Invariance in Policy Optimisation and Partial Identifiability in Reward Learning

Joar Skalse [*1 2] **Matthew Farrugia-Roberts** [*3] **Stuart Russell** [4] **Alessandro Abate** [1] **Adam Gleave** [5]

Abstract

It is often very challenging to manually design reward functions for complex, real-world tasks. To solve this, one can instead use *reward learning* to infer a reward function from data. However, there are often multiple reward functions that fit the data equally well, even in the infinite-data limit. This means that the reward function is only *partially identifiable*. In this work, we formally characterise the partial identifiability of the reward function given several popular reward learning data sources, including expert demonstrations and trajectory comparisons. We also analyse the impact of this partial identifiability for several downstream tasks, such as policy optimisation. We unify our results in a framework for comparing data sources and downstream tasks by their invariances, with implications for the design and selection of data sources for reward learning.

1. Introduction

Many problems can be represented as sequential decision-making tasks, where the goal is to maximise a numerical *reward function* over several steps (Sutton & Barto, 2018). However, for real-world tasks, it is often challenging to design a reward function that reliably incentivizes the right behaviour (Amodei et al., 2016; Leike et al., 2018; Dulac-Arnold et al., 2019). One approach to solving this problem is to use *reward learning* algorithms, which aim to learn a reward function from data. Reward learning algorithms can be based on many different data sources, including expert demonstrations (Ng & Russell, 2000), preferences over trajectories (Christiano et al., 2017), and many others (Jeon et al., 2020). These algorithms can learn a reward function for tasks where it would be infeasible to specify a reward manually (e.g., Abbeel et al., 2010; Christiano et al., 2017; Singh et al., 2019; Stiennon et al., 2020).

There will often be multiple reward functions that are consistent with a given data source, even in the limit of infinite data. For example, two reward functions may induce exactly the same expert behaviour; in that case, no amount of expert demonstrations can distinguish them. Similarly, two reward functions may induce exactly the same preferences over trajectories; in that case, no amount of preference data can distinguish them. This means that the reward function is ambiguous, or *partially identifiable*, based on these data sources. For most data sources, this issue has been acknowledged, but its extent has not been characterised. In this work, we formally characterise the ambiguity of the reward function for several popular data sources, including expert demonstrations (§3.1) and trajectory comparisons (§3.3). Our results describe the infinite-data bounds for the information that can be recovered from these types of data.

Identifying a reward function uniquely is often unnecessary, because all plausible reward functions might lead to the same outcome in a given application. For example, if we want to learn a reward function in order to compute an optimal policy, then it is enough to learn a reward function that has the same optimal policies as the true reward function. In general, ambiguity is not problematic if all compatible reward functions lead to identical downstream outcomes. Therefore, we also characterise the *ambiguity tolerance* for various applications (especially policy optimisation). This allows us to evaluate whether or not the ambiguity of a given data source is problematic for a given application.

Ambiguity and ambiguity tolerance are formally related. Both concern *invariances* of objects that can be computed from reward functions to *transformations* of those reward functions. Thus, our main contribution is to catalogue the invariances of various mathematical objects derived from the reward function. In Section 2.1, we explore a *partial order* on these invariances and its implications for selecting and evaluating data sources, addressing an open problem in reward learning (Leike et al., 2018, §3.1).

[*]Equal contribution [1]Department of Computer Science, Oxford University [2]Future of Humanity Institute, Oxford University [3]School of Computing and Information Systems, the University of Melbourne [4]Center for Human-Compatible Artificial Intelligence, University of California, Berkeley [5]FAR AI, Inc. Correspondence to: Joar Skalse <joar.skalse@cs.ox.ac.uk>, Matthew Farrugia-Roberts <matt.farrugia@unimelb.edu.au>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1.1. Related Work

Inverse reinforcement learning (IRL; Russell, 1998) is a central example of reward learning. An IRL algorithm attempts to infer a reward function from demonstrations by an expert, by inverting a model of the expert's *planning algorithm* (Armstrong & Mindermann, 2018; Shah et al., 2019). Existing work partially characterises the inherent ambiguity of expert demonstrations for certain planning algorithms (Ng & Russell, 2000; Cao et al., 2021) and classes of tasks (Dvijotham & Todorov, 2010; Kim et al., 2021). We extend these results by considering a more expressive space of reward functions, more planning algorithms, and arbitrary stochastic tasks.

IRL is related to *dynamic discrete choice* (Rust, 1994; Aguirregabiria & Mira, 2010), a problem where identifiability has been studied extensively (e.g., Aguirregabiria, 2005; Srisuma, 2015; Arcidiacono & Miller, 2020). We study a simpler setting with known tasks. IRL also relates to *preference elicitation* (Rothkopf & Dimitrakakis, 2011) and *inverse optimal control* (Ab Azar et al., 2020). Preferences over sequential trajectories are not typically considered as a data source in these fields.

Reward learning methods have also been proposed for many other data sources (Jeon et al., 2020). A popular and effective option is *preferences over trajectories* (Akrour et al., 2012; Christiano et al., 2017). Unlike for IRL, the ambiguity arising from these data sources has not been formally characterised in any previous work. We contribute a formal characterisation of the ambiguity for central models of evaluative feedback, including trajectory preferences.

Several studies have explored learning from a combination of both expert demonstrations and preferences (Ibarz et al., 2018; Palan et al., 2019; Bıyık et al., 2022; Koppol et al., 2020), or other multimodal data sources (Tung et al., 2018; Krasheninnikov et al., 2019; Jeon et al., 2020). One motivation is that different data sources may provide complementary reward information (Koppol et al., 2020), and therefore eliminate some ambiguity. Similarly, Amin et al. (2017) and Cao et al. (2021) observe reduced ambiguity by combining behavioural data across multiple *tasks*. We provide a general framework for understanding these results.

The primary application of a learnt reward function is to compute an optimal policy (Abbeel & Ng, 2004; Wirth et al., 2017). Ng et al. (1999) proved that *potential-shaping transformations* always preserve the set of optimal policies. We extend this result, characterising the full set of transformations that preserve optimal policies in each task, including for additional policy optimisation techniques such as maximum causal entropy reinforcement learning.

Ambiguity corresponds to the *partial identifiability* (Lewbel, 2019) of the reward function modelled as a latent parameter. A common response to partial identifiability in reward learning has been to impose additional constraints or assumptions until the data identifies the reward function uniquely. Following Manski (1995; 2003) and Tamer (2010), we instead *describe* ambiguity *given* various constraints and assumptions. This gives practitioners results that are appropriate for the data they in fact have, and the ambiguity tolerance of their actual application.

1.2. Preliminaries

We consider an idealised setting with finite, observable, infinite-horizon sequential decision-making environments, formalised as *Markov Decision Processes* (MDPs; Sutton & Barto, 2018, §3). An MDP is a tuple $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma)$ where $\mathcal{S}$ and $\mathcal{A}$ are finite sets of environment *states* and agent *actions*; $\tau : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ encodes the *transition distributions* governing the environment dynamics; $\mu_0 \in \Delta(\mathcal{S})$ is an initial state distribution; $R : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ is a *reward function*;[1] and $\gamma \in (0, 1)$ is a reward *discount rate*. We distinguish states in the support of μ_0 as *initial states*. In some cases, we also wish to include *terminal states*. These states have the property that $\tau(s|s, a) = 1$ and $R(s, a, s) = 0$ for all a. A *policy* is a function $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that encodes the behaviour of an agent in an MDP.

We represent the *transition* from state s to state s' using action a as the tuple $x = (s, a, s')$. We classify (s, a, s') as *possible* in an MDP if s' is in the support of $\tau(s, a)$, otherwise it is *impossible*. A *trajectory* is an infinite sequence of concatenated transitions $\xi = (s_0, a_0, s_1, a_1, s_2, \ldots)$, and a *trajectory fragment* of length n is a finite sequence of n concatenated transitions $\zeta = (s_0, a_0, s_1, \ldots, a_{n-1}, s_n)$. A trajectory or trajectory fragment is *possible* if all of its transitions are possible, and is *impossible* otherwise. A trajectory or trajectory fragment is *initial* if its first state is initial. We say that a state or transition is *reachable* if it is part of some possible and initial trajectory.

Given an MDP, we define the *return function* G as the cumulative discounted reward of entire trajectories and trajectory fragments: $G(\zeta) = \sum_{t=0}^{|\zeta|-1} \gamma^t R(s_t, a_t, s_{t+1})$ for a trajectory fragment ζ of length $|\zeta|$, and similarly for trajectories.

A policy π and a transition distribution τ together induce a distribution of trajectories starting from each state. We denote such a trajectory starting from s with the random variable Ξ_s, and its remaining components with random variables A_0, S_1, A_1, S_2, and so on.

Given an MDP and a policy π, the *value function* encodes

[1] Notably, we consider deterministic reward functions that may depend on a transition's successor state. Alternative spaces of reward functions are often considered (such as functions from $\mathcal{S}$ or $\mathcal{S}\times\mathcal{A}$, or distributions). The chosen space has straightforward consequences for invariances, which we discuss in Appendix E.

the expected return from a state, $V_\pi(s) = \mathbb{E}_{\Xi_s \sim \pi, \tau}\big[G(\Xi_s)\big]$, and the *Q-function* of π encodes the expected return given an initial action, $Q_\pi(s,a) = \mathbb{E}_{\Xi_s \sim \pi, \tau}\big[G(\Xi_s) \mid A_0 = a\big]$. Q_π and V_π satisfy a *Bellman equation*:

$$Q_\pi(s,a) = \mathbb{E}_{S' \sim \tau(s,a)}\big[R(s,a,S') + \gamma V_\pi(S')\big], \quad (1)$$

$$V_\pi(s) = \mathbb{E}_{A \sim \pi(s)}\big[Q_\pi(s,A)\big], \quad (2)$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Their difference, $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$, is the *advantage function* of π.

We further define a *policy evaluation function*, $\mathcal{J}$, encoding the expected return from following a particular policy in an MDP, $\mathcal{J}(\pi) = \mathbb{E}_{S_0 \sim \mu_0}\big[V_\pi(S_0)\big]$. $\mathcal{J}$ induces an order over policies. A policy maximising $\mathcal{J}$ is an *optimal policy*, denoted $\pi_\star$. Similarly, $Q_\star$, $V_\star$, and $A_\star$ denote the Q-, value, and advantage functions of an optimal policy. Since $\mathcal{J}$ may be multimodal, we often discuss the *set* of optimal policies. However, $Q_\star$, $V_\star$, and $A_\star$ are each unique.

Besides optimal policies, we also consider policies resulting from alternative objectives. Given an *inverse temperature* parameter $\beta > 0$, we define the *Boltzmann-rational* policy (Ramachandran & Amir, 2007), denoted $\pi^\star_\beta$, with a softmax distribution over the optimal advantage function:

$$\pi^\star_\beta(a \mid s) = \frac{\exp\big(\beta A_\star(s,a)\big)}{\sum_{a' \in \mathcal{A}} \exp\big(\beta A_\star(s,a')\big)}. \quad (3)$$

The *Maximal Causal Entropy* (MCE) policy (Ziebart, 2010; Haarnoja et al., 2017) is given by

$$\pi^{\mathrm{H}}_\beta(a \mid s) = \frac{\exp\big(\beta Q^{\mathrm{H}}_\beta(s,a)\big)}{\sum_{a' \in \mathcal{A}} \exp\big(\beta Q^{\mathrm{H}}_\beta(s,a')\big)}, \quad (4)$$

where Q^{H}_β is the *soft Q-function*, a regularised variant of the Q-function. Haarnoja et al. (2017, Theorem 2 and Appendix A.2) show that Q^{H}_β is the unique function satisfying

$$Q^{\mathrm{H}}_\beta(s,a) = \mathbb{E}\Big[R(s,a,S') + \gamma \frac{1}{\beta} \log \sum_{a' \in \mathcal{A}} \exp\big(\beta Q^{\mathrm{H}}_\beta(S',a')\big)\Big]. \quad (5)$$

The MCE policy results from maximising a policy evaluation function with an entropy regularisation term with weight $\alpha = \beta^{-1}$ (Haarnoja et al., 2017). The Boltzmann-rational policy can also be connected to a kind of (per-timestep) entropy regularisation (Haarnoja et al., 2017).

2. Our Framework

In this section, we describe a framework for analysing partial identifiability, that we will then use throughout the paper. This framework, illustrated in Figure 1, makes it easy to compare different data sources. The framework also simplifies reasoning about whether or not the ambiguity of a given data source could be problematic for a given application.

2.1. The Reward Learning Lattice

Given sets of states $\mathcal{S}$ and actions $\mathcal{A}$, let $\mathcal{R}$ be the set of reward functions definable over $\mathcal{S}$ and $\mathcal{A}$, that is, functions of type $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. Moreover, let X be some set of objects that can be computed from reward functions.

Definition 2.1. Given a function $f : \mathcal{R} \to X$, the *invariance partition* of f is the partition of $\mathcal{R}$ according to the equivalence relation $\sim$ where $R_1 \sim R_2$ if and only if $f(R_1) = f(R_2)$.

