A general framework for reward function distances

Erik Jenner *
UC Berkeley
jenner@berkeley.edu

Joar Skalse *
Oxford University
joar.skalse@cs.ox.ac.uk

Adam Gleave
UC Berkeley
gleave@berkeley.edu

Abstract

In reward learning, it is helpful to be able to measure *distances* between reward functions, for example to evaluate learned reward models. Using simple metrics such as L^2 distances is not ideal because reward functions that are equivalent in terms of their optimal policies can nevertheless have high L^2 distance. EPIC [5] and DARD [19] are distances specifically designed for reward functions that address this by being *invariant* under certain transformations that leave optimal policies unchanged. However, EPIC and DARD are designed in an ad-hoc manner, only consider a subset of relevant reward transformations, and suffer from serious pathologies in some settings. In this paper, we define a general class of reward function distance metrics, of which EPIC is a special case. This framework lets as address all these issues with EPIC and DARD, and allows for the development of reward function distance metrics in a more principled manner.

1 Introduction

Specifying a good reward function can be difficult, especially as reinforcement learning tasks become more complex. In many cases, specifying a reward function by hand is in fact infeasible, so the reward must be *learned* instead [12, 1, 4, 2, 20]. Of course, such learned reward functions can be wrong, making *evaluation* of learned reward functions and of reward learning algorithms crucial. One approach is to interpret reward models [14, 11, 7]. But another important tool to have for evaluation are *distance metrics* between reward functions [5, 19]. For example, they can be used to benchmark reward learning algorithms by comparing learned reward functions to the ground truth.

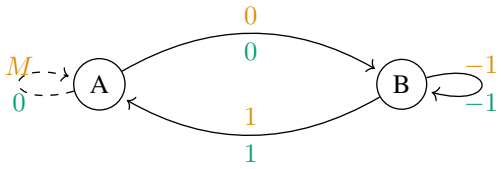
We could of course simply use the L^2 distance between reward functions. But this is unsatisfactory, because two reward functions can induce the exact same ordering of policies by expected returns while having arbitrarily high L^2 distance. For example, applying potential shaping [13] to a reward function, or scaling by a positive constant, never changes its policy ordering. To deal with this, existing metrics [5, 19] aim to be invariant to transformations which do not change the policy order, so that rewards with the same policy order get zero distance. However, these existing metrics do not consider all relevant transformations and suffer from other issues (see section 3).

Our contribution is to introduce a very general framework for distance metrics between reward functions. This framework contains EPIC as a special case but also allows invariance to a wider range of transformations. Additionally, this general setting lets us fix pathologies in EPIC and DARD that we identify, and prove stronger regret bounds than the existing one. Our framework also highlights that existing metrics have been chosen somewhat arbitrarily, and suggests certain natural design choices, which existing metrics have not used.

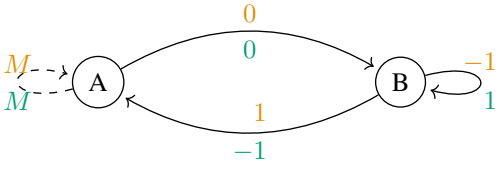
Related work Our work builds on existing reward function distances, EPIC [5] and DARD [19], which we discuss more in section 2. Also important for us are discussions of equivalence relations between reward functions [3, 9, 16, 15] and in particular work on potential shaping [13, 8].

*Equal contribution.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) Rewards equivalent on distribution, but EPIC distance high



(b) Rewards very different on distribution, but EPIC distance low

Figure 1: Toy MDPs where the EPIC distance is unreasonably high or low because of a mismatch between the canonicalization and coverage distribution. There are two states, A and B, with two actions in each state (all deterministic). The numbers along transitions describe two reward functions (one in orange, one in green). The canonicalization distribution is assumed to be uniform over all four transitions. The coverage distribution is uniform over all but the dashed transition on the left, which has probability zero. EPIC distances become arbitrarily bad as $M \to \infty$.

2 Background

Equivalent reward transformations Scaling a reward function by any positive constant clearly does not affect the ordering of policies by expected returns, under arbitrary transition dynamics. Similarly, given any function $\Phi : \mathcal{S} \to \mathbb{R}$ (a *potential*), we can shape a reward function R with $R'(s, a, s') := R(s, a, s') + \gamma\Phi(s') - \Phi(s)$. This is called *potential shaping* [13] and R again induces the same policy ordering as R'. R and R' can thus be considered *equivalent* in a very strong sense. We get additional equivalence relations if we e.g. fix the transition dynamics, see appendix B.

Existing reward distance metrics EPIC [5] was the first distance metric invariant under potential shaping and positive scaling. It uses a three-step process: Given two reward functions R_1 and R_2, EPIC (1) applies a mapping C to both reward functions that is invariant under potential shaping, the "canonicalization", (2) normalizes $C(R_1)$ and $C(R_2)$ under a weighted L^2 norm, and (3) takes a weighted L^2 distance between these canonicalized and normalized reward functions. We can write this formally as

$$D(R_1, R_2) = \left\| \frac{C(R_1)}{\|C(R_1)\|_2} - \frac{C(R_2)}{\|C(R_2)\|_2} \right\|_2 . \tag{1}$$

The L^2 norms are weighted using a coverage distribution $\mathcal{D}$ over transitions. The canonicalization C is defined as

$$C_{\text{EPIC}}(R)(s, a, s') := R(s, a, s') + \mathbb{E}\left[\gamma R(s', A, S') - R(s, A, S') - \gamma R(S, A, S')\right] , \tag{2}$$

where the expectation is over $S, A, S' \sim \mathcal{D}_\mathcal{S} \times \mathcal{D}_\mathcal{A} \times \mathcal{D}_\mathcal{S}$ for some distribution $\mathcal{D}_\mathcal{S}$ over states and $\mathcal{D}_\mathcal{A}$ over actions. If R_1 and R_2 are related by potential shaping, then $C(R_1) = C(R_2)$, so $D(R_1, R_2) = 0$.

More recently, DARD [19] aims to address an important limitation of EPIC, namely that the expectation in eq. (2) treats the next state S' as independent from S. This means EPIC cannot encode much knowledge about transition dynamics into the distribution used for canonicalization. DARD allows S' to depend on S and is otherwise identical to EPIC except for a slight difference in the canonicalization function C.

3 Limitations of existing distance metrics

As already mentioned, and as [19] point out, EPIC is limited by only allowing distributions for canonicalization where S' is independent of S. One subtlety is that the distribution $\mathcal{D}$ used to compute norms in eq. (1) can be arbitrary and thus encode more information about the environment dynamics. However, if $\mathcal{D}$ is chosen to be different from the canonicalization distribution $\mathcal{D}_\mathcal{S} \times \mathcal{D}_\mathcal{A} \times \mathcal{D}_\mathcal{S}$, then the main theoretical result on EPIC, a certain regret bound [5], no longer holds. In fact, EPIC can display pathological behavior in this setting, as we illustrate in fig. 1: EPIC distances can be dominated by the effect of transitions that have probability zero under the coverage distribution, since canonicalization takes them into account anyway. This can lead to EPIC distances arbitrarily close to zero for very dissimilar reward functions, or arbitrarily close to one for equivalent reward functions.

