \( D_{L^1, D_{\pi}}(R_A^S, R_B^S) \to 0 \). So we can conclude that as \( D_{\text{EPIC}}(R_A, R_B) \to 0 \):

\[ |G_{R_A}(\pi) - G_{R_B}(\pi)| \to 0. \]

\( \square \)