Invariance partitions characterise the ambiguity of reward learning data sources, and also the ambiguity tolerance of different applications. To see this, let us first build an abstract model of a reward learning algorithm. Let $R^\star$ be the true reward function. We model the data source as a function $f : \mathcal{R} \to X$, for some data space X, so that the learning algorithm observes $f(R^\star)$. Note that $f(R^\star)$ could be a distribution, which models the case where the data comprises a set of samples from some source, but it could also be some finite object. A reasonable learning algorithm should converge to a reward function R' that is compatible with the observed data, that is, such that $f(R') = f(R^\star)$. This means that the invariance partition of f groups together all reward functions that the learning algorithm could converge to. When it comes to applications, let $g : \mathcal{R} \to Y$ be the function whose output we wish to compute. If $R^\star$ is the true reward function, then it is acceptable to instead learn a reward function R' as long as $g(R') = g(R^\star)$. This means that the invariance partition of g groups together all reward functions that it would be acceptable to learn. The information contained in the data source f is guaranteed to be sufficient for computing g if $f(R') = f(R^\star) \implies g(R') = g(R^\star)$.

To make this more intuitive, let us give an example. Consider first a reward learning data source, such as trajectory comparisons. In this case, we can let X be the set of all (strict, partial) orderings of the set of all trajectories, and f be the function that returns the ordering of the trajectories that is induced by the trajectory return function, G. Let $R^\star$ be the true reward function. In the limit of infinite data, the reward learning algorithm will learn a reward function R' that induces the same trajectory ordering as $R^\star$, which means that $f(R') = f(R^\star)$. Furthermore, if we want to use the learnt reward function to compute a policy, then we may consider the function $g : \mathcal{R} \to \Pi$ that takes a reward function R, and returns a policy $\pi^\star$ that is optimal under R (in a given environment). Then as long as $f(R') = f(R^\star) \implies g(R') = g(R^\star)$, we will compute a policy that is optimal under the true reward $R^\star$.

The characterisation of ambiguity and tolerance in terms of partitions of $\mathcal{R}$ suggests a natural *partial order* on data sources and applications. Namely, we use the *partition refinement* relation on their invariances:

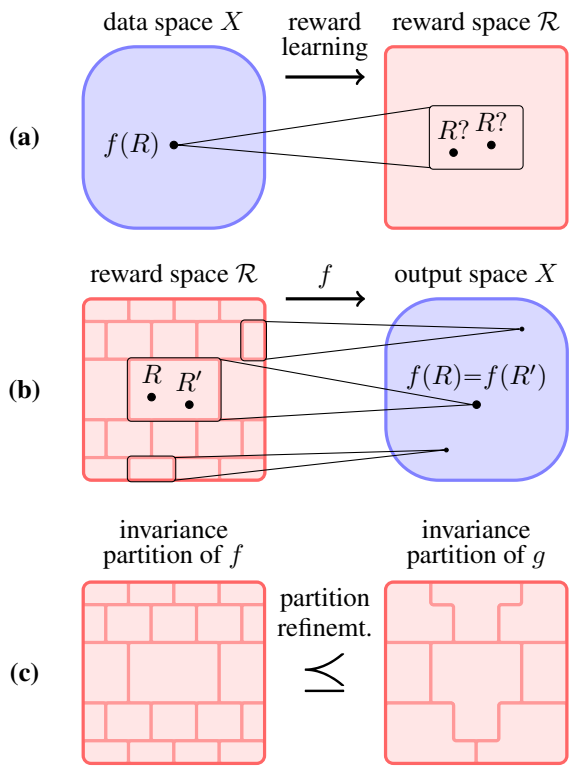

Figure 1. An overview of our framework. **(a)** A reward learning algorithm infers a reward function from data, assuming the data has been generated from some *data source*, modelled as a function $f : \mathcal{R} \to X$. Depending on f, multiple reward functions $R \in \mathcal{R}$ may be consistent with the observed data $f(R) \in X$. **(b)** By analysing the function f one can partition the reward space into groups of reward functions that lead to the same output. We call this partition the *invariance partition* of the function f. This partition can be effectively described by the set of transformations of the reward function that do not change the output of the function f. **(c)** Such a partition characterises the ambiguity of reward learning based on a data source f. Moreover, it characterises the tolerance to ambiguity of computing $f(R)$ (where R is a learnt reward function). We can compare data sources and applications by their partitions using the *partition refinement* relation, that is, $f \preceq g$ if and only if $\forall R_1, R_2 \in \mathcal{R}, f(R_1) = f(R_2) \implies g(R_1) = g(R_2)$. If f and g are data sources, then $f \preceq g$ means that f contains no more ambiguity than g. If g is a downstream application, then $f \preceq g$ means g can tolerate the ambiguity in a reward function learnt from data source f.

Definition 2.2. Consider two functions $f : \mathcal{R} \to X$ and $g : \mathcal{R} \to Y$. If $f(R_1) = f(R_2) \implies g(R_1) = g(R_2)$ for all $R_1, R_2 \in \mathcal{R}$, we write $f \preceq g$ and say f is *no more ambiguous* than g. If $f \preceq g$ but not $g \preceq f$, then we write $f \prec g$ and say that f is *strictly less ambiguous* than g.

We can use this to formalise some important relationships. Given two reward learning data sources f and g, if $f \prec g$ then we get strictly more information about the underlying reward function by observing data from f than we get by observing data from g. Moreover, given a downstream application h, $f \preceq h$ is precisely the condition of h tolerating the ambiguity of the data source f. It is also worth noting that the invariances of an object are inherited by all objects which can be computed from it. Formally:

Proposition 2.3. *Consider two functions $f : \mathcal{R} \to X$ and $g : \mathcal{R} \to Y$. If there exists a function $h : X \to Y$ such that $h \circ f = g$, then $f \preceq g$.*

This means that if there is an intermediate object (such as, for example, the Q-function) that is too ambiguous for a given application, then this will also hold for all objects that in principle could be computed from it.

Our framework places all reward learning data sources and applications in a lattice structure. In particular, the invariance partition of $f : \mathcal{R} \to X$ is a partition of $\mathcal{R}$, and the set of all these partitions forms a (bounded) lattice via the partition refinement relation. This, together with the properties of the refinement relation that we have just outlined, will make it simple and intuitive to reason about ambiguity.

2.2. Reward Transformations

Throughout the rest of this paper, we will characterise the invariances of functions $f : \mathcal{R} \to X$ in terms of the transformations of R that preserve $f(R)$. Formally:

Definition 2.4. A *reward transformation* is a map $t : \mathcal{R} \to \mathcal{R}$. We say that the *invariances* of f is a set of reward transformations T if for all $R_1, R_2 \in \mathcal{R}$, we have that $f(R_1) = f(R_2)$ if and only if there is a $t \in T$ such that $t(R_1) = R_2$. We then say that f *determines* R *up to* T.

When talking about a particular kind of object, we will for the sake of brevity usually leave the function f implicit, and instead just mention the relevant object. For example, we might say that "the Boltzmann-rational policy determines R up to T". This should be understood as saying that "f determines R up to T, where f is the function that takes a reward and returns the corresponding Boltzmann-rational policy". It is also worth noting that f and T often will be parameterised by τ, μ_0, or γ; this dependence will similarly not always be fully spelt out.

We will sometimes express the invariances of a function f in terms of several sets of reward transformations – for example, we might say that "f determines R up to T_1 and T_2". This should be understood as saying that f determines R up to T, where T is the set of all transformations that can be formed by composing transformations in T_1 and T_2.

Our results are expressed in terms of several fundamental sets of reward transformations, that we will now define. Before considering novel transformations, first recall *potential shaping*, introduced by Ng et al. (1999) and widely known

to preserve optimal policies in all MDPs. We explore some properties of potential shaping in Appendix B.

Definition 2.5 (Potential Shaping)**.** A *potential function* is a function $\Phi : \mathcal{S} \to \mathbb{R}$, where $\Phi(s) = 0$ if s is a terminal state. If $\Phi(s) = k$ for all initial states then we say that Φ is *k-initial*. Let R_1 and R_2 be reward functions. Given a discount γ, we say R_2 is produced by *(k-initial) potential shaping* of R_1 if $R_2(s, a, s') = R_1(s, a, s') + \gamma \cdot \Phi(s') - \Phi(s)$ for some (k-initial) potential function Φ.

We next introduce a number of new transformations.

Definition 2.6 (S'-Redistribution)**.** Let R_1 and R_2 be reward functions. Given transition dynamics τ, we say that R_2 is produced by *S'-redistribution* of R_1 if $\mathbb{E}_{S' \sim \tau(s,a)}\big[R_1(s, a, S')\big] = \mathbb{E}_{S' \sim \tau(s,a)}\big[R_2(s, a, S')\big]$.

S'-redistribution is simply any transformation that preserves $\mathbb{E}_{S' \sim \tau(s,a)}\big[R(s, a, S')\big]$.[2] For example, if at least two states s'_1, s'_2 are in the support of $\tau(s, a)$, then S'-redistribution could increase $R(s, a, s'_1)$ and decrease $R(s, a, s'_2)$, as long as $\mathbb{E}_{S' \sim \tau(s,a)}\big[R(s, a, S')\big]$ stays constant. Also note that S'-redistribution allows R to be changed arbitrarily for impossible transitions.

Definition 2.7 (Monotonic Transformations)**.** Let R_1 and R_2 be reward functions. We say that R_2 is produced by a *zero-preserving monotonic transformation* (ZPMT) of R_1 if for all x and $x' \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, $R_1(x) \leqslant R_1(x')$ if and only if $R_2(x) \leqslant R_2(x')$, and $R_1(x) = 0$ if and only if $R_2(x) = 0$. Moreover, we say that R_2 is produced by *positive linear scaling* of R_1 if $R_2 = c \cdot R_1$ for some positive constant c.

A zero-preserving monotonic transformation is simply a monotonic transformation that maps zero to itself. Positive linear scaling is a special case.

Definition 2.8 (Optimality-Preserving Transformation)**.** Let R_1 and R_2 be reward functions. Given transition dynamics τ and discount rate γ, we say R_2 is produced by an *optimality-preserving transformation* of R_1 if there is a function $\Psi : \mathcal{S} \to \mathbb{R}$ such that $\mathbb{E}_{S' \sim \tau(s,a)}\big[R_2(s, a, S') + \gamma \cdot \Psi(S')\big] \leqslant \Psi(s)$ for all s, a, with equality if and only if $a \in \arg\max_a A_{1\star}(s, a)$.

Here Ψ acts as a new value function, and the last condition ensures that R_1 and R_2 share the same optimal actions (and therefore the same optimal policies, cf. Theorem 3.2).

Definition 2.9 (Masking)**.** Let R_1 and R_2 be reward functions. Given a transition set $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we say that R_2 is produced by a *mask of* $\mathcal{X}$ from R_1 if $R_1(s, a, s') = R_2(s, a, s')$ for all $(s, a, s') \notin \mathcal{X}$.

[2] S'-redistribution depends crucially on the reward function's dependence on the successor state. This set of transformations collapses to the identity for simpler spaces of reward functions, as we explore in Appendix E.

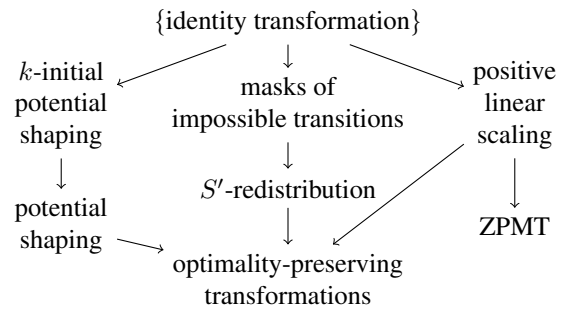

Figure 2. Subset relationships between the main sets of reward transformations (given a fixed transition distribution, initial state distribution, and discount rate). $A \to B$ denotes $A \subseteq B$.

Masking allows the reward to vary freely for some set of transitions $\mathcal{X}$. It is also worth noting that some of these sets of reward transformations are subsets of other sets. For example, all masks of impossible transitions are instances of S'-redistribution, and any potential shaping transformation is also optimality preserving, etc. All of these relationships are mapped out in Figure 2.

3. Invariances of Reward-Related Objects

In this section, we catalogue the *invariances* of various important objects that can be derived from reward functions. As discussed in Section 2.1, these invariances describe both the ambiguity when these objects are used as data sources for reward learning, and the ambiguity tolerance when these objects are computed as part of an application. Proofs and additional results are given in Appendices A to C.

3.1. Invariances of Policies

We first characterise the invariances of different types of *policies*. As discussed in Section 2.1, this describes both the ambiguity of various inverse reinforcement learning algorithms, and the ambiguity tolerance of various applications in policy optimisation. We begin with the Q-function:

Theorem 3.1. *Given an MDP and a policy π, the Q-function Q_π of π determines R up to S'-redistribution. The optimal Q-function, $Q_\star$, and the soft Q-function Q^{H}_β (for any β), both have precisely the same invariances.*

As noted in Proposition 2.3, this invariance is inherited by any object that can be derived from a Q-function. Next, we turn our attention to optimal policies. We say that an optimal policy is *maximally supportive* if it takes all optimal actions with positive probability.