By making S' depend on S in eq. (2), DARD can make the coverage and canonicalization distribution the same while still encoding useful information about environment dynamics. However, DARD in fact uses some transitions that are *not* sampled from the specified canonicalization distribution. Because of that, the examples from fig. 1 apply to DARD without any modifications. Consequently, the DARD paper does not include a regret bound, just like EPIC with differing coverage and canonicalization distribution. An additional issue specific to DARD is that the canonicalized reward function $C_{\text{DARD}}(R)$ is generally *not* in the same potential shaping equivalence class as R.

A separate limitation of both EPIC and DARD is that they are only invariant under potential shaping and positive linear scaling, and no other equivalence relations. But for e.g. fixed environment dynamics, there are other transformations that leave optimal policies unchanged [16]. So given some knowledge about the transition dynamics, like Wulfe et al. [19] assume, we would ideally like reward distance metrics to be invariant to these additional transformations.

In the remainder of this paper, we introduce a much more general framework for reward distance metrics, which addresses all of these limitations.

4 Generalized framework: EPIC-like distances

At a very high level, our framework uses the same three steps as EPIC and DARD: canonicalization, normalization, and taking a distance. However, we significantly generalize every single one of these steps. First, let us specify what we in general mean by "canonicalization":

Definition 1. We say that $R_1 \sim_{PS} R_2$ if R_1 and R_2 differ by potential shaping, and that $R_1 \sim_{S'R} R_2$ if $\mathbb{E}_{S' \sim \tau(s,a)}[R_1(s,a,S')] = \mathbb{E}_{S' \sim \tau(s,a)}[R_2(s,a,S')]$. Given an equivalence relation $\sim$ between reward functions, a function $c : \mathcal{R} \to \mathcal{R}$ is a *canonicalization function* for $\sim$ if and only if (1) $c(R_1) = c(R_2) \iff R_1 \sim R_2$, and (2) $c(R) \sim R$ for all reward functions $R \in \mathcal{R}$.

For now, we will only consider canonicalization for $\sim_{PS}$ or $\sim_{S'R}$ (or both), but other options are considered in Appendix B. EPIC and DARD both canonicalize for $\sim_{PS}$. However, note that DARD violates (2), and so is not technically a canonicalization. We can canonicalize for $\sim_{S'R}$ using e.g. $c(R)(s,a,s') = \mathbb{E}_{S' \sim \tau(s,a)}[R(s,a,S')]$, and for both $\sim_{PS}$ and $\sim_{S'R}$ by combining this with e.g. C_{EPIC}. Note that we must know τ to canonicalize for $\sim_{S'R}$, but that no knowledge of τ is needed to canonicalize for $\sim_{PS}$.

For normalization, EPIC and DARD use a weighted L^2 norm. We generalize this too:

Definition 2. A function $n : \mathcal{R} \to \mathbb{R}$ is a *normalization function* if (1) $n(R) = 0 \iff R = 0$, and (2) $n(\alpha \cdot R) = \alpha \cdot n(R)$ for any $\alpha \geq 0$.

Note that any norm is a normalization function, though normalization functions are more general.

Finally, EPIC and DARD also use the L^2 norm for the final step, measuring distances. We generalize this to allow any function that bounds some norm. This gives us the class of reward function distance metrics that we are considering:

Definition 3. An *EPIC-like distance* is a function $d : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ of the form

$$d(R_1, R_2) = m(s(R_1), s(R_2)) = m\left(\frac{c(R_1)}{n(c(R_1))}, \frac{c(R_2)}{n(c(R_2))}\right) \tag{3}$$

where c is a canonicalization function, n is a normalization function, and m gives an upper bound for some norm. That is, there is a norm p and a constant K_m such that $p(R_1, R_2) \leq K_m \cdot m(R_1, R_2)$ for all $R_1, R_2 \in \text{Im}(s)$. $s(R) := \frac{c(R)}{n(c(R))}$ is the corresponding *standardization function*.

Note that d is generally not a pseudo-metric, like EPIC, but does become a pseudo-metric if m is one. In Appendix C, we discuss several examples of EPIC-like distances, including with canonicalization functions for all equivalence relations analyzed by [16]. In Appendix F, we show that all EPIC-like distances are topologically equivalent.

5 Regret Bound

Next, we show that *any* distance covered by Definition 3 induces a regret bound:

Theorem 1. *Let d be an EPIC-like distance, given by c, n, and m. Suppose the equivalence relation $\sim$ of c satisfies that, if $R_1 \sim R_2$, then there is a k such that $J_1(\pi) = J_2(\pi) + k$ for all π. Suppose there is a constant K_n such that $n(c(R)) \leq K_n \cdot n(R)$ for all R. Then there is a constant K_d such that for any rewards R_1, R_2, and any policies π_1, π_2, if $J_2(\pi_2) \geq J_2(\pi_1)$ then*

$$J_1(\pi_1) - J_1(\pi_2) \leq K_d \cdot n(R_1) \cdot d(R_1, R_2). \tag{4}$$

This theorem establishes a bound on the regret that is incurred under reward R_1 if a policy π_1 is improved to π_2 under a different reward R_2. Note that, as a special case, we could let π_i be an optimal policy under R_i; this way, we recover the regret bound proven by [5]. But Theorem 1 is more general both in that it does not assume optimality, and in that it applies to *any* EPIC-like distance, rather than only EPIC. The proof is given in the appendix.

Note that $\sim_{PS}$ always satisfies that, if $R_1 \sim_{PS} R_2$, then there is a k such that $J_1(\pi) = J_2(\pi) + k$ for all π, whereas $\sim_{S'R}$ only satisfies this condition for the τ it is defined against.

6 How to pick a canonicalization function

To design an EPIC-like distance, we must choose a canonicalization function c, a normalization function n, and the distance m. A reasonable default for m and n is e.g. a weighted L^2-norm. In this section, we discuss how to choose the canonicalization function, c.

We first need to decide which *equivalence relation* to canonicalize for. A reasonable desideratum is that $s(R) = s(R')$ if and only if R and R' induce the same policy ordering over some class of environments.[2] If we do not want to make any assumptions about the transition dynamics, then potential shaping and positive scaling are the only transformations that preserve policy ordering [15]. This still holds for certain limited partial knowledge about transition dynamics [8]. If we *do* know the transition dynamics then $\sim_{S'R}$ also preserves policy ordering, but no other transformations do [15]. The normalization step with n takes care of positive scaling. Therefore, our canonicalization function c should always canonicalize with respect to $\sim_{PS}$, and additionally with respect to $\sim_{S'R}$ if we know the transition dynamics τ. Note that these are the optimal choices for both cases — there is no larger equivalence class which would satisfy the conditions of Theorem 1.