Theorem 3.2. *Given an MDP, a maximally supportive optimal policy $\pi^\star$ determines R up to optimality-preserving transformations. The set of* all *optimal policies has precisely the same invariances.*

This theorem answers two questions at once. First, it tells us the exact ambiguity tolerance that we have if we want to use the learnt reward function to compute an optimal policy. Second, it also tells us what is the ambiguity of inverse reinforcement learning algorithms that observe an optimal policy, such as those developed by Ng & Russell (2000) and Abbeel & Ng (2004).

There are many inverse reinforcement learning algorithms that do not assume that the observed demonstrator policy is optimal. For example, Ramachandran & Amir (2007) and Ziebart et al. (2008) assume that the demonstrator policy is *Boltzmann-rational*, and Ziebart et al. (2010) assume that it is *causal entropy maximising*. We catalogue the invariances of both these types of policies.

Theorem 3.3. *Given an MDP and an inverse temperature parameter β, the Boltzmann-rational policy $\pi^\star_\beta$ determines R up to S'-redistribution and potential shaping. The MCE policy π^{H}_β has precisely the same invariances.*

3.2. Invariances of Trajectories

Note that Theorem 3.2 and 3.3 describe the invariances of *policies*, rather than *trajectories sampled from policies*. The invariances of the trajectories are largely the same as of the policies themselves, but with some minor caveats. In the infinite-data limit, trajectories sampled from a policy reveal the *distribution* of trajectories induced by the policy, and therefore *the policy itself* for all states reachable via its supported actions.

Boltzmann-rational policies and MCE policies support *all* actions, so the trajectory distribution determines the policy for all *reachable* states. It follows that trajectory sampling introduces invariance solely to changes in the reward of unreachable transitions.

Theorem 3.4. *Given an MDP and an inverse temperature parameter β, the distribution of trajectories $\Delta^\star_\beta$ induced by the Boltzmann-rational policy $\pi^\star_\beta$, or $\Delta^{\mathrm{H}}_\beta$ induced by the MCE policy π^{H}_β, determines R up to S'-redistribution, potential shaping, and a mask of unreachable transitions.*

Similarly, trajectories sampled from an optimal policy reveal the policy in those states that the policy visits. However, here there are some subtleties coming from the fact that an optimal policy may not visit all reachable states. This allows the reward to vary greatly, but not arbitrarily, in the states that are reachable but not visited by the optimal policy. For this purpose, we introduce a more general notion of "optimality-preserving transformation" that is parameterised by a set-valued function $\mathcal{O}$ that specifies the optimal actions.

Definition 3.5 (General Optimality-Preserving Transformation)**.** Let R and R' be reward functions. Given a function $\mathcal{O} : \mathcal{S} \to \mathcal{P}(\mathcal{A})\backslash\{\varnothing\}$, transition dynamics τ, and discount rate γ, we say R' is produced from R by a *(general) optimality-preserving transformation* with $\mathcal{O}$ if there is a function $\Psi : \mathcal{S} \to \mathbb{R}$ such that $\mathbb{E}_{S'\sim\tau(s,a)}\big[R'(s,a,S') + \gamma\Psi(S')\big] \leqslant \Psi(s)$ for all s, a, with equality if and only if $a \in \mathcal{O}(s)$.

The difference between Definitions 2.8 and 3.5 lies in the introduction of $\mathcal{O}$. In use, we constrain $\mathcal{O}$ to be $\arg\max_a A_\star(s,a)$ for s in some subset of states (rather than for all states, which is the case for Definition 2.8). If $\mathcal{O}$ were unconstrained, the set would contain all possible transformations. We can now state the invariances:

Theorem 3.6. *Given an MDP, consider the distribution of trajectories induced by a maximally supportive optimal policy, $\Delta_\star$. Let $\mathfrak{S}$ be the set of visited states. Let $\mathfrak{O}$ be the set of functions $\mathcal{O}$ defined on $\mathcal{S}$ such that $\mathcal{O}(s) = \arg\max_a A_\star(s,a)$ for $s \in \mathfrak{S}$. $\Delta_\star$ determines R up to general optimality-preserving transformations for all $\mathcal{O} \in \mathfrak{O}$.*

Note that a mask of unreachable transitions is included as a subset of the transformations permitted by Theorem 3.6. However, a mask of the complement of $\mathfrak{S}$ is *not* included. As $\mathcal{O}$ is unconstrained outside $\mathfrak{S}$ the reward is effectively unconstrained in those states, except that the reward of transitions out of $\mathfrak{S}$ may have to "compensate" for the value of their successor states to prevent new actions that lead out of $\mathfrak{S}$ from becoming optimal.

3.3. Invariances of Trajectory Evaluation

The return function captures the reward accumulated over a trajectory, and is the basis for many reward learning algorithms. In this section, we catalogue the invariances of the return function and related objects.

Theorem 3.7. *Given an MDP, the return function restricted to possible and initial trajectories, G_ξ, determines R up to zero-initial potential shaping and a mask of* unreachable *transitions.*

For reward learning, observing G_ξ would correspond to the case where the reward learning algorithm observes a trajectory or episode, together with the exact numerical value of the reward for that trajectory or episode.

A popular data source for reward learning is pairwise comparisons between trajectories, where a human is repeatedly shown two example behaviours, and asked to rank which of them is better (Akrour et al., 2012; Christiano et al., 2017). It is common to model the comparisons as being based on the trajectory return, but with accompanying *decision noise*, to account for the fact that the human labeller will sometimes make mistakes. One option is to model this noise as following a *Boltzmann distribution*, which says that the labeller is more likely to make a mistake if the trajectories have a similar value. Formally, given an MDP and an inverse temperature $\beta > 0$, let $\preceq^\zeta_\beta$ be a distribution over each

pair of *possible* trajectory fragments, ζ_1, ζ_2, such that

$$\mathbb{P}(\zeta_1 \preceq^{\zeta}_{\beta} \zeta_2) = \frac{\exp(\beta G(\zeta_2))}{\exp(\beta G(\zeta_1)) + \exp(\beta G(\zeta_2))}\,,$$

and let $\preceq^{\xi}_{\beta}$ be the analogous distributions for possible, initial *trajectories*. Intuitively, the data source $\preceq^{\zeta}_{\beta}$ corresponds to the case where a noisy expert ranks all trajectory fragments, and where the more valuable ζ_1 is relative to ζ_2, the more likely the expert is to select ζ_1 over ζ_2.

Theorem 3.8. *Given an MDP, the distribution of comparisons of possible trajectory fragments, $\preceq^{\zeta}_{\beta}$, determines R up to a mask of* impossible *transitions.*

Theorem 3.9. *Given an MDP, the distribution of comparisons of possible and initial trajectories, $\preceq^{\xi}_{\beta}$, determines R up to k-initial potential shaping and a mask of* unreachable *transitions.*

Note that the limited invariances of $\preceq^{\zeta}_{\beta}$ arises from the very flexible comparisons permitted, including, for example, comparisons between individual transitions and empty trajectories. It is also worth noting that these invariances rely very heavily on the precise structure of the decision noise, and our assumption of infinite data. First of all, in the infinite-data limit, each pair of trajectories is sampled multiple times, which means that we obtain the exact value of $\mathbb{P}(\zeta_1 \preceq^{\zeta}_{\beta} \zeta_2)$ for each ζ_1 and ζ_2. Moreover, since $\mathbb{P}(\zeta_1 \preceq^{\zeta}_{\beta} \zeta_2)$ depends on the relative difference between the values of ζ_1 and ζ_2, it reveals cardinal information about the reward. This is a property of Boltzmann noise that will not hold for many other kinds of decision noise.

It is also possible to model trajectory comparisons as noiseless comparisons based on the return of these trajectories. The infinite-data limit then corresponds to the order induced by the return function. Formally, define the *noiseless order of possible trajectory fragments* as a relation, $\preceq^{\zeta}_{\star}$, on possible trajectory fragments:

$$\zeta_1 \preceq^{\zeta}_{\star} \zeta_2 \Leftrightarrow G(\zeta_1) \leqslant G(\zeta_2)\,.$$

Analogously, define the *noiseless order of possible and initial trajectories*, $\preceq^{\xi}_{\star}$, on possible, initial trajectories. Here, the precise invariances will depend on the MDP.

Theorem 3.10. *We have the following bounds on the invariances of the noiseless order of possible trajectory fragments, $\preceq^{\zeta}_{\star}$. In all MDPs:*

(1) $\preceq^{\zeta}_{\star}$ is invariant to positive linear scaling and a mask of impossible transitions; and

(2) $\preceq^{\zeta}_{\star}$ is not invariant to transformations other than zero-preserving monotonic transformations or masks of impossible transitions.

Moreover, there exist MDPs attaining each of these bounds.

Note that $\preceq^{\zeta}_{\beta}$ is less ambiguous than $\preceq^{\zeta}_{\star}$, even though $\preceq^{\zeta}_{\beta}$ corresponds to noisy comparisons, and $\preceq^{\zeta}_{\star}$ corresponds to noiseless comparisons. The reason for this is, again, that the Boltzmann noise reveals cardinal information about the reward function, whereas noiseless comparisons only reveal ordinal information.

We give a lower bound on the invariances of the noiseless order of possible and initial trajectories, $\preceq^{\xi}_{\star}$. Since $\preceq^{\xi}_{\star}$ can be derived from $\preceq^{\xi}_{\beta}$, it inherits the latter's invariances. Moreover, $\preceq^{\xi}_{\star}$ is always invariant to positive linear scaling.

Theorem 3.11. *Given an MDP, the noiseless order of possible and initial trajectories, $\preceq^{\xi}_{\star}$, is invariant to k-initial potential shaping, positive linear scaling, and a mask of unreachable transitions.*

Note that Theorem 3.11 gives a bound on the ambiguity of $\preceq^{\xi}_{\star}$, rather than an exact characterisation. Therefore, it does *not* rule out additional invariances (unlike our other results).

We next give the transformations that preserve preferences over *lotteries* (distributions) of trajectories. Formally, let $\preceq^{\xi}_{\mathcal{D}}$ be the relation on distributions over possible initial trajectories given by

$$\mathcal{D}_1 \preceq^{\xi}_{\mathcal{D}} \mathcal{D}_2 \Leftrightarrow \mathbb{E}_{\Xi\sim\mathcal{D}_1}\big[G(\Xi)\big] \leqslant \mathbb{E}_{\Xi\sim\mathcal{D}_2}\big[G(\Xi)\big].$$

It is possible to model such preferences as vNM-rational choices between lotteries over trajectories, with the return function G as the utility function. Such preferences are well known to be invariant to positive affine transformations of the utility function, and no other transformations (von Neumann & Morgenstern, 1947, Appendix A). We show that positive affine transformations of G correspond to k-initial potential shaping and positive linear scaling of R.

Theorem 3.12. *Given an MDP, $\preceq^{\xi}_{\mathcal{D}}$ determines R up to k-initial potential shaping, positive linear scaling, and a mask of unreachable transitions.*

3.4. Computing The Reward Learning Lattice

As discussed in Section 2.1, the reward invariances of all of these objects allow us to place them in a lattice structure. We show part of this lattice in Figure 3. This allows us to represent many important relationships between data sources and applications in a graphical way. In particular, recall that if $f \preceq g$ then we get more information about the underlying reward function by observing data from f than we get by observing data from g. Moreover, given some downstream application h, we have that h tolerates the ambiguity of the data source f exactly when $f \preceq h$. Recall also that the invariances of an object are inherited by all objects which can be computed from it (Proposition 2.3). This allows one to check whether a data source provides enough information for a given application based on whether the data source is above the application in the lattice.

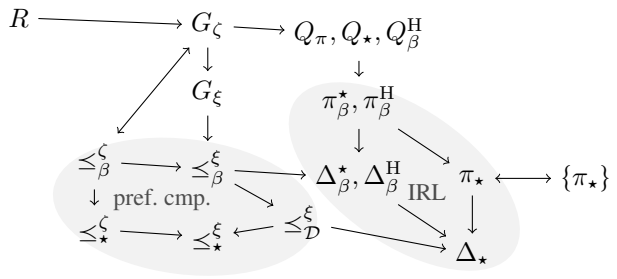

Figure 3. The invariances of objects that can be computed from a reward function in a fixed environment induce a partial order over these objects. Here, a directed path from X to Y means that $X \preceq Y$, which in turn implies that X is no more ambiguous than Y, that Y is tolerant to X's ambiguity, and that Y can be computed from X. The objects are the reward function itself (R); Q-functions ($Q_\pi, Q_\star, Q^{\mathrm{H}}_\beta$); Boltzmann-rational policies ($\pi^\star_\beta$), MCE policies (π^{H}_β), maximally supportive optimal policies ($\pi_\star$) and their trajectory distributions ($\Delta^\star_\beta, \Delta^{\mathrm{H}}_\beta, \Delta_\star$); the return function restricted to partial and full trajectories (G_ζ, G_ξ); Boltzmann-distributed (β) and noiseless ($\star$) comparisons between these trajectories ($\preceq^\zeta_\beta$, $\preceq^\xi_\beta$, $\preceq^\zeta_\star$, $\preceq^\xi_\star$) and lotteries over trajectories ($\preceq^\xi_{\mathcal{D}}$); and the set of optimal policies ($\{\pi_\star\}$).