After deciding on an equivalence class, we need to pick a specific canonicalization function. One approach is to pick a canonicalization which makes the regret bound from Theorem 1 as low as possible. As we discuss in appendix E, the constant K_d in eq. (4) is proportional to $K_n := \sup_R \frac{n(c(R))}{n(R)}$. The best value we can achieve is $K_n = 1$, since $c(c(R)) = c(R)$. So it makes sense to pick a *minimal* canonicalization c that achieves this bound if possible, i.e. $n(c(R)) \leq n(R)$ for all $R \in \mathcal{R}$. As we show in Appendix B, a minimal canonicalization always exists if n is continuous.

Minimal canonicalization functions are not always unique, but do turn out to be unique in many important cases. For example, if n is the L^2-norm, then there is a unique minimal canonicalization for potential shaping, which can be interpreted as the unique "divergence-free" reward function in its equivalence class[8]. Note that EPIC does not use this minimal canonicalization function. On the other hand, EPIC's canonicalization function is cheaper to compute than divergence-free canonicalization (which requires solving a linear problem). We hope future work will find choices that are as easy to compute as the EPIC canonicalization but as principled as the minimal one.

7 Conclusion

We have defined a general framework for reward function distance metrics that are invariant under certain equivalence relations. For metrics in this framework, we generalized the regret bound from Gleave et al. [5]. We also gave many examples for possible instantiations of this framework. Which of these many distance metrics should be used in practice is a question that future work will need to investigate empirically. Our contribution is to show that many viable alternatives for existing metrics exist and even have important advantages. We thus highlight the need for such future investigations and provide a framework in which they can take place.

[2] Note that $d(R, R') = 0 \iff s(R) = s(R')$ if m is a metric.

References

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

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

[3] Haoyang Cao, Samuel N. Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. In *NeurIPS*, 2021.

[4] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

[5] Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *ICLR*, 2021.

[6] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018.

[7] Erik Jenner and Adam Gleave. Preprocessing reward functions for interpretability. In *NeurIPS Cooperative AI workshop*, 2021.

[8] Erik Jenner, Herke van Hoof, and Adam Gleave. Calculus on MDPs: Potential shaping as a gradient, 2022. URL `https://arxiv.org/abs/2208.09570`.

[9] Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. Reward identification in inverse reinforcement learning. In *ICML*, 2021.

[10] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, 2018.

[11] Eric J. Michaud, Adam Gleave, and Stuart Russell. Understanding learned reward functions. In *NeurIPS Deep RL workshop*, 2020.

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

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

[14] Jacob Russell and Eugene Santos. Explaining reward functions in markov decision processes. In *International Florida Artificial Intelligence Research Society Conference*, 2019.

[15] Joar Skalse and Alessandro Abate. Misspecification in inverse reinforcement learning, forthcoming.

[16] Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning, 2022.

[17] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2020.

[18] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback, 2021.

[19] Blake Wulfe, Logan Michael Ellis, Jean Mercat, Rowan Thomas McAllister, and Adrien Gaidon. Dynamics-aware comparison of learned reward functions. In *ICLR*, 2022.

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

A Relation to existential risk reduction

It seems likely that we will at some point in the foreseeable future develop AI systems that would in principle be able to disempower humanity, if they chose to. This means we need to ensure that AI systems are aligned with our goals, in order to avoid an existential catastrophe from conflict with misaligned AI.

Reward learning is currently the leading approach for aligning AI systems with complex human goals, rather than simplistic objectives that would be unsafe to optimize. Although more sophisticated approaches may be required to align sufficiently advanced AI systems, many proposals for these still require reward learning as a subcomponent [10, 6, 18].

An important obstacle to alignment, including to the reward learning approach, is Goodhart's law. Although optimizing a (learned) proxy reward function will at first improve outcomes to our actual preferences, at some point further optimization of the proxy will lead to *worse* outcomes. Current reward models are often not sufficiently good proxies to withstand even moderately strong optimization pressure. For example, KL penalties have to be used when fine-tuning language models using RL, in order to avoid reward hacking [17].

In order to develop more robust reward models, which can be optimized safely to a larger extent, good evaluation of reward learning algorithms is crucial. As argued by Gleave et al. [5], simply evaluating the performance policies optimized using the learned reward models has numerous drawbacks. Using reward distance metrics instead can give a more reliable evaluation criteria. Although EPIC and DARD provide a good starting point as reward distance metrics, there is probably room for improvement, given how new this subfield is. From the perspective of Goodhart's law, an especially important aspect is the regret bound, which we significantly generalize compared to the one given in the EPIC paper. We hope that the framework we introduce will lead to a more comprehensive study of reward distance metrics, and ultimately to better evaluation of learned reward models and of reward learning algorithms.

B Generalized Canonicalization Functions

In the main paper, we have mainly used canonicalization functions that remove potential shaping by mapping all reward functions that differ by potential shaping to a single representative in their equivalence class. In this section, we will show how to canonicalize even broader equivalence classes, and thereby achieve even tighter regret bounds.

If two reward functions R_1 and R_2 induce the same ordering of policies, then they should be considered *equivalent*, and have $d(R_1, R_2) = 0$. Moreover, it has been shown that R_1 and R_2 induce the same ordering of policies for all transition functions τ if and only if R_1 and R_2 differ by *potential shaping* and *positive linear scaling* (see Skalse and Abate [15]). We therefore use a canonicalization function c to remove potential shaping, and a normalisation function n to remove positive scaling. This ensures that $s(R_1) = s(R_2)$ if and only if R_1 and R_2 are equivalent for all τ.

If we do not wish to make any assumptions about τ, then this is the best we can do. That is, if R_1 and R_2 do not differ by potential shaping and positive linear scaling, then there is some τ for which they induce different policy orderings. However, can we do better than this if we know τ? The answer is yes; if $\mathbb{E}_{S' \sim \tau(s,a)}[R_1(s,a,S')] = \mathbb{E}_{S' \sim \tau(s,a)}[R_2(s,a,S')]$, then R_1 and R_2 induce the same ordering of policies under τ. Following Skalse et al. [16], we refer to this as S'*-redistribution*. We can canonicalize for S'-redistribution using e.g. $c(R)(s,a,s') = \mathbb{E}_{S' \sim \tau(s,a)}[R(s,a,S')]$. Is it possible to do even better than this? The answer is no:

Proposition 1. *Two reward functions R_1 and R_2 induce the same ordering of policies if and only if they differ by potential shaping, S'-redistribution, and positive linear scaling.*

Proof. See Skalse and Abate [15]. □

This means that if we use a canonicalization function c that removes potential shaping and S'-redistribution, and a normalisation function n that removes positive scaling, then $s(R_1) = s(R_2)$ if and only if R_1 and R_2 are equivalent under τ. In other words, if we do not know τ, then the best we can do is to canonicalize for potential shaping, and normalise for positive scaling, and if

we know τ, then the best we can do is to additionally canonicalize for S'-redistribution. However, if we have *partial* knowledge of τ, then there may be more we can do. For example, if we know that all transitions in some set X are impossible under τ, then we could set $R(s, a, s') = 0$ for all $(s, a, s') \in X$. This calls for a more general definition of canonicalization functions:

Definition 4. An *environment* E consists of a transition function τ and initial state distribution μ_0. Say that $R_1 \sim R_2$ in $E = (\tau, \mu_0)$ if there is a constant k such that $J_1(\pi) = J_2(\pi) + k$ for all π, where J_1 and J_2 are evaluated under τ and μ_0. We say that $c : \mathcal{R} \to \mathcal{R}$ is a canonicalization function for E if $c(R_1) = c(R_2)$ implies $R_1 \sim R_2$ in E, and $R \sim c(R)$ in E. Moreover, we say that c is a canonicalization function for a set of environments X if it is a canonicalization function for each environment in X.