It is worth noting that the trajectories of optimal policies ($\Delta_\star$ in Figure 3) are more ambiguous than almost all other data sources, which also means that they tolerate the ambiguity of almost all other data sources. In other words, the ambiguity of most data sources is unproblematic if the downstream task is to compute and then deploy an optimal policy. However, this notably excludes noiseless comparison data (because this data only includes ordinal information, which is insufficient in stochastic environments). Moreover, Figure 3 assumes a fixed environment. Applications in different environments may tolerate different ambiguity (Section 4.2).

4. Additional Results

In this section, we provide some additional formal results, first concerning the case where we combine data from multiple different sources, and then concerning the case where we carry out reward learning in one MDP but use the learnt reward function in a different MDP. We provide proofs of Theorems 4.1 and 4.2 in Appendix D.

4.1. Combining Data Sources

There has been recent interest in the prospect of combining information from multiple different sources for the purpose of reward learning (Krasheninnikov et al., 2019; Jeon et al., 2020). Our results offer additional insights for this setting. In particular, ambiguity refinement is a *partial* order, with some data sources being *incomparable*. We note that such incomparable ambiguity is *complementary* ambiguity, in that by combining the associated data sources, we can reduce the overall ambiguity about the latent reward.

Theorem 4.1. *Given data sources X and Y, let (X, Y) denote the combined data source formed from X and Y. If X and Y are incomparable, then $(X, Y) \prec X$ and $(X, Y) \prec Y$.*

This suggests that we can reduce the ambiguity of reward learning methods by using a mixture of data sources with complementary ambiguity. Unfortunately, most data sources appear to have similar kinds of ambiguity, assuming a fixed MDP. However, our results suggest that ambiguity could be reduced by incorporating data from *multiple* MDPs, along the lines of Amin et al. (2017) and Cao et al. (2021).

4.2. Transfer Learning

It is interesting to consider the setting where the reward is learnt in one MDP, but used in a different MDP. This captures, for example, the common sim-to-real setting, where learning occurs in a simulation whose dynamics differ slightly from those of the real environment. It also models the case where training data for the reward learning algorithm is not available for the deployment environment, but is available for a different environment.

We can begin by noting that no guarantees can be obtained if the two MDPs have very different reachable state spaces; in that case, the learnt reward function will mostly depend on the inductive bias of the learning algorithm. However, we will demonstrate that similar problems can occur even with much smaller differences between the learning environment and the deployment environment:

Theorem 4.2. *Let $\mathcal{L} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be any function, R_1 any reward function, and τ_1, τ_2 any transition distributions. Then there exists a reward function R_2, produced from R_1 by S'-redistribution under τ_1, such that $\mathbb{E}_{S' \sim \tau_2(s,a)}\big[R_2(s, a, S')\big] = \mathcal{L}(s, a)$ for all s, a such that $\tau_1(s, a) \neq \tau_2(s, a)$.*

To unpack this, let R_1 be the true reward function, τ_1 be the transition dynamics of the training environment, and τ_2 be the transition dynamics of the deployment environment. Theorem 3.1 then says, roughly, that if the reward learning algorithm is invariant to S'-redistribution, and τ_1 and τ_2 differ for enough states, then the learnt reward function is essentially unconstrained in the deployment environment, meaning that no guarantees could be obtained. Moreover, note that Theorem 3.1 and Proposition 2.3 imply that this result extends to *any* object that can be computed from a Q-function, which is a very broad class. Theorem 4.2 then suggests that any such data source is too ambiguous to guarantee transfer to a different environment. Note that this strong result relies on the formulation of rewards as depending on the successor state (see Appendix E).

5. Discussion

In this section, we will discuss the implications of our results, and some limitations and directions for future work.

5.1. Limitations and Future Work

We have catalogued the invariances of many important objects. However, there are some interesting objects that are not covered by our analysis. To start with, Theorem 3.11 only gives a bound on the invariance partition of noiseless trajectory comparisons, rather than an exact characterisation. It would be interesting to derive this invariance partition exactly. It would also be interesting to characterise what happens if various restrictions are imposed on the trajectories and trajectory fragments considered throughout Section 3.3, such as, for example, a minimum and maximum length of the trajectories. Moreover, future work could characterise the ambiguity of the *policy ordering* induced by the policy evaluation function. Theorem 3.2 characterises ambiguity tolerance of computing an optimal policy. However, in practice, one often uses a reinforcement learning algorithm that is not guaranteed to find a globally optimal policy, so it may be prudent to preserve the entire policy order, rather than just the set of maximising policies.

Moreover, our results primarily concern the asymptotic behaviour of reward learning algorithms, in the limit of infinite data. In practice, data sets are finite and, when data collection is expensive, may be quite small. An important direction for future work is to characterise how much information is contained in data sets of varying sizes and data sources. This would enable practitioners to determine the most sample-efficient data source.

Furthermore, our results rely on the assumption that data in fact is generated according to the process that is assumed by the reward learning algorithm. However, in reality, it is unlikely that these assumptions will hold perfectly (Orsini et al., 2021). For example, human demonstrations are rarely perfectly optimal or Boltzmann-rational. This means that the learning algorithms are often *misspecified* in practice. Studying this setting formally is another important direction for future work.

5.2. Conclusion and Implications

We have precisely characterised the ambiguity of many reward learning data sources, and the ambiguity tolerance of many applications using reward functions. These results can be used to easily answer whether or not a given reward learning data source is appropriate for a given application, and to compare different reward learning data sources against each other. This is valuable to help practitioners choose what reward learning method to use for a given problem.

Our results are also helpful for developing new reward learning algorithms. In developing a reward learning algorithm, it is relevant to ask how effective this algorithm is relative to an optimal algorithm for the given data source. Our results characterise the information that is available in each data source, which makes it possible to evaluate a reward learning algorithm against this theoretical upper bound.

Our framework enables direct comparisons between different data sources. We find that some data sources are strictly less informative than others. For example, noiseless preference comparisons are strictly less informative than return labels. By contrast, other data sources are incomparable and have complementary ambiguity. For example, the ambiguity of the Q-values is incomparable to the ambiguity of episode return values (the former are invariant to S'-redistribution but not to potential shaping, whereas the latter are invariant to some potential shaping but not to S'-redistribution). As discussed in Section 4.1, these results can be used to identify promising opportunities for combining information from multiple sources.

We find that in a fixed environment, many reward learning data sources are sufficient for the purpose of computing an optimal policy, including labelled trajectories, Boltzmann comparisons of trajectories and trajectory fragments, and trajectories sampled from any of the three types of policies that are common in IRL (optimal, Boltzmann-rational, or causal entropy maximising). However, this notably excludes noiseless comparisons between trajectory fragments in some MDPs, since zero-preserving monotonic transformations are not, in general, optimality preserving. These relationships are summarised in Figure 3. Theorem 4.2 also shows that a very wide range of reward learning data sources will be too ambiguous to facilitate transfer to new environments.

Acknowledgements

JS, MFR, and AG contributed to this research while affiliated with the Center for Human-Compatible Artificial Intelligence, UC Berkeley.

The authors thank Daniel Filan, Erik Jenner, Cassidy Laidlaw, and Smitha Milli for feedback on earlier versions of the manuscript and Daniel Murfet for insightful discussions about the project. We also thank our anonymous reviewers for suggestions which have helped to improve the clarity of our results and notation.

References

Ab Azar, N., Shahmansoorian, A., and Davoudi, M. From inverse optimal control to inverse reinforcement learning: A historical review. *Annual Reviews in Control*, 50:119–138, 2020. doi: 10.1016/j.arcontrol.2020.06.001.

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 1–8, Banff, Alberta, Canada, July 2004. Association for Computing Machinery. doi: 10.1145/1015330.1015430.

Abbeel, P., Coates, A., and Ng, A. Y. Autonomous helicopter aerobatics through apprenticeship learning. *The International Journal of Robotics Research*, 29(13):1608–1639, 2010. doi: 10.1177/0278364910371999.

Aguirregabiria, V. Nonparametric identification of behavioral responses to counterfactual policy interventions in dynamic discrete decision processes. *Economics Letters*, 87(3):393–398, 2005. doi: 10.1016/j.econlet.2004.12.014.

Aguirregabiria, V. and Mira, P. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010. doi: 10.1016/j.jeconom.2009.09.007.

Aigner, M. *Combinatorial Theory*. Classics in Mathematics. Springer Berlin Heidelberg, reprint of the 1979 edition, 1996. ISBN 3540617876.

Akrour, R., Schoenauer, M., and Sebag, M. APRIL: Active preference learning-based reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2012, Proceedings, Part II*, volume 7524 of *Lecture Notes in Computer Science*, pp. 116–131, Bristol, UK, 2012. Springer. doi: 10.1007/978-3-642-33486-3_8.

Amin, K., Jiang, N., and Singh, S. P. Repeated inverse reinforcement learning. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, volume 30, pp. 1813–1822, Long Beach, California, USA, 2017. Curran Associates, Inc., Red Hook, NY, USA.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint*, arXiv:1606.06565 [cs.AI], 2016.

Arcidiacono, P. and Miller, R. A. Identifying dynamic discrete choice models off short panels. *Journal of Econometrics*, 215(2):473–485, 2020. doi: 10.1016/j.jeconom.2018.12.025.

Armstrong, S. and Mindermann, S. Occam's razor is insufficient to infer the preferences of irrational agents. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, volume 31, pp. 5603–5614, Montréal, Canada, 2018. Curran Associates, Inc., Red Hook, NY, USA.

Bıyık, E., Losey, D. P., Palan, M., Landolfi, N. C., Shevchuk, G., and Sadigh, D. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *The International Journal of Robotics Research*, 41(1):45–67, 2022. doi: 10.1177/02783649211041652.

Cao, H., Cohen, S. N., and Szpruch, L. Identifiability in inverse reinforcement learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, volume 34, pp. 12362–12373, Virtual, 2021. Curran Associates, Inc., Red Hook, NY, USA.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, volume 30, pp. 4302–4310, Long Beach, California, USA, 2017. Curran Associates, Inc., Red Hook, NY, USA.

Dabney, W., Rowland, M., Bellemare, M. G., and Munos, R. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32(1), pp. 2892–2901, New Orleans, Lousiana, USA, April 2018. AAAI Press.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. Challenges of real-world reinforcement learning. In *Reinforcement Learning for Real Life, International Conference on Machine Learning Workshop*, Long Beach, California, USA, June 2019. doi: 10.48550/arXiv.1904.12901.

Dvijotham, K. and Todorov, E. Inverse optimal control with linearly-solvable MDPs. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 335–342, Haifa, Israel, June 2010. Omnipress, Madison, Wisconsin, USA.

Fishburn, P. C. *Utility Theory for Decision Making*. John Wiley & Sons, Inc., 1970. ISBN 0471260606.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361, Sydney, Australia, August 2017. PMLR.

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. Reward learning from human preferences and demonstrations in Atari. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, volume 31, pp. 8022–8034, Montréal, Canada, 2018. Curran Associates, Inc., Red Hook, NY, USA.

Jeon, H. J., Milli, S., and Dragan, A. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, volume 33, pp. 4415–4426, Virtual, 2020. Curran Associates, Inc., Red Hook, NY, USA.

Kim, K., Garg, S., Shiragur, K., and Ermon, S. Reward identification in inverse reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5496–5505, Virtual, July 2021. PMLR.

Koppol, P., Admoni, H., and Simmons, R. Iterative interactive reward learning. In *Participatory Approaches to Machine Learning, International Conference on Machine Learning Workshop*, Virtual, July 2020.

Krasheninnikov, D., Shah, R., and van Hoof, H. Combining reward information from multiple sources. In *Workshop on Learning with Rich Experience (LIRE), NeurIPS 2019*, Vancouver, Canada, December 2019. *arXiv preprint,* arXiv:2103.12142 [cs.LG].

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: A research direction. *arXiv preprint*, arXiv:1811.07871 [cs.LG], 2018.

Lewbel, A. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57 (4):835–903, 2019. doi: 10.1257/jel.20181361.

Manski, C. F. *Identification Problems in the Social Sciences*. Harvard University Press, 1995. ISBN 9780674442849.

Manski, C. F. *Partial Identification of Probability Distributions*. Springer, 2003. ISBN 0387004548.

Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H., and Tanaka, T. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 799–806, Haifa, Israel, June 2010a. Omnipress, Madison, Wisconsin, USA.

Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H., and Tanaka, T. Parametric return density estimation for reinforcement learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 368–375, Catalina Island, California, USA, 2010b. AUAI Press, Arlington, Virginia, USA.

Ng, A. Y. and Russell, S. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, volume 1, pp. 663–670, Stanford, California, USA, 2000. Morgan Kaufmann Publishers Inc., San Francisco, California, USA.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287, Bled, Slovenia, 1999. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Orsini, M., Raichuk, A., Hussenot, L., Vincent, D., Dadashi, R., Girgin, S., Geist, M., Bachem, O., Pietquin, O., and Andrychowicz, M. What matters for adversarial imitation learning? In *Proceedings of the 35th Conference on Neural Information Processing Systems*, volume 34, pp. 14656–14668, Virtual, 2021. Curran Associates, Inc., Red Hook, NY, USA.

Palan, M., Landolfi, N. C., Shevchuk, G., and Sadigh, D. Learning reward functions by integrating human demonstrations and preferences. In *Proceedings of Robotics: Science and Systems*, Freiburg im Breisgau, Germany, June 2019. doi: 10.15607/RSS.2019.XV.023.

Ramachandran, D. and Amir, E. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pp. 2586–2591, Hyderabad, India, 2007. Morgan Kaufmann Publishers Inc., San Francisco, California, USA.

Rothkopf, C. A. and Dimitrakakis, C. Preference elicitation and inverse reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2011, Proceedings, Part III*, volume 6913 of *Lecture Notes in Computer Science*, pp. 34–48, Athens, Greece, 2011. Springer. doi: 10.1007/978-3-642-23808-6_3.

Russell, S. Learning agents for uncertain environments (extended abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pp. 101–103, Madison, Wisconsin, USA, 1998. Association for Computing Machinery. doi: 10.1145/279943.279964.

Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, third edition, 2009. ISBN 9780136042594.

Rust, J. Structural estimation of Markov decision processes. In Engle, R. F. and McFadden, D. L. (eds.), *Handbook of Econometrics*, volume 4, pp. 3081–3143. Elsevier, 1994. doi: 10.1016/S1573-4412(05)80020-0.

Shah, R., Gundotra, N., Abbeel, P., and Dragan, A. On the feasibility of learning, rather than assuming, human biases for reward inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5670–5679, Long Beach, California, USA, June 2019. PMLR.

Singh, A., Yang, L., Hartikainen, K., Finn, C., and Levine, S. End-to-end robotic reinforcement learning without reward engineering. In *Proceedings of Robotics: Science and Systems*, Freiburg im Breisgau, Germany, June 2019. doi: 10.15607/RSS.2019.XV.073.

Srisuma, S. Identification in discrete Markov decision models. *Econometric Theory*, 31(3):521–538, 2015. doi: 10.1017/S0266466614000437.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize from human feedback. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, volume 33, pp. 3008–3021, Virtual, 2020. Curran Associates, Inc., Red Hook, NY, USA.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, second edition, 2018. ISBN 9780262352703.

Tamer, E. Partial identification in econometrics. *Annual Review of Economics*, 2:167–195, 2010. doi: 10.1146/annurev.economics.050708.143401.

Tung, H.-Y., Harley, A. W., Huang, L.-K., and Fragkiadaki, K. Reward learning from narrated demonstrations. In *Proceedings: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7004–7013, Salt Lake City, Utah, USA, June 2018. IEEE Computer Society, Los Alamitos, CA, USA. doi: 10.1109/CVPR.2018.00732.

von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton University Press, second revised edition, 1947.

Wiewiora, E., Cottrell, G. W., and Elkan, C. Principled methods for advising reinforcement learning agents. In *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 792–799, Washington, D.C., USA, August 2003. AAAI Press.

Wirth, C., Akrour, R., Neumann, G., and Fürnkranz, J. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(1):4945–4990, January 2017.

Ziebart, B. D. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, 2010.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *23rd AAAI Conference on Artificial Intelligence*, volume 8, pp. 1433–1438, 2008.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1255–1262, Haifa, Israel, June 2010. Omnipress, Madison, Wisconsin, USA.

A. Directory of Results

Table 1. An overview of all invariance results. Symbols: ($\equiv$) The object is invariant to the transformations; ($\equiv^*$) The object is invariant to the transformations as a special case of the invariances listed in the theorem; ($\times$) The object is not generally invariant to the transformations (more precisely, it is invariant only to those that can be represented as a combination of the listed transformations); ($\times/\equiv$) The extent of the object's invariance to the transformations depends on the MDP (as in Theorem 3.7). Blank cells indicate unresolved invariances (as in Theorem 3.11, see also Remark C.12).

Reward-derived object		Thm. & proof link	Class of transformations										
			Identity	0-initial Φ shaping	k-initial Φ shaping	Φ shaping	S' -redistr.	Pos. lin. scaling	0-preserving monotonic	Optimality-preserving	Gen. opt.-preserving	Impossible mask	Unreachable mask
R	Reward function	–	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$
Q_π	Q-function for policy π	3.1$^\square$	$\equiv^*$	$\times$	$\times$	$\times$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\times$
$Q_\star$	Optimal Q-function	3.1$^\square$	$\equiv^*$	$\times$	$\times$	$\times$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\times$
Q_β^{H}	Soft Q-function	3.1$^\square$	$\equiv^*$	$\times$	$\times$	$\times$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\times$
$\pi_\beta^\star$	Boltzmann-rational policy	3.3$^\square$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\times$
π_β^{H}	MCE policy	3.3$^\square$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\times$
$\pi_\star$	Maximally supportive optimal policy	3.2$^\square$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\times$	$\equiv$	$\times$	$\equiv^*$	$\times$
$\Delta_\beta^\star$	Trajectory distribution induced by $\pi_\beta^\star$	3.4$^\square$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\equiv$
$\Delta_\beta^{\mathrm{H}}$	Trajectory distribution induced by π_β^{H}	3.4$^\square$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\equiv$
$\Delta_\star$	Trajectory distribution induced by $\pi_\star$	3.6$^\square$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\times$	$\equiv^*$	$\equiv$	$\equiv^*$	$\equiv^*$
G_ζ	Return on possible trajectory fragments	C.11$^\square$	$\equiv^*$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\equiv$	$\times$
G_ξ	Return on possible, initial trajectories	3.7$^\square$	$\equiv^*$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\equiv$
$\preceq_\beta^\zeta$	Boltzmann comparisons of possible fragments	3.8$^\square$	$\equiv^*$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\equiv$	$\times$
$\preceq_\beta^\xi$	Boltzmann comparisons of possible initial trajectories	3.9$^\square$	$\equiv^*$	$\equiv^*$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\equiv$
$\preceq_\star^\zeta$	Noiseless order of possible fragments	3.10$^\square$	$\equiv^*$	$\times$	$\times$	$\times$	$\times$	$\equiv$	$\times/\equiv$	$\times$	$\times$	$\equiv$	$\times$
$\preceq_\star^\xi$	Noiseless order of possible initial trajectories	3.11$^\square$	$\equiv^*$	$\equiv^*$	$\equiv$			$\equiv$				$\equiv^*$	$\equiv$
$\preceq_{\mathcal{D}}^\xi$	Order of distributions over possible initial trajectories	3.12$^\square$	$\equiv^*$	$\equiv^*$	$\equiv$	$\times$	$\times$	$\equiv$	$\times$	$\times$	$\times$	$\equiv^*$	$\equiv$
$\{\pi_\star\}$	Set of optimal policies	3.2$^\square$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\times$	$\equiv$	$\times$	$\equiv^*$	$\times$
A_π	Advantage function for arbitrary policy π	C.1	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\times$
$A_\star$	Optimal advantage function	C.1	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\times$
$\pi_\beta^{\pi_0}$	Boltzmann policy for arbitrary base policy π_0	C.2	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\times$
$\Delta_\beta^{\pi_0}$	Trajectory distribution induced by $\pi_\beta^{\pi_0}$	C.6	$\equiv^*$	$\equiv^*$	$\equiv^*$	$\equiv$	$\equiv$	$\times$	$\times$	$\times$	$\times$	$\equiv^*$	$\equiv$

B. Properties of Fundamental Reward Transformations

We begin with some supporting results concerning the basic reward transformations used in Section 3 to characterise the invariances of various objects derived from the reward function.

The following result captures how potential shaping affects various reward-related functions.

Lemma B.1. *Consider M and M', two MDPs differing only in their reward functions, respectively R and R'. Denote the return function, Q-function, value function, policy evaluation function, and advantage function of M' by G', Q'_π, V'_π, $\mathcal{J}'$, and A'_π. If R' is produced by potential shaping of R with a potential function Φ, then:*

(1) for a trajectory fragment $\zeta = (s_0, a_0, s_1, \ldots, s_n)$, $G'(\zeta) = G(\zeta) + \gamma^n \Phi(s_n) - \Phi(s_0)$;

(2) for a trajectory $\xi = (s_0, a_0, \ldots)$, $G'(\xi) = G(\xi) - \Phi(s_0)$;

(3) for a state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, $Q'_\pi(s,a) = Q_\pi(s,a) - \Phi(s)$;

(4) for a state $s \in \mathcal{S}$, $V'_\pi(s) = V_\pi(s) - \Phi(s)$;

(5) for a policy π, $\mathcal{J}'(\pi) = \mathcal{J}(\pi) - \mathbb{E}_{S_0 \sim \mu_0}\big[\Phi(S_0)\big]$; and

(6) for a state $s \in \mathcal{S}$, and action $a \in \mathcal{A}$, $A'_\pi(s,a) = A_\pi(s,a)$.

Proof. (1) is given by a straightforward telescopic argument. For (2), take the limit as the length of a prefix goes to infinity, whereupon $\gamma^n \Phi(s_n)$ goes to zero ($\gamma < 1$ by definition, and $\Phi(s_n)$ is bounded since its domain is finite). (3) and (4) were proved for optimal policies by Ng et al. (1999), and they also observed that the extension to arbitrary policies is straightforward (it follows immediately from (2), for example). (5) is immediate from (4). (6) follows from (3) and (4) as the shifts of $-\Phi(s)$ to both the Q- and value functions cancel eachother. □

In the next result we show that potential shaping induces a similar state-dependent shift in the soft Q-function as well.

Lemma B.2. *Consider M_1 and M_2, two MDPs differing only in their reward functions, respectively R_1 and R_2. Denote the soft Q-function of M_1 by $Q^{\mathrm{H}}_{\beta,1}$, and of M_2 by $Q^{\mathrm{H}}_{\beta,2}$. If R_2 is produced by potential shaping of R_1 with a potential function Φ, then for all states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$, $Q^{\mathrm{H}}_{\beta,2}(s,a) = Q^{\mathrm{H}}_{\beta,1}(s,a) - \Phi(s)$.*

Proof. We will appeal to to uniqueness of the soft Q-function. By definition, $R_2(s,a,s') = R_1(s,a,s') + \gamma \cdot \Phi(s') - \Phi(s)$. Combining with Equation (5), we have for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$:

$$
\begin{aligned}
Q^{\mathrm{H}}_{\beta,1}(s,a) &= \mathbb{E}_{S' \sim \tau(s,a)}\big[R_1(s,a,S') + \gamma \frac{1}{\beta} \log \sum_{a' \in \mathcal{A}} \exp \beta Q^{\mathrm{H}}_{\beta,1}(S',a')\big] \\
&= \mathbb{E}_{S' \sim \tau(s,a)}\big[R_2(s,a,S') - \gamma \cdot \Phi(S') + \Phi(s) + \gamma \frac{1}{\beta} \log \sum_{a' \in \mathcal{A}} \exp \beta Q^{\mathrm{H}}_{\beta,1}(S',a')\big] \\
\rightarrow \quad Q^{\mathrm{H}}_{\beta,1}(s,a) - \Phi(s) &= \mathbb{E}_{S' \sim \tau(s,a)}\big[R_2(s,a,S') + \gamma \frac{1}{\beta} \log \sum_{a' \in \mathcal{A}} \exp \beta \left(Q^{\mathrm{H}}_{\beta,1}(S',a') - \Phi(S')\right)\big].
\end{aligned}
$$

We see that $Q^{\mathrm{H}}_{\beta,1}(s,a) - \Phi(s)$ satisfies Equation (5) for $Q^{\mathrm{H}}_{\beta,2}(s,a)$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Since the soft Q-function is the unique solution to this equation, we conclude $Q^{\mathrm{H}}_{\beta,2}(s,a) = Q^{\mathrm{H}}_{\beta,1}(s,a) - \Phi(s)$. □

We next show that k-initial potential shaping and linear scaling correspond to affine transformations of G.

Lemma B.3. *Let $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma)$ be an MDP, R' a reward function, and $k \in \mathbb{R}$ a constant. Then we have that $G'(\xi) = G(\xi) - k$ for all possible and initial trajectories ξ, if and only if R' is produced from R by k-initial potential shaping and a mask of unreachable transitions.*

Proof. The converse follows from Lemma B.1 and that, by definition, varying the reward for unreachable transitions does not affect the return of any possible, initial trajectories.

For the forward direction, we show that the constant difference between G' and G on possible initial trajectories implies a constant difference between the returns of possible trajectories from any given reachable state, and that this state-dependent difference defines a k-initial potential function that transforms R into R'.