This definition allows for canonicalization functions that take into account as much information about τ as possible. Canonicalizing larger sets of reward functions can make the resulting regret bound tighter, but it will also make it applicable in fewer environments.

Next, the equivalence classes we have discussed can be canonicalized with a linear transformation:

Proposition 2. *There exists a linear canonicalization function for potential shaping, S'-redistribution, and impossible transitions.*

Proof. The obvious way to canonicalise impossible transitions X is to set

$$c(R)(s, a, s') = 0 \text{ for } \langle s, a, s' \rangle \in X.$$

This is a linear transformation. The obvious way to canonicalize for S'-redistribution is to set

$$c(R)(s, a, s') = \mathbb{E}_{S' \sim \tau(s,a)}[R(s, a, S')].$$

This is also a linear transformation. Finally, the canonicalization function C_{EPIC} used by EPIC is an example of a linear function that canonicalizes for potential shaping. □

Our proofs will not require that c is linear, but it is good to know that a linear c exists. We next introduce a notion of "minimal" canonicalization functions. This is motivated by the following:

Proposition 3. *If n is continuous, and if S is an affine subspace of $\mathcal{R}$, then there is an $R_S \in S$ such that $n(R_S) \leq n(R)$ for all $R \in S$.*

Proof. Let $N \subseteq \mathcal{R}$ be the unit ball of n. If n is continuous then there is a unique smallest value $c \in \mathbb{R}$ such that $c \cdot N$ intersects S. Any reward function R_S in this intersection satisfies $n(R_S) \leq n(R)$ for all $R \in S$. □

Note that R_S may not be unique! Note also that if n is a norm then n is continuous, since $\mathcal{R}$ is finite-dimensional. Next, note that for any canonicalisation function c for one of the equivalence classes we have discussed, and any reward function R, the set of all reward functions R' such that $c(R') = c(R)$ forms an affine space. Proposition 3 thus implies that we, given a continuous normalisation function n, can define a (not necessarily unique) canonicalisation function c which sends each R to (one of) the smallest reward functions (as measured by n) in its equivalence class.

Definition 5. Given a normalisation function n, a canonicalisation function c is *minimal* if $n(c(R)) \leq n(R')$ for all R' such that $c(R) = c(R')$.

Note that Proposition 3 implies that there always exists a minimal canonicalisation function for any continuous n. Note also that this function may or may not be unique; for example, it is unique if n is the L^2-norm, but not if it is the L^∞-norm.

C Examples

In this section, we give a few examples of EPIC-like distances. Our first example is of course EPIC, whose canonicalisation function c is C_{EPIC}, whose normalisation function n is the $\mathcal{D}$-weighted L^2 norm, and whose m is also the $\mathcal{D}$-weighted L^2 norm. Next, our definition of an EPIC-like distance allows each of these parts to be changed. For example, we could normalise using the L^1 norm, or the L^∞ norm, and so on.

There are also many choices for the canonicalisation function. For example, we can generalize EPIC by shaping with the value function of an arbitrary policy π under arbitrary transition dynamics τ: $c(R)(s,a,s') = R(s,a,s') + \gamma v^\pi(s') - v^\pi(s)$. EPIC is the special case where $\pi(a|s) = \mathcal{D}_\mathcal{A}(a)$ and $\tau(s'|s,a) = \mathcal{D}_\mathcal{S}(s')$. So just like DARD, this is a strict generalization of EPIC, but it avoids the pathologies of DARD discussed in section 3.

To formalize this example, we can define the outflow of a reward function as $\mathrm{out}(R) : \mathcal{S} \to \mathcal{S}$ with

$$\mathrm{out}(R)(s) := \sum_{a,s'} \mathcal{D}(s,a,s') R(s,a,s') \tag{5}$$

for a coverage distribution $\mathcal{D}$. Following Jenner et al. [8], we will also write

$$\mathrm{grad}(\Phi)(s,a,s') := \gamma\Phi(s') - \Phi(s) \tag{6}$$

for the potential shaping term. We then have the following result:

Proposition 4. *Let $\mathcal{D}$ be any distribution over transitions, R a reward function, and $\gamma < 1$. Then there is a unique potential Φ such that $\mathrm{out}(R + \mathrm{grad}\,\Phi) = 0$ and $\Phi(s) = 0$ for all s for which the marginal probability is zero: $\sum_{a,s'} \mathcal{D}(s,a,s') = 0$.*

Proof. Writing out the condition $\mathrm{out}(R + \mathrm{grad}\,\Phi) = 0$ explicitly, we get

$$\sum_{a,s'} \mathcal{D}(s,a,s')(R(s,a,s') + \gamma\Phi(s') - \Phi(s)) = 0\,. \tag{7}$$

We can rearrange this slightly to obtain

$$\Phi(s) \sum_{a,s'} \mathcal{D}(s,a,s') = \sum_{a,s'} \mathcal{D}(s,a,s')(R(s,a,s') + \gamma\Phi(s'))\,. \tag{8}$$

As a sidenote to provide intuition, note what happens if $\sum_{a,s'} \mathcal{D}(s,a,s') > 0$ for all states s: we can divide by this and get

$$\Phi(s) = \sum_{a,s'} \mathcal{D}(a,s'|s)(R(s,a,s') + \gamma\Phi(s'))\,, \tag{9}$$

which is precisely the Bellman equation for the value function v^π if we factorize $\mathcal{D}(a,s'|s) =: \pi(a|s)p(s'|a,s)$ (which is always uniquely possible). So the unique potential Φ that makes $R + \mathrm{grad}\,\Phi$ outflow-free would then be simply the value function.

Intuitively, if $\sum_{a,s'} \mathcal{D}(s,a,s') = 0$ for some state s, then we're missing the information necessary to compute the value of s, and $\Phi(s)$ thus becomes arbitrary. By fixing the potential of these states to zero, we recover a unique solution. This is similar to setting the value of terminal states to zero for solving the Bellman equation. The remainder of the proof essentially formalizes this idea.

We interpret the potential Φ as a vector in $\mathcal{R}^n$, where n is the number of states. Without loss of generality, we order the states $\mathcal{S} = \{s_1, \ldots, s_n\}$ such that $\mathcal{D}(s_i) > 0$ for $i \leq m$ and $\mathcal{D}(s_i) = 0$ for $i > m$ (where m is the number of states with non-zero marginal).