Consider an arbitrary reachable state $s \in \mathcal{S}$. Let ξ_s be some possible trajectory starting in s, and define $\Delta_{\xi_s} = G(\xi_s) - G'(\xi_s)$, the difference in return ascribed to this trajectory by G and G'. We show that Δ_{ξ_s} is independent of ξ_s given s. To extend ξ_s into an initial trajectory, let ζ_s be some possible, initial, trajectory fragment ending in s (at least one exists, since s is reachable; let its length be n). Let $\zeta_s + \xi_s$ denote the concatenation of ζ_s and ξ_s. Then,

$$\begin{aligned}
\Delta_{\xi_s} &= G(\xi_s) - G'(\xi_s) \\
&= \frac{G(\zeta_s + \xi_s) - G(\zeta_s)}{\gamma^n} - \frac{G'(\zeta_s + \xi_s) - G'(\zeta_s)}{\gamma^n} && (\dagger) \\
&= \frac{k - G(\zeta_s) + G'(\zeta_s)}{\gamma^n}\,. && (\ddagger)
\end{aligned}$$

To reach (†), note that by definition of return, $G(\zeta_s + \xi_s) = G(\zeta_s) + \gamma^n G(\xi_s)$ (and likewise for G'), and recall that we have defined $\gamma > 0$. To reach (‡), note that since $\zeta_s + \xi_s$ is an initial trajectory, we have by assumption $G(\zeta_s + \xi_s) - G'(\zeta_s + \xi_s) = k$. Note (‡) shows that Δ_{ξ_s} is independent of ξ_s except for a possible dependence on ξ_s's starting state s (arising through ζ_s).

Thus, we may associate a unique $P(s) = \Delta_{\xi_s}$ with each reachable s. Then $P(s)$ is a k-initial potential function on reachable states. In particular, $P(s) = \Delta_{\xi_s} = k$ if s is initial as then we may choose ζ_s to be empty with $G(\zeta_s) = G'(\zeta_s) = 0$ and $n = 0$. Furthermore, from the definition of terminal states we must have that $P(s) = \Delta_{\xi_s} = 0$ for terminal s.

Moreover, for reachable transitions, R' is given by k-initial potential shaping of R with $\Phi(s) = P(s)$. Consider a reachable transition (s, a, s'). Let ξ and ξ' be possible trajectories such that $\xi = (s, a, s') + \xi'$. Then,

$$\begin{aligned}
R(s,a,s') + \gamma P(s') - P(s) &= R(s,a,s') + \gamma(G(\xi') - G'(\xi')) - (G(\xi) - G'(\xi)) \\
&= G'(\xi) - \gamma G'(\xi') + (R(s,a,s') + \gamma G(\xi')) - G(\xi) \\
&= G'(\xi) - \gamma G'(\xi') + G(\xi) - G(\xi) \\
&= G'(\xi) - \gamma G'(\xi') \\
&= R'(s,a,s')\,.
\end{aligned}$$

Any variation in reward for unreachable transitions can be accounted for by a mask. □

Lemma B.4. *Let $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma)$ be an MDP, R' a reward function, and $c \in \mathbb{R}$ a constant. Then $G'(\xi) = c \cdot G(\xi)$ for all possible initial trajectories ξ, if and only if R' is produced from R by zero-initial potential shaping, linear scaling by a factor of c, and a mask of all unreachable transitions.*

Proof. It is sufficient to show that the first condition is equivalent to R' being produced from $c \cdot R$ by zero-initial potential shaping and a mask of all unreachable transitions (in particular, any sequence of the above three transformations from R can be converted into a sequence where the linear scaling happens first).

Denote by G_c the return function of the scaled reward function $c \cdot R$. It is straightforward to show that $c \cdot G(\xi) = G_c(\xi)$ for all ξ. Then our first condition, $G'(\xi) = c \cdot G(\xi)$ for all possible initial trajectories, is equivalent to having $G'(\xi) = G_c(\xi)$ for these trajectories.

By Lemma B.3 (with $k = 0$) this is equivalent to R' being produced from $c \cdot R$ by zero-initial potential shaping and a mask of all unreachable transitions. This completes the proof. □

C. Proofs for Section 3 Results

We provide proofs for the theoretical results presented in the main paper along with several general supporting lemmas from which these results follow.

We distribute proofs of the results from Section 3.1 across three subsections. Appendix C.1 proves results for the invariances of (soft) Q-functions (Theorem 3.1). Appendix C.2 proves results concerning alternative policies and their trajectory distributions (Theorems 3.3 and 3.4). Appendix C.3 proves the results relating to optimal policies and their trajectory distributions (Theorems 3.2 and 3.6).

C.1. Proofs for Section 3.1 Results Concerning Q-functions

Theorem 3.1. *Given an MDP and a policy π, the Q-function Q_π of π determines R up to S'-redistribution. The optimal Q-function, $Q_\star$, and the soft Q-function Q^{H}_β (for any β), both have precisely the same invariances.*

Proof. Q_π is the only function which satisfies the Bellman equation (1) for all $s \in \mathcal{S}$, $a \in \mathcal{A}$:

$$Q_\pi(s,a) = \mathbb{E}_{S'\sim\tau(s,a),A'\sim\pi(S')}\big[R(s,a,S') + \gamma \cdot Q_\pi(S',A')\big].$$

This equation can be rewritten as

$$\mathbb{E}_{S'\sim\tau(s,a)}\big[R(s,a,S')\big] = Q_\pi(s,a) - \gamma \cdot \mathbb{E}_{S'\sim\tau(s,a),A'\sim\pi(S')}\big[Q_\pi(S',A')\big].$$

Since Q_π is the only function which satisfies this equation for all $s \in \mathcal{S}, a \in \mathcal{A}$, we have that the values of the left-hand side for each $s \in \mathcal{S}, a \in \mathcal{A}$ together determine Q_π, and vice versa. Since the left-hand side values are preserved by S'-redistribution of R, and no other transformations (cf. Definition 2.6), we have that Q_π is preserved by S'-redistribution of R, and no other transformations.

$Q_\star = Q_{\pi_\star}$ where $\pi_\star$ is any optimal policy derived from $Q_\star$, so the invariances of $Q_\star$ follow as a special case.

The proof for Q^{H}_β is essentially the same. Q^{H}_β is the only function that satisfies Equation (5) for all $s \in \mathcal{S}$, $a \in \mathcal{A}$:

$$Q^{\mathrm{H}}_\beta(s,a) = \mathbb{E}_{S'\sim\tau(s,a)}\big[R(s,a,S') + \gamma\frac{1}{\beta}\log\sum_{a'\in\mathcal{A}}\exp\beta Q^{\mathrm{H}}_\beta(S',a')\big].$$

This can be rewritten as

$$\mathbb{E}_{S'\sim\tau(s,a)}\big[R(s,a,S')\big] = Q^{\mathrm{H}}_\beta(s,a) - \gamma \cdot \mathbb{E}_{S'\sim\tau(s,a)}\big[\frac{1}{\beta}\log\sum_{a'\in\mathcal{A}}\exp\beta Q^{\mathrm{H}}_\beta(S',a')\big].$$

Since Q^{H}_β is the only function which satisfies this equation for all $s \in \mathcal{S}, a \in \mathcal{A}$, we have that the values of the left-hand side for each $s \in \mathcal{S}, a \in \mathcal{A}$ together determine Q^{H}_β, and vice versa. Since the left-hand side values are preserved by S'-redistribution of R, and no other transformations (cf. Definition 2.6), we have that Q^{H}_β is preserved by S'-redistribution of R, and no other transformations. □

C.2. Proofs for Results Concerning Alternative Policies

We split the proof of Theorem 3.3 into two proofs, Theorem C.3 and Theorem C.4, below.

In order to derive the invariances of the Boltzmann-rational policy, we analyse a more general softmax-based policy we call a *Boltzmann policy*, of which the Boltzmann-rational policy is a special case. Given a base policy π_0, and an *inverse temperature* parameter $\beta > 0$, we define the *Boltzmann policy* with respect to π_0, denoted $\pi^{\pi_0}_\beta$, using the softmax function:

$$\pi^{\pi_0}_\beta(a \mid s) = \frac{\exp\big(\beta A_{\pi_0}(s,a)\big)}{\sum_{a'\in\mathcal{A}}\exp\big(\beta A_{\pi_0}(s,a')\big)}. \tag{6}$$

The *Boltzmann-rational* policy, $\pi^\star_\beta$, is the Boltzmann policy with respect to optimal policies (cf. 3).

We begin with Lemma C.1, characterising the invariance of the advantage functions from which Boltzmann policies are derived. This in turn supports Lemma C.2, characterising the invariances of arbitrary Boltzmann policies.

Lemma C.1. *Given an MDP and a policy π, the advantage function for π, A_π, determines R up to S'-redistribution and potential shaping. The optimal advantage $A_\star$ has the same invariances.*

Proof. A_π can be derived from Q_π, given π (by Equation (1), $A_\pi(s,a) = Q_\pi(s,a) - \mathbb{E}_{A\sim\pi(s)}\big[Q_\pi(s,A)\big]$). Thus A_π is invariant to S'-redistribution following Theorem 3.1. Moreover, by Lemma B.1, potential shaping causes no change in A_π. That is, A_π is also invariant to potential shaping.

Conversely, let R and R' be such that $A_\pi = A'_\pi$. Define $\Phi : \mathcal{S} \to \mathbb{R}$ such that $\Phi(s) = \mathbb{E}_{A\sim\pi(s)}\big[Q_\pi(s,A) - Q'_\pi(s,A)\big]$. This Φ satisfies the requirements of a potential function (all Q-values from terminal states are zero). Potential shaping R with Φ yields a new reward function, denoted $R^{(\Phi)}$, with Q-function denoted $Q^{(\Phi)}_\pi$. Then observe, for each $s \in \mathcal{S}, a \in \mathcal{A}$:

$$\begin{aligned}
Q^{(\Phi)}_\pi &= Q_\pi(s,a) - \Phi(s) && \text{(by Lemma B.1)}\\
&= (Q_\pi(s,a) - \mathbb{E}_{A\sim\pi(s)}\big[Q_\pi(s,A)\big]) + \mathbb{E}_{A\sim\pi(s)}\big[Q'_\pi(s,A)\big]\\
&= A_\pi(s,a) + \mathbb{E}_{A\sim\pi(s)}\big[Q'_\pi(s,A)\big]\\
&= A'_\pi(s,a) + \mathbb{E}_{A\sim\pi(s)}\big[Q'_\pi(s,A)\big] && (A_\pi = A'_\pi \text{ by assumption})\\
&= (Q'_\pi(s,a) - \mathbb{E}_{A\sim\pi(s)}\big[Q'_\pi(s,A)\big]) + \mathbb{E}_{A\sim\pi(s)}\big[Q'_\pi(s,A)\big]\\
&= Q'_\pi(s,a)\,.
\end{aligned}$$

That is, $R^{(\Phi)}$ and R' share a Q-function. Thus, by Theorem 3.1, R' is given by S'-redistribution from $R^{(\Phi)}$.

The optimal advantage function's invariances arise as a special case, since $A_\star = A_{\pi_\star}$, where $\pi_\star$ is any optimal policy derived from $A_\star$. □

Lemma C.2. *Given an MDP, an inverse temperature parameter β, and a base policy π_0, the Boltzmann policy $\pi^{\pi_0}_\beta$ determines R up to S'-redistribution and potential shaping.*

Proof. By Equation (6), $\pi^{\pi_0}_\beta$ can be derived from A_{π_0}. Thus $\pi^{\pi_0}_\beta$ is invariant to S'-redistribution and potential shaping by Lemma C.1.

Conversely, we show that A_{π_0} can be derived from $\pi^{\pi_0}_\beta$ in turn. Therefore $\pi^{\pi_0}_\beta$ can have no more invariances than A_{π_0}, amounting to S'-redistribution and potential shaping by Lemma C.1.

For each $s \in \mathcal{S}, a \in \mathcal{A}$, observe:

$$\begin{aligned}
\pi^{\pi_0}_\beta(a \mid s) &= \frac{\exp\big(\beta A_{\pi_0}(s,a)\big)}{\sum_{a'\in\mathcal{A}} \exp\big(\beta A_{\pi_0}(s,a')\big)}\\
\rightarrow \quad A_{\pi_0}(s,a) &= \frac{1}{\beta}\log \pi^{\pi_0}_\beta(a \mid s) + \frac{1}{\beta}\log \sum_{a'\in\mathcal{A}} \exp\big(\beta A_{\pi_0}(s,a')\big)\,. \qquad (\dagger)
\end{aligned}$$

We have not yet solved for A_{π_0}, since it still occurs on both sides of (†). However, we can eliminate the RHS occurrence by appealing to the following identity (that the advantage has zero mean in each state $s \in \mathcal{S}$):

$$\begin{aligned}
\mathbb{E}_{A\sim\pi_0(s)}\big[A_{\pi_0}(s,A)\big] &= \mathbb{E}_{A\sim\pi_0(s)}\big[Q_{\pi_0}(s,A) - \mathbb{E}_{A'\sim\pi_0(s)}\big[Q_{\pi_0}(s,A')\big]\big]\\
&= \mathbb{E}_{A\sim\pi_0(s)}\big[Q_{\pi_0}(s,A)\big] - \mathbb{E}_{A\sim\pi_0(s)}\big[Q_{\pi_0}(s,A)\big]\\
&= 0\,.
\end{aligned}$$

Taking the expectation of both side of (†) therefore yields:

$$\begin{aligned}
\mathbb{E}_{A\sim\pi_0(s)}\big[A_{\pi_0}(s,A)\big] &= \mathbb{E}_{A\sim\pi_0(s)}\big[\frac{1}{\beta}\log \pi^{\pi_0}_\beta(A \mid s)\big] + \mathbb{E}_{A\sim\pi_0(s)}\big[\frac{1}{\beta}\log \sum_{a'\in\mathcal{A}} \exp\big(\beta A_{\pi_0}(s,a')\big)\big]\\
\rightarrow \quad 0 &= \mathbb{E}_{A\sim\pi_0(s)}\big[\frac{1}{\beta}\log \pi^{\pi_0}_\beta(A \mid s)\big] + \frac{1}{\beta}\log \sum_{a'\in\mathcal{A}} \exp\big(\beta A_{\pi_0}(s,a')\big)\\
\rightarrow \quad \frac{1}{\beta}\log \sum_{a'\in\mathcal{A}} \exp\big(\beta A_{\pi_0}(s,a')\big) &= -\mathbb{E}_{A\sim\pi_0(s)}\big[\frac{1}{\beta}\log \pi^{\pi_0}_\beta(A \mid s)\big]\,. \qquad (\ddagger)
\end{aligned}$$

Combining (‡) with (†) gives us an expression for A_{π_0} in terms only of $\pi_\beta^{\pi_0}$, as required:

$$A_{\pi_0}(s,a) = \frac{1}{\beta}\log \pi_\beta^{\pi_0}(a \mid s) - \mathbb{E}_{A\sim\pi_0(s)}\big[\frac{1}{\beta}\log \pi_\beta^{\pi_0}(A \mid s)\big]\,.$$

□

Our main result concerning Boltzmann-rational policies follows immediately.