We then additionally define the expected reward vector $R^+ \in \mathbb{R}^m$ as

$$R^+_s := \sum_{a,s'} \mathcal{D}(a,s'|s) R(s,a,s') \tag{10}$$

for $s \in \{s_1, \ldots, s_m\}$, and the transition matrix $P \in \mathbb{R}^{m \times n}$ as

$$P_{ss'} := \sum_a \mathcal{D}(s',a|s)\,. \tag{11}$$

Finally, we split up Φ and P into two parts at the index m:

$$\Phi = \begin{pmatrix} \Phi^+ \\ \Phi^0 \end{pmatrix} \tag{12}$$

and

$$P = \begin{pmatrix} P^+ & P^0 \end{pmatrix}\,. \tag{13}$$

So e.g. Φ^+ is the potential of all states s with $\mathcal{D}(s) > 0$.

Note that the outflow condition eq. (8) above always holds for states with $\mathcal{D}(s) = 0$. So a potential Φ^+ leads to an outflow-free reward if and only if eq. (8) holds for $s \in \{s_1, \dots, s_m\}$. We can write these m linear equations in the n unknowns Φ in matrix form as follows:

$$\Phi^+ = R^+ + \gamma P \Phi = R^+ + \gamma (P^+ \Phi^+ + P^0 \Phi^0) \,. \tag{14}$$

Here we made use of the fact that for these m equations, we *can* divide by $\mathcal{D}(s)$, as discussed above. We now rearrange this as

$$(I - \gamma P^+)\Phi^+ = R^+ + \gamma P^0 \Phi^0 \,. \tag{15}$$

Now note that each row in P^+ sums to at most one (since each row in P sums to exactly one). This means that all eigenvalues of P^+ have absolute value at most one, i.e. we have $\rho(P^+) \leq 1$ for the spectral radius of P^+. Thus, $\rho(\gamma P^+) \leq \gamma < 1$, so $I - \gamma P^+$ is invertible (with the inverse given by a Neumann series). This means that for any choice of Φ^0, there is a unique solution for Φ^+ to satisfy the outflow condition. □

The corresponding canonicalization is then $c(R) := R + \operatorname{grad} \Phi$. As mentioned in the proof, the potential Φ that is used to obtain $c(R)$ from R can be interpreted as the value function v^π of a policy and transition dynamics implicitly defined by $\mathcal{D}$. This is also the reason that a unique $c(R)$ exists—this is equivalent to saying that the Bellman equation for the value function of a policy always has a unique solution. Setting Φ to zero for states with zero marginal probability is somewhat arbitrary: for any choice of potential on these states, there is exactly one way to fill in the remaining potential to make the reward outflow-free. Zero is merely the simplest choice. The condition $\gamma < 1$ could also be replaced with a suitable condition on the canonicalization distribution, similar to how the Bellman equation can be uniquely solved if either $\gamma < 1$ or if all trajectories end in an absorbing state with probability one.

We can assume without much of a restriction that all marginal probabilities are non-zero (otherwise, we could simply remove the zero-probability states altogether). Then proposition 4 simplifies:

Corollary 1. *Assume that $\sum_{a,s'} \mathcal{D}(s, a, s') > 0$ for all states s. Then every potential shaping equivalence class of reward functions has exactly one outflow-free representative.*

As already mentioned in the main paper, another example of a canonicalization function for potential shaping is divergence-free canonicalization [8].

[16] also discuss constant shifts (a special case of potential shaping) and masks of invalid transitions. These are similarly easy to canonicalize, for example using $C(R) := R - \min R$ for constant shift, and by setting rewards of invalid transitions to zero for masking.

As a slightly more complex example, consider monotonic transformations of reward functions, i.e. $R_1 \sim R_2$ if $R_1(x) \leq R_1(x') \iff R_2(x) \leq R_2(x')$ for any transitions x, x'. One way to canonicalize this equivalence relation is to set $C(R)(x) := \#\{\tilde{x} \mid R(\tilde{x} \leq R(x)\}$. This discards the exact magnitudes of rewards and picks a reasonable canonical representation of their ordering.

We can also canonicalize the final equivalence relation discussed by Skalse et al. [16]: R_1 and R_2 are related by an *optimality-preserving* transformation for given transition dynamics and initial state distribution if and only if they both induce the same optimal policy set. Note that this is the case precisely when the optimal Q-functions of both rewards have the same maximizing actions in each state. So to canonicalize a reward function R, we can first canonicalize its optimal Q-function $Q^\star(s, \cdot)$ for each state s, e.g. by setting it to 1 for optimal actions and 0 for non-optimal ones. This canonicalized optimal Q-function specifies a reward up to S'-redistribution [16, Theorem 3.1], so we can use the canonicalization for S'-redistribution from above to obtain a canonicalized reward function.

An important note is that the regret bound does not apply to canonicalization of monotonic and optimality-preserving transformations, since these can affect returns in other ways than just by adding a constant. Distances using these canonicalizations should thus only be used in applications where no regret bound is needed. In exchange, these canonicalizations can recognize even more reward functions as equivalent, which may be desirable in some cases.

C.1 Interesting choices of m

It is possible to get quite creative with the choice of m. Consider:

Definition 6. Given a linear canonicalisation function c, let the angle-based metric ANG^c be

$$\mathrm{ANG}^c(R_1, R_2) = \mathrm{ang}(c(R_1), c(R_2)),$$

where $\mathrm{ang} : \mathbb{R}^n \times \mathbb{R}^n \to [0, \pi]$ is the function that returns the angle between non-zero vectors. For the zero vector, let $\mathrm{ang}(\vec{0}, \vec{0}) = 0$ and $\mathrm{ang}(\vec{0}, R) = \pi/3$ for $R \neq \vec{0}$.

It may not be immediately obvious why ANG^c is an EPIC-like distance. To see this, first note that ANG^c is unaffected by the scale of R_1 and R_2. We may therefore equivalently express it as $\mathrm{ang}(s(R_1), s(R_2))$, where s canonicalises using c and normalises using the L^2 norm. Next, note that we now have that $L^2(R_1, R_2) = 2\sin(\mathrm{ang}(R_1, R_2)/2)$ for all $R_1, R_2 \in \mathrm{Im}(s)$. This means that ang bounds L^2 (using $K_m = 1$), and so ANG^c is an EPIC-like distance. Next, let us give another example of a "creative" choice of m:

Definition 7. Let s be a standardisation function, and let $\mathcal{D}$ be a distribution over $(\mathcal{S}\times\mathcal{A}\times\mathcal{S})^\star$ that gives support to all possible transitions. Then the sample-based distance measure $\mathrm{SAM}^{s,\mathcal{D}}$ is

$$\mathrm{SAM}^{s,\mathcal{D}}(R_1, R_2) = \mathbb{E}_{\xi\sim\mathcal{D}}|G_1^S(\xi) - G_2^S(\xi)|.$$

Here m is implicitly given by the L^1-norm, weighted by the discounted cumulative probability with which $\mathcal{D}$ visits each transition. In other words, $W(s, a, s') = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\xi\sim\mathcal{D}}(s_t, a_t, s_{t+1} = s, a, s')$. If we are only concerned with a single transition function τ, then $\mathcal{D}$ could be the trajectory distribution of any policy that visits all reachable transitions with positive probability. If we are concerned with all transition functions, then $\mathcal{D}$ must give positive support to all transitions (though not necessarily at every time step). Note that we do not need to assume that $\mathcal{D}$ corresponds to a policy; it could be non-stationary, give support to impossible transitions, or even support trajectories that do not form a path, etc. We will give one final example of a creative EPIC-like distance:

Definition 8. Given a fixed τ, μ_0, and γ, let s be the standardisation function that normalises using $n(R) = \max_\pi J(\pi) - \min_\pi J(\pi)$, and canonicalizes with a function c that ensures that $\min_\pi J(\pi) = 0$. Then the continuous order-based distance measure CORD is given by

$$\mathrm{CORD}(R_1, R_2) = \max_\pi |J_1^s(\pi) - J_2^s(\pi)|.$$

CORD can be seen as a continuous measure of the extent to which the policy orderings of R_1 and R_2 differ. If J is normalised to $[0, 1]$, then $J^s(\pi)$ can be seen as an index corresponding to π's place in the ordering. CORD then gives the maximal value by which the index of any policy differs between R_1 and R_2. To see that CORD is EPIC-like, note that if π is a policy that visits every transition with positive probability, then $p(R_1, R_2) = |J_1(\pi) - J_2(\pi)|$ is a norm (equivalent to a weighted L^1-norm, and ignoring the transitions that are impossible under τ), and that $p \leq m$.

D Regret Bound Proofs

First, we will establish a few general lemmas that concern how we can bound the regret between different reward functions in various ways. To start with, note that the difference in reward obtained by some particular policy π under two different reward functions (under arbitrary transition dynamics) can be bounded in terms of the L^∞-distance between those reward functions.

Lemma 1. *For any reward functions R_1 and R_2, and any policy π, we have*

$$|J_1(\pi) - J_2(\pi)| \leq \left(\frac{1}{1-\gamma}\right) L^\infty(R_1, R_2).$$

Proof.

$$
\begin{aligned}
|J_1(\pi) - J_2(\pi)| &= |\mathbb{E}_{\xi\sim\pi}[\sum_{t=0}^{\infty}\gamma^t R_1(s_t,a_t,s_{t+1})] - \mathbb{E}_{\xi\sim\pi}[\sum_{t=0}^{\infty}\gamma^t R_2(s_t,a_t,s_{t+1})]| \\
&= \sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{\xi\sim\pi}[|R_1(s_t,a_t,s_{t+1}) - R_2(s_t,a_t,s_{t+1})|] \\
&\leq \sum_{t=0}^{\infty}\gamma^t L^\infty(R_1,R_2) = \left(\frac{1}{1-\gamma}\right) L^\infty(R_1,R_2).
\end{aligned}
$$

□

This is the tightest bound possible, if all we know about R_1 and R_2 is their L^∞-distance. However, we could compute an even tighter bound, if we examine R_1 and R_2 more closely (e.g., by tracing the maximal difference between them along any path through $\mathcal{S}$ and $\mathcal{A}$). Nonetheless, this bound will be sufficient for our purposes. We next show that there is a similar bound for *any* norm.

Lemma 2. *If p is a* norm *then there is a constant K_p such that $|J_1(\pi) - J_2(\pi)| \leq K_p \cdot p(R_1,R_2)$.*

Proof. If p and q are norms on a finite-dimensional vector space, then there are constants k and K such that

$$k \cdot p(x) \leq q(x) \leq K \cdot p(x).$$

Since $\mathcal{S}$ and $\mathcal{A}$ are finite, $\mathcal{R}$ is a finite-dimensional vector space. This means that there is a constant K such that $L^\infty(R_1,R_2) \leq K \cdot p(R_1,R_2)$. Together with Lemma 1, this implies that

$$|J_1(\pi) - J_2(\pi)| \leq \left(\frac{1}{1-\gamma}\right) \cdot K \cdot m(R_1,R_2).$$

Letting $K_p = \left(\frac{K}{1-\gamma}\right)$ completes the proof. □

Next, we derive a lemma that allows us to go from bounds of this form to regret bounds.

Lemma 3. *Let R_1 and R_2 be reward functions, and π_1 and π_2 be policies. If $|J_1(\pi) - J_2(\pi)| \leq U$ for $\pi \in \{\pi_1,\pi_2\}$, and if $J_2(\pi_2) \geq J_2(\pi_1)$, then we have that*

$$J_1(\pi_1) - J_1(\pi_2) \leq 2 \cdot U.$$

Proof. First note that U must be non-negative. Next, note that if $J_1(\pi_1) < J_1(\pi_2)$ then $J_1(\pi_1) - J_1(\pi_2) \leq 0$, and so the lemma holds. Now consider the case when $J_1(\pi_1) \geq J_1(\pi_2)$:

$$
\begin{aligned}
J_1(\pi_1) - J_1(\pi_2) &= J_1(\pi_1) - J_2(\pi_2) + J_2(\pi_2) - J_1(\pi_2) \\
&\leq |J_1(\pi_1) - J_2(\pi_2)| + |J_2(\pi_2) - J_1(\pi_2)|
\end{aligned}
$$

Our assumptions imply that $|J_2(\pi_2) - J_1(\pi_2)| \leq U$. We will next show that $|J_1(\pi_1) - J_2(\pi_2)| \leq U$ as well. Our assumptions imply that

$$
\begin{aligned}
&|J_1(\pi_1) - J_2(\pi_1)| \leq U \\
\implies &J_2(\pi_1) \geq J_1(\pi_1) - U \\
\implies &J_2(\pi_2) \geq J_1(\pi_1) - U
\end{aligned}
$$

Here the last implication uses the fact that $J_2(\pi_2) \geq J_2(\pi_1)$. A symmetric argument also shows that $J_1(\pi_1) \geq J_2(\pi_2) - U$ (recall that we assume that $J_1(\pi_1) \geq J_1(\pi_2)$). Together, this implies that $|J_1(\pi_1) - J_2(\pi_2)| \leq U$. We have thus shown that if $J_1(\pi_1) \geq J_1(\pi_2)$ then

$$|J_1(\pi_1) - J_2(\pi_2)| + |J_2(\pi_2) - J_1(\pi_2)| \leq 2 \cdot U,$$

and so the lemma holds. This completes the proof. □

This lemma establishes a bound on the regret that is incurred under reward R_1 if a policy π_1 is optimised to π_2 under a different reward R_2, making no assumptions about how much optimisation is performed. Here U could come from Lemma 1 or 2, or it could be derived in some other way.

D.1 Removing Standardisation

Let us now briefly recall our strategy for constructing reward distance functions. We have seen that if p is a norm (or gives an upper bound for a norm) then $p(R_1, R_2)$ gives a regret bound for R_1 and R_2 that holds in all environments. Moreover, since distinct reward functions can be *equivalent* in various classes of environments, we can improve on the bound provided by $p(R_1, R_2)$ by creating a standardisation function s, which maps all reward functions in the same equivalence class to a single representative, and then measure the distance $p(s(R_1), s(R_2))$. This gives us a regret bound for $s(R_1)$ and $s(R_2)$. To give a full regret bound, we additionally need a way to bound the regret of R_1 and R_2 in terms of the regret of $s(R_1)$ and $s(R_2)$. We will now discuss this issue.