Theorem C.3. *Given an MDP and an inverse temperature parameter β, the Boltzmann-rational policy $\pi_\beta^\star$, determines R up to S'-redistribution and potential shaping.*

Proof. The Boltzmann-rational policy $\pi_\beta^\star$ determines its own base policy $\pi_\star$. This is because the maximum probability actions in $\pi_\beta^\star$ are precisely those actions with maximal optimal advantage $A_\star$ ($\arg\max_{a\in\mathcal{A}} A_\star(s) = \arg\max_{a\in\mathcal{A}} \pi_\beta^\star(s,a)$). We can break ties arbitrarily, as any optimal base policy will lead to the same Boltzmann-rational policy. So, given $\pi_\beta^\star$, we are effectively also given a base policy, and the invariances of $\pi_\beta^\star$ therefore follow as a special case of Lemma C.2. □

We turn to prove the corresponding result about MCE policies, which follows a similar line of reasoning relative to the soft Q-function. We use an elementary property of the softmax function, which we state and derive as Lemma C.5 for the convenience of the unfamiliar reader.

Theorem C.4. *Given an MDP and an inverse temperature β, the MCE policy π_β^{H} determines R up to S'-redistribution and potential shaping.*

Proof. π_β^{H} is given by applying the softmax function to Q_β^{H}. Recall (or see Lemma C.5, below) that the softmax function is invariant to a constant shift, and no other transformations. This means that π_β^{H} is invariant to exactly those transformations that induce constant shifts in Q_β^{H} for each state.

S'-redistribution induces no shift in Q_β^{H} by Theorem 3.1. By Lemma B.2, potential shaping induces a state-dependent constant shift. Thus, π_β^{H} is invariant to S'-redistribution and potential shaping.

Conversely, we show that any state-dependent constant shift in Q_β^{H} can be described by these two kinds of transformations. Therefore, they are the only invariances. Let $B : \mathcal{S} \to \mathbb{R}$, and suppose R_1 and R_2 are two reward functions such that the corresponding soft Q-functions satisfy $Q_{\beta,1}^{\mathrm{H}}(s,a) = Q_{\beta,2}^{\mathrm{H}}(s,a) + B(s)$. Then,

$$\begin{aligned}
\mathbb{E}_{S'\sim\tau(s,a)}\big[R_1(s,a,S')\big] &= Q_{\beta,1}^{\mathrm{H}}(s,a) - \mathbb{E}_{S'\sim\tau(s,a)}\big[\gamma\frac{1}{\beta}\log\sum_{a'\in A}\exp\beta Q_{\beta,1}^{\mathrm{H}}(S',a')\big] \\
&= Q_{\beta,2}^{\mathrm{H}}(s,a) + B(s) - \mathbb{E}_{S'\sim\tau(s,a)}\big[\gamma\frac{1}{\beta}\log\sum_{a'\in A}\exp\beta\left(Q_{\beta,2}^{\mathrm{H}}(S',a') + B(S')\right)\big] \\
&= Q_{\beta,2}^{\mathrm{H}}(s,a) + B(s) - \mathbb{E}_{S'\sim\tau(s,a)}\big[\gamma\frac{1}{\beta}\log\left(\sum_{a'\in A}\exp\beta Q_{\beta,2}^{\mathrm{H}}(S',a')\right) + \gamma B(S')\big] \\
&= \mathbb{E}_{S'\sim\tau(s,a)}\big[R_2(s,a,S') + B(s) - \gamma B(S')\big]\,.
\end{aligned}$$

Now set $\Phi(s) = -B(s)$, and we can see that the difference between R and R' is described by potential shaping and S'-redistribution. □

Lemma C.5. *Consider two functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ defined on a finite set $\mathcal{X}$. Then the softmax distributions over f and $f + g$ agree, that is, for all $x \in \mathcal{X}$,*

$$\frac{\exp(f(x)+g(x))}{\sum_{x'\in\mathcal{X}}\exp(f(x')+g(x'))} = \frac{\exp(f(x))}{\sum_{x'\in\mathcal{X}}\exp(f(x'))}\,, \tag{7}$$

if and only if g is a constant function over $\mathcal{X}$.

Proof. This is an elementary property of the softmax function. The forward direction can be seen by manipulating Equation (7) as follows:

$$\frac{\exp(f(x)+g(x))}{\exp(f(x))} = \frac{\sum_{x'\in\mathcal{X}}\exp(f(x')+g(x'))}{\sum_{x'\in\mathcal{X}}\exp(f(x'))}$$
$$\rightarrow \quad g(x) = \log\left(\frac{\sum_{x'\in\mathcal{X}}\exp(f(x')+g(x'))}{\sum_{x'\in\mathcal{X}}\exp(f(x'))}\right)$$

which is constant in x.

The converse can be seen as follows. Assume $g(x) = G$, a constant. Then,

$$\frac{\exp(f(x)+g(x))}{\sum_{x'\in\mathcal{X}}\exp(f(x')+g(x'))} = \frac{\exp(f(x))\cdot\exp(G)}{(\sum_{x'\in\mathcal{X}}\exp(f(x')))\cdot\exp(G)} = \frac{\exp(f(x))}{\sum_{x'\in\mathcal{X}}\exp(f(x'))}\,.$$

□

We continue with results about the trajectories derived from these alternative policies. We once again prove Theorem 3.4 in two parts (Theorems C.7 and C.8). For Boltzmann-rational trajectories, we once again provide a more general lemma concerning arbitrary Boltzmann policies (6).

Lemma C.6. *Given an MDP M, an inverse temperature β, and a base policy π_0, the distribution of trajectories, $\Delta_\beta^{\pi_0}$, induced by the Boltzmann policy $\pi_\beta^{\pi_0}$ acting in MDP M determines R up to S'-redistribution, potential shaping, and a mask of* unreachable *transitions.*

Proof. That the distribution is invariant to S'-redistribution and potential shaping follows from Lemma C.2. The distribution is also invariant to changes in the reward for transitions out of unreachable states, since these rewards cannot affect the policy for reachable states. As a result, the distribution is additionally invariant to a mask of unreachable transitions.

The trajectory distribution can be factored into the separate distributions $\pi_\beta^{\pi_0}(s) \in \Delta(\mathcal{A})$ for each reachable state s by conditioning on a supported prefix trajectory fragment that leads to s and marginalising over subsequent states and actions. Via a similar argument to the proof of Lemma C.2, the distribution determines the reward function for transitions (out of these reachable states) up to potential shaping and S'-redistribution (as they affect reachable states). □

Theorem C.7. *Given an MDP M and an inverse temperature parameter β, the distribution of trajectories, $\Delta_\beta^\star$, induced by the Boltzmann-rational policy $\pi_\beta^\star$ acting in MDP M, determines R up to S'-redistribution, potential shaping, and a mask of unreachable transitions.*

Proof. As in Theorem 3.3, the invariances for the Boltzmann-rational policy's trajectories arises as a special case. □

We turn to prove the corresponding result about MCE policies, which follows a similar line of reasoning as for the Boltzmann trajectories, but relative to the MCE policy instead.

Theorem C.8. *Given an MDP M and an inverse temperature parameter β, the distribution of trajectories, $\Delta_\beta^{\mathrm{H}}$, induced by the MCE policy π_β^{H} acting in MDP M determines R up to S'-redistribution, potential shaping, and a mask of unreachable transitions.*

Proof. Directly analogous to the proof of Lemma C.6, (relative to Theorem C.4). □

C.3. Proofs for Results Concerning Optimal Policies

Our results concerning the invariance of optimal policies and their trajectories follow from the following general result connecting general optimality-preserving transformations to the set of optimal actions in some subset of states.

The key idea of the proof is to establish a link between the value-bounding function Ψ (Definition 2.8) and the optimal value function for R' via the Bellman optimality equation. We note that the definition of (general) optimality-preserving transformations is designed specifically to elicit this link.

Lemma C.9. *Given an MDP M, suppose we have the set of optimal actions for each state in a subset of states $\mathfrak{S} \subseteq \mathcal{S}$. Let $\mathfrak{O}$ be the set of (set-valued) functions $\mathcal{O} : \mathcal{S} \to \mathcal{P}(\mathcal{A})\backslash\{\varnothing\}$ such that $\mathcal{O}(s) = \arg\max_{a\in\mathcal{A}} A_\star(s,a)$ for all $s \in \mathfrak{S}$ (but where $\mathcal{O}$ is unconstrained outside $\mathfrak{S}$). Then, these optimal action sets determine R up to general optimality-preserving transformations with $\mathcal{O} \in \mathfrak{O}$.*

Proof. Suppose R' is obtained from M's reward R via a general optimality-preserving transformation with some $\mathcal{O} \in \mathfrak{O}$. Let Ψ be the corresponding value-bounding function, that is, a function $\Psi : \mathcal{S} \to \mathbb{R}$ satisfying, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\mathbb{E}_{S'\sim\tau(s,a)}\big[R'(s,a,S') + \gamma \cdot \Psi(S')\big] \leqslant \Psi(s)\,, \tag{8}$$

with equality if and only if $a \in \mathcal{O}(s)$. Since $\mathcal{O}(s)$ is nonempty (by definition), we have for all $s \in \mathcal{S}$

$$\Psi(s) = \max_{a\in\mathcal{A}} \left(\mathbb{E}_{S'\sim\tau(s,a)}\big[R'(s,a,S') + \gamma \cdot \Psi(S')\big]\right)\,.$$

This recursive condition on Ψ is the Bellman optimality equation for the unique optimal value function, $V'_\star$, of the MDP with transformed reward R'. Therefore, $\Psi(s) = V'_\star(s)$ for all $s \in \mathcal{S}$, and we can rewrite Equation (8) as

$$\mathbb{E}_{S'\sim\tau(s,a)}\big[R'(s,a,S') + \gamma \cdot V'_\star(S')\big] \leqslant V'_\star(s)\,, \tag{9}$$

with equality only for $a \in \mathcal{O}(s)$.

Now, consider a state $s \in \mathfrak{S}$. By assumption, for this s, $\mathcal{O}(s) = \arg\max_{a\in\mathcal{A}} A_\star(s,a)$. Then for this state, the actions that attain the optimal value bound in Equation (9) are these same optimal actions. Therefore, R' induces the same sets of optimal actions from states in $\mathcal{S}$.

Conversely, consider a second MDP M', differing from M only in its reward function, R'. Assume the set of optimal actions in states in $\mathfrak{S}$ agrees with the optimal actions in M for those states. Let $V'_\star$ and $A'_\star$ denote the optimal value and advantage functions for M'. The Bellman optimality equation for M' ensures that, for $s \in \mathcal{S}$,

$$V'_\star(s) = \max_{a\in\mathcal{A}} \left(\mathbb{E}_{S'\sim\tau(s,a)}\big[R'(s,a,S') + \gamma \cdot V'_\star(S')\big]\right) \tag{10}$$

with the maximum attained precisely by the actions $a \in \arg\max_{a\in\mathcal{A}}(A'_\star(s,a))$. Setting $\mathcal{O}(s) = \arg\max_{a\in\mathcal{A}}(A'_\star(s,a))$, Equation (10) can be rewritten as

$$\mathbb{E}_{S'\sim\tau(s,a)}\big[R'(s,a,S') + \gamma \cdot V_\star(S')\big] \leqslant V_\star(s) \tag{11}$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, with equality if and only if $a \in \mathcal{O}(s)$.