Some care is required when we perform both canonicalisation and normalisation. The reason for this is that the canonicalisation c might (and typically will) change the n-size of the reward vector. We therefore need to bound $n(c(R))$ in terms of $n(R)$. When we have such a bound, we can derive a regret bound with the following strategy:

Lemma 4. *If there is a K_n such that $n(c(R)) \leq K_n \cdot n(R)$ for all R, and the equivalence relation $\sim$ of c satisfies that if $R_1 \sim R_2$ then there is a k such that $J_1(\pi) = J_2(\pi) + k$ for all π, then for any reward R and any policies π, π', we have that*

$$J^S(\pi) - J^S(\pi') \leq U \implies J(\pi) - J(\pi') \leq K_n \cdot n(R) \cdot U.$$

Proof. First recall that $s(R) = \left(\frac{c(R)}{n(c(R))}\right)$. This means that

$$J^S(\pi) = \left(\frac{1}{n(c(R))}\right)(J(\pi) + k),$$

which further implies that

$$J^S(\pi) - J^S(\pi') = \left(\frac{1}{n(c(R))}\right)(J(\pi) - J(\pi'))$$

since the k-terms cancel out. By rearranging, we get that

$$J(\pi) - J(\pi') = n(c(R))(J^S(\pi) - J^S(\pi')).$$

Since $n(c(R)) \leq K_n \cdot n(R)$, and since $J^S(\pi) - J^S(\pi') \leq U$, this implies that

$$J(\pi) - J(\pi') \leq K_n \cdot n(R) \cdot U.$$

This completes the proof. □

This now raises the question; when can we be sure that such a K_n exists? Again, we will show that this exists under very general conditions. We begin by showing that it is sufficient for c to be linear and n to be continuous.

Lemma 5. *If a canonicalisation function c is linear, and a normalisation function n is continuous, then there is a K_n such that $n(c(R)) \leq K_n \cdot n(R)$.*

Proof. We begin by noting that if c is linear, then for any positive $\alpha \in \mathbb{R}$,

$$\frac{n(c(\alpha \cdot R))}{n(\alpha \cdot R)} = \left(\frac{\alpha}{\alpha}\right)\frac{n(c(R))}{n(R)} = \frac{n(c(R))}{n(R)},$$

since n is absolutely homogeneous. The maximal value of $\frac{n(c(R))}{n(R)}$ must therefore occur on the unit ball in n. Next, since the unit ball of n is a compact set, and since $\frac{n(c(R))}{n(R)}$ is continuous, the extreme value theorem implies that $\frac{n(c(R))}{n(R)}$ must take on some maximal value K on this domain. This implies that there is a K such that $n(c(R)) \leq K_n \cdot n(R)$ for all R. □

This is sufficient to ensure that such a constant K_n exists in all cases we are concerned with. It is also worth noting that there are other ways to get a bound on $n(c(R))$ in terms of $n(R)$. For example, if we know τ, μ_0, and γ, and use the normalisation function

$$n(R) = \max_\pi J(\pi) - \min_\pi J(\pi),$$

then $n(c(R)) = n(R)$ for all valid canonicalisation functions. Therefore, in that case we have that $K_n = 1$ for any choice of c (linear or not). Also note that if c is a minimal canonicalisaton for n, then $n(c(R)) \leq n(R)$. Therefore, in these cases we also get a bound with $K_n = 1$. Again, recall that a minimal canonicalisation function always exists for any continuous n.

D.2 Full Regret Bound

We now have all the pieces necessary to prove the regret bound.

Theorem 2. *Let d be an EPIC-like distance, given by c, n, and m. Suppose the equivalence relation $\sim$ of c satisfies that, if $R_1 \sim R_2$, then there is a k such that $J_1(\pi) = J_2(\pi) + k$ for all π. Suppose there is a constant K_n such that $n(c(R)) \leq K_n \cdot n(R)$ for all R. Then there is a constant K_d such that for any rewards R_1, R_2, and any policies π_1, π_2, if $J_2(\pi_2) \geq J_2(\pi_1)$ then*

$$J_1(\pi_1) - J_1(\pi_2) \leq K_d \cdot n(R_1) \cdot d(R_1, R_2).$$

Proof. Recall that $d(R_1, R_2) = m(s(R_1), s(R_2))$. We will begin by establishing a regret bound in terms of the standardised reward functions $s(R_1)$ and $s(R_2)$, and then translate this into a regret bound in terms of R_1 and R_2. To do this, first recall that m bounds some norm p. Since p is a norm, we can apply Lemma 2 to conclude that there is a constant K_p such that for any π,

$$|J_1^S(\pi) - J_2^S(\pi)| \leq K_p \cdot p(s(R_1), s(R_2)).$$

Recall that $p(R_1, R_2) \leq K_m \cdot m(R_1, R_2)$ for some constant K_m. Therefore,

$$\begin{aligned} K_p \cdot p(s(R_1), s(R_2)) &\leq K_p \cdot K_m \cdot m(s(R_1), s(R_2)) \\ &= K_{mp} \cdot d(R_1, R_2) \end{aligned}$$

where $K_{mp} = K_p \cdot K_m$. We have thus established that

$$|J_1^S(\pi) - J_2^S(\pi)| \leq K_{mp} \cdot d(R_1, R_2)$$

for any π. Next, note that if $J_2(\pi_2) \geq J_2(\pi_1)$ then $J_2^S(\pi_2) \geq J_2^S(\pi_1)$. We can therefore apply Lemma 3 and conclude that

$$J_1^S(\pi_1) - J_1^S(\pi_2) \leq 2 \cdot K_{mp} \cdot d(R_1, R_2).$$

We have assumed that there is a constant K_n such that $n(c(R)) \leq K_n \cdot n(R)$ for all R. We can therefore apply Lemma 4, and conclude that

$$J_1(\pi) - J_2(\pi) \leq K_n \cdot n(R_1) \cdot 2 \cdot K_{mp} \cdot d(R_1, R_2).$$

Setting $K_d = 2 \cdot K_n \cdot K_{mp}$ completes the proof. □

E Concrete regret bound constants

Theorem 2 assumes existence of a suitable constant K_n and shows existence of another constant K_p. The specific values of these constants depend on the details of the distance we are using. In this section, we give specific bounds for a few of the examples we have seen.

From the proof of theorem 2, we can see that K_d comes from three constants: $K_d = 2 \cdot K_n \cdot K_m \cdot K_p$. Typically, m will itself be a norm, in which case we can set $p = m$ and get $K_m = 1$. As we can see from the proof of lemma 2, $K_p = \frac{K}{1-\gamma}$, where K is chosen such that $L^\infty(R_1, R_2) \leq K \cdot m(R_1, R_2)$.

If m is the L^∞ norm, we can simply use $K = 1$. Gleave et al. [5, Lemma A.11] show that for m a $\mathcal{D}$-weighted L^1 norm, we need $K\mathcal{D}(s, a, s') \geq 2\mathcal{D}_{\pi,t}(s, a, s')$ for all time steps t and policies π for which we want the regret bound to hold. At the very least, we can achieve $K = 2|\mathcal{S}||\mathcal{A}||\mathcal{S}|$ with a uniform distribution $\mathcal{D}$. For a $\mathcal{D}$-weighted L^2 norm, the same K can be used as long as we normalize $\mathcal{D}$, due to Hölder's inequality.

In summary, we get:

- $K_d = \frac{2K_n}{1-\gamma}$ for an L^∞ norm

- $K_d = \frac{4K_n \cdot K}{1-\gamma}$ with $K\mathcal{D}(s,a,s') \geq \mathcal{D}_{\pi,t}(s,a,s')$ for a $\mathcal{D}$-weighted L^1 or L^2 norm, with $\mathcal{D}$ some probability distribution over transitions.

Now let us discuss the existence of K_n. This will depend on the choice of n and of the canonicalization function c. We will focus on n being a $\mathcal{D}$-weighted L^2 norm, since this is what EPIC uses. We then get the following results:

Proposition 5. *Let n be a $\mathcal{D}$-weighted L^2 norm. Then we have $n(c(R)) \leq K_n \cdot n(R)$ with*

1. $K_n = 4$ *for* $c = C_{EPIC}$
2. $K_n = 1$ *for divergence-free canonicalization*
3. $K_n = \left(1 + \frac{1+\gamma}{1-\gamma}\right)$ *for shaping with a value function (outflow-free canonicalization)*

This assumes in each case that the distribution used for canonicalization is also $\mathcal{D}$. For value-based shaping and divergence-free canonicalization, this is always a valid choice, while for EPIC, it only works if $\mathcal{D}$ has independent state, action, and next state. If a different distribution is used for canonicalization than for computing n, we get additional factors in K_n, which become infinite if the two distributions have different support. In terms of regret bounds, it is thus best to choose both distributions to be the same.

Proof. For EPIC, this claim corresponds to Lemma A.12 in Gleave et al. [5].

For divergence-free canonicalization, $K_n = 1$ suffices because c picks the representative with minimal L^2 norm in its equivalence class, so in particular $\|c(R)\|_2 \leq \|R\|_2$.

Finally, write $C_{\text{out}}(R) = R + \operatorname{grad} \Phi$ and use the notation for Φ^+, P^+, and R^+ introduced in the proof of proposition 4. Then,

$$\|C_{\text{out}}(R)\| \leq \|R\| + \|\operatorname{grad} \Phi\| \leq \|R\| + (1+\gamma)\|\Phi\| \,. \tag{16}$$

A word on $\|\Phi\|$: we can either interpret Φ as a function on transitions, which happens to only depend on s (not a or s'), or we can interpret it as a vector in $\mathbb{R}^n$, as in the proof of proposition 4, in which case we use the marginal of $\mathcal{D}$ to compute the norm. Crucially, the norm is the same in both cases.

As discussed in the proof of proposition 4, Φ can be split into Φ^+ and Φ^0, but the latter was defined to be zero there, so we get $\|\Phi\| = \|\Phi^+\|$. Φ^+ is explicitly given by

$$\Phi^+ = (I - \gamma P^+)^{-1} R^+ \,. \tag{17}$$

We can write the inverse as a Neumann series,

$$(I - \gamma P^+)^{-1} = \sum_{k=0}^{\infty} (\gamma P^+)^k \,, \tag{18}$$

which lets us bound its operator norm as

$$\begin{aligned} \left\|(I - \gamma P^+)^{-1}\right\| &\leq \sum_{k=0}^{\infty} \left\|(\gamma P^+)^k\right\| \\ &\leq \sum_{k=0}^{\infty} \gamma^k \left\|P^+\right\|^k \\ &\leq \sum_{k=0}^{\infty} \gamma^k \\ &= \frac{1}{1-\gamma} \,. \end{aligned} \tag{19}$$

Here, we used first the triangle inequality, then the submultiplicativity of the operator norm, and finally the fact that the operator norm of P^+ is at most 1.

If we combine this bound on the operator norm with eq. (17), we get

$$\left\|\Phi^+\right\| \leq \frac{1}{1-\gamma} \left\|R^+\right\| \leq \frac{1}{1-\gamma} \|R\| \,. \tag{20}$$

There is a slight subtlety here: $\|R^+\|$ is a norm on states, and $\|R\|$ on transitions. We made use of

$$\begin{aligned} \left\|R^+\right\|^2 &= \sum_s \mathcal{D}(s) \left(\sum_{a,s'} \mathcal{D}(a,s'|s) R(s,a,s') \right)^2 \\ &\leq \sum_s \mathcal{D}(s) \sum_{a,s'} \mathcal{D}(a,s'|s) R(s,a,s')^2 \\ &= \|R\|^2 . \end{aligned} \tag{21}$$

Putting everything together, we get

$$\|c(R)\| \leq \|R\| + (1+\gamma)\|\Phi\| \leq \|R\| + \frac{1+\gamma}{1-\gamma}\|R\| . \tag{22}$$

□

F Topological Equivalence Proof

Definition 3 specifies a very diverse class of distance metrics. A natural question is then just how much these distance metrics can vary. In this section, we will partially answer this question, and show that all EPIC-like distances are equivalent in a certain sense.

Theorem 3. *If d_1 and d_2 are continuous EPIC-like distances which canonicalize the same equivalence classes, then there exist positive $L, U \in \mathbb{R}^{>0}$ such that for all R_1, R_2,*

$$L \cdot d_1(R_1, R_2) \leq d_2(R_1, R_2) \leq U \cdot d_1(R_1, R_2).$$

Proof. Let s be an arbitrary continuous standardisation function, which canonicalises for the same equivalence classes as d_1 and d_2. For example, s could correspond to s_1's canonicalisation, and normalisation with L^2. Next, consider the function $f(R_1, R_2) = \frac{d_2(R_1,R_2)}{d_1(R_1,R_2)}$ for $R_1 \neq R_2$, and 0 otherwise. If d_1 and d_2 are continuous, then so is f. Note that if m_2 gives a linear upper bound for some norm then $d_2(R_1, R_2) \neq 0$ whenever $R_1 \neq R_2$, for all $R_1, R_2 \in \mathrm{Im}(s)$. Moreover, if s is continuous then $\mathrm{Im}(s) \times \mathrm{Im}(s)$ is a compact space. This means that we can apply the extreme value theorem, and conclude that f takes on a maximal value U and a minimal value L on $\mathrm{Im}(s) \times \mathrm{Im}(s)$. Multiplying all sides by $d_1(R_1, R_2)$ completes the proof. □

This shows that all EPIC-like distances are topologically equivalent. Therefore, while some EPIC-like distances might induce tighter regret bounds, or be easier or faster to compute, etc, they are ultimately broadly similar, at least in that they induce the same topology on $\mathcal{R}$.