Now, for $s \in \mathfrak{S}$, we have $\arg\max_{a\in\mathcal{A}}(A'_\star(s,a)) = \arg\max_{a\in\mathcal{A}}(A_\star(s,a))$, because M and M' have matching sets of optimal actions for these states (by assumption). Then, Equation (11) shows that R' is produced from R by a general optimality-preserving transformation with $\mathcal{O}(s) = \arg\max_{a\in\mathcal{A}}(A'_\star(s,a))$ (and $\Psi(s) = V'_\star(s)$). □

We are now in a position to prove Theorems 3.2 and 3.6:

Theorem 3.2. *Given an MDP, a maximally supportive optimal policy $\pi^\star$ determines R up to optimality-preserving transformations. The set of* all *optimal policies has precisely the same invariances.*

Proof. By assumption, our optimal policies are maximally supportive. Therefore, their support determines the set of optimal actions from all states. Also by assumption, our maximally supportive optimal policies are determined by the set of optimal actions in each state. Therefore, a maximally supportive optimal policy has the same reward information as the set of optimal policies in each state. Its invariances follow as a special case of Lemma C.9, with $\mathfrak{S} = \mathcal{S}$. □

Theorem 3.6. *Given an MDP, consider the distribution of trajectories induced by a maximally supportive optimal policy, $\Delta_\star$. Let $\mathfrak{S}$ be the set of visited states. Let $\mathfrak{O}$ be the set of functions $\mathcal{O}$ defined on $\mathcal{S}$ such that $\mathcal{O}(s) = \arg\max_a A_\star(s,a)$ for $s \in \mathfrak{S}$. $\Delta_\star$ determines R up to general optimality-preserving transformations for all $\mathcal{O} \in \mathfrak{O}$.*

Proof. The distribution of trajectories can be factored into separate distributions $\pi_\star(s) \in \Delta(\mathcal{A})$ for each state $s \in \mathfrak{S}$ (in a manner similar to Lemma C.6, as proved above). As above, these individual distributions determine and are determined by the set of optimal actions within each of those states. The invariance result therefore follows from Lemma C.9. □

Remark C.10. As mentioned in Section 3, when there are multiple optimal policies, invariances depend on how the given policy is chosen. The proofs above reveal that our assumptions are crucial in connecting maximally supportive optimal policies to optimal action sets. We comment on the motivation for these assumptions and, following our theme of cataloguing partial identifiability, we sketch how the result would change without them.

Assumption (1), that the given policy is maximally supportive, allows us to rule out unsupported actions as suboptimal. Additional reward transformations could become permissible otherwise. As a well-known example, the zero reward function is consistent with any policy if unsupported actions could also be optimal (Ng & Russell, 2000). This more general case is difficult to analyse within our framework, because it is not well-described by transformations or an equivalence relation. The assumption may be demanding, but the consequences of misspecification are mild in the case of policy optimisation – at least the learnt reward function won't allow any suboptimal actions to become optimal.

Assumption (2), that a given policy is computed only from the set of optimal actions in each state, appears to be common. The purpose of this technical assumption is to rule out pathological schemes for encoding additional reward information through the selection of the policy. Through such schemes one could in principle encode the full reward function, for example into the infinite decimal representation of the probability of taking one action over another in some state. Such a selection scheme, even if it was not known to the learner, would remove invariances, as transformations that change the reward function but not the set of optimal states would change the given policy.

C.4. Proofs for Section 3.3 Results

Theorem C.11. *Given an MDP, the return function restricted to possible trajectory fragments, G_ζ, determines R up to a mask of* impossible *transitions.*

Proof. The result is immediate, since the restricted domain still includes all possible transitions (as length one trajectory fragments with return equal to the reward of the transition), and no fragments with impossible transitions. □

Theorem 3.7. *Given an MDP, the return function restricted to possible and initial trajectories, G_ξ, determines R up to zero-initial potential shaping and a mask of* unreachable *transitions.*

Proof. The result follows from Lemma B.3 with $k = 0$. □

Theorem 3.8. *Given an MDP, the distribution of comparisons of possible trajectory fragments, $\preceq_\beta^\zeta$, determines R up to a mask of* impossible *transitions.*

Proof. Since $\preceq_\beta^\zeta$ can be derived from G_ζ, it is invariant to a mask of impossible transitions by Theorem C.11. Conversely, $\preceq_\beta^\zeta$ determines R for all *possible* transitions. This is because $R(s, a, s')$ is encoded in the Boltzmann distribution of comparisons between the length zero trajectory fragment $\zeta_0 = (s)$ and the length one trajectory fragment $\zeta_1 = (s, a, s')$, and can be recovered as follows:

$$\begin{aligned}
\mathbb{P}(\zeta_0 \preceq_\beta^\zeta \zeta_1) &= \frac{\exp(\beta G(\zeta_1))}{\exp(\beta G(\zeta_0)) + \exp(\beta G(\zeta_1))} \\
&= \frac{\exp(\beta R(s, a, s'))}{\exp(\beta \cdot 0) + \exp(\beta R(s, a, s'))} \\
\rightarrow \quad R(s, a, s') &= \frac{1}{\beta} \cdot \log \left(\frac{\mathbb{P}(\zeta_0 \preceq_\beta^\zeta \zeta_1)}{1 - \mathbb{P}(\zeta_0 \preceq_\beta^\zeta \zeta_1)} \right).
\end{aligned}$$

Therefore $\preceq_\beta^\zeta$ is invariant to precisely a mask of impossible transitions. □

Theorem 3.9. *Given an MDP, the distribution of comparisons of possible and initial trajectories, $\preceq_\beta^\xi$, determines R up to k-initial potential shaping and a mask of* unreachable *transitions.*

Proof. Note that as $\preceq_\beta^\xi$ can be derived from G_ξ, by Theorem 3.7, $\preceq_\beta^\xi$ is invariant to zero-initial potential shaping and a mask of unreachable transitions. It is additionally invariant to k-initial potential shaping for arbitrary constants $k \in \mathbb{R}$, and no other transformations: G_ξ can be recovered from $\preceq_\beta^\xi$ up to a constant (we can compare all possible initial trajectories

to an arbitrary reference trajectory and recover their relative return using a similar manipulation as above, but we can't determine the return of the reference trajectory). From there, the precise invariance follows from Lemma B.3. □

Theorem 3.10. *We have the following bounds on the invariances of the noiseless order of possible trajectory fragments, $\preceq_\star^\zeta$. In all MDPs:*

(1) $\preceq_\star^\zeta$ is invariant to positive linear scaling and a mask of impossible transitions; and

(2) $\preceq_\star^\zeta$ is not invariant to transformations other than zero-preserving monotonic transformations or masks of impossible transitions.

Moreover, there exist MDPs attaining each of these bounds.

Proof. For (1), positive linear scaling of reward by a constant c leads to the same scaling of the return of each trajectory fragment, and this always preserves the relation $\preceq_\star^\zeta$, since for any $c > 0$, $c \cdot G(\zeta_1) \leqslant c \cdot G(\zeta_2) \Leftrightarrow G(\zeta_1) \leqslant G(\zeta_2)$ for all pairs of trajectory fragments ζ_1, ζ_2. Moreover, $\preceq_\star^\zeta$ inherits invariance to a mask of impossible transitions from G_ζ (Theorem C.11).

For (2), let R' be produced from R via some transformation that is *neither* a mask of impossible transitions *nor* a zero-preserving monotonic transformation. It must be that either R' fails to preserve the ordinal comparison of two possible transitions, or that it fails to preserve the set of zero-reward possible transitions, compared to R. In the first case, consider two possible transitions whose rewards are not preserved, x_1 and x_2. Without loss of generality, suppose $R(x_1) \leqslant R(x_2)$ but $R'(x_1) > R'(x_2)$. This corresponds to a change in $\preceq_\star^\zeta$'s comparison of the length one trajectories formed from x_1 and x_2, namely $x_1 \preceq_\star^\zeta x_2$ from true to false. Similarly, in the second case, the comparisons between the transition whose reward became or ceased to be zero and a length one trajectory (with return 0) will have changed. Therefore, $\preceq_\star^\zeta$ is not invariant to such transformations.

The bound (1) is attained by the following MDP invariant precisely to positive linear scaling and a mask of impossible transitions. Let $\mathcal{S} = \{s\}$, $\mathcal{A} = \{a_1, a_2\}$, $R(s, a_1, s) = 1$, and $R(s, a_2, s) = 1 + \gamma$. Since $R(s, a_2, s) = R(s, a_1, s) + \gamma R(s, a_1, s)$, the corresponding order relation will contain both $(s, a_2, s) \preceq_\star^\zeta (s, a_1, s, a_1, s)$ and $(s, a_1, s, a_1, s) \preceq_\star^\zeta (s, a_2, s)$. This property requires that $R(s, a_1, s) = (1 + \gamma) \cdot R(s, a_2, s)$, which is preserved only by linear scaling of R. (Non-positive linear scaling is already ruled out by (2)).

The bound (2) is attained by the following MDP invariant to arbitrary zero-preserving monotonic transformations. Let $\mathcal{S} = \{s_1, s_2\}$, $\mathcal{A} = \{a\}$, with possible transitions (s_1, a, s_2) and (s_2, a, s_2), and $R(s_1, a, s_2) > R(s_2, a, s_2) = 0$. Any zero-preserving monotonic transformation of R preserves the ordering of all *possible* trajectory fragments, namely that all nonempty trajectories starting in s_1 have positive return and all other possible trajectories have zero return. □

Theorem 3.11. *Given an MDP, the noiseless order of possible and initial trajectories, $\preceq_\star^\xi$, is invariant to k-initial potential shaping, positive linear scaling, and a mask of unreachable transitions.*

Proof. The pairwise Boltzmann distributions of $\preceq_\beta^\xi$ can be used to derive the noiseless comparisons of $\preceq_\star^\xi$, since the relative return of each pair of trajectories is encoded in each $\mathbb{P}(\xi_1 \preceq_\beta^\zeta \xi_2)$:

$$\xi_1 \preceq_\star^\xi \xi_2 \Leftrightarrow \big(G(\xi_1) \leqslant G(\xi_2)\big) \Leftrightarrow \big(\exp(\beta G(\xi_1)) \leqslant \exp(\beta G(\xi_2))\big) \Leftrightarrow \left(\tfrac{1}{2} \leqslant \mathbb{P}(\xi_1 \preceq_\beta^\zeta \xi_2)\right).$$

Therefore, $\preceq_\star^\xi$ is invariant to k-initial potential shaping and a mask of unreachable transitions by Theorem 3.9.

That $\preceq_\star^\xi$ is also invariant to positive linear scaling follows from a similar argument as for the first bound in Theorem 3.10, proved above. □

Remark C.12. Theorem 3.11 is a lower bound on the full set of invariances of the noiseless order of possible and initial trajectories. We note the following:

- It is not a tight bound: At least in some MDPs, the order is invariant to additional transformations.
- A slightly tighter lower bound can be achieved by establishing that $\preceq_\star^\xi$ can be derived from $\preceq_\star^\zeta$: Consider, for a given trajectory ξ, the sequence of 'prefix' trajectory fragments $\xi^{(0)}, \xi^{(1)}, \xi^{(2)}, \ldots$, with each $\xi^{(n)}$ comprising the first n

transitions of ξ. By definition $G(\xi) = \lim_{n\to\infty} G(\xi^{(n)})$, and so for each pair of trajectories ξ_1, ξ_2, we have $\xi_1 \preceq_\star^\xi \xi_2$ if and only if $\xi_1^{(n)} \preceq_\star^\zeta \xi_2^{(n)}$ for infinitely many n. While this is not a practical method to compute the trajectory order $\preceq_\star^\xi$ from the fragment order $\preceq_\star^\zeta$, it counts as a derivation in that it is sufficient to show that if a transformation does not change the fragment order $\preceq_\star^\zeta$, it cannot change the trajectory order $\preceq_\star^\xi$ either. Therefore, in particular, $\preceq_\star^\xi$ inherits invariance to ZPMTs in *some* MDPs from $\preceq_\star^\zeta$. This tightens the bound, at least in some MDPs.

- The previous point does not imply that the trajectory order $\preceq_\star^\xi$ inherits the fragment order $\preceq_\star^\zeta$'s *non*-invariances. A case in point is that $\preceq_\star^\xi$ is invariant to k-initial potential shaping and a mask of unreachable transitions, where $\preceq_\star^\zeta$ is not (Theorem 3.10). It is not yet clear if there are MDPs where $\preceq_\star^\xi$ is invariant to no ZPMTs other than positive linear scaling, or even to not all ZPMTs, and there may be invariances of $\preceq_\star^\xi$ that require new transformation classes to describe. However, Theorem 3.11 and this remark give us enough information to confidently position $\preceq_\star^\xi$ in our partial order of reward-derived objects.

Theorem 3.12. *Given an MDP, $\preceq_\mathcal{D}^\xi$ determines R up to k-initial potential shaping, positive linear scaling, and a mask of unreachable transitions.*

Proof. It is clear that preferences between lotteries over a choice set are preserved by positive affine transformations of the value (and no other transformations). In particular, the converse is a consequence of the well-known VNM utility theorem (von Neumann & Morgenstern, 1947). The proof by von Neumann & Morgenstern (1947) covers a finite number of outcomes, and the result also holds for an infinite number of outcomes (see, e.g., Fishburn, 1970).

Thus, our result is immediate from Lemmas B.3 and B.4, which together state that these positive affine transformations of the return function correspond exactly to k-initial potential shaping, positive linear scaling, and a mask of unreachable transitions. □

D. Proofs for Section 4 Results

Theorem 4.1. *Given data sources X and Y, let (X, Y) denote the combined data source formed from X and Y. If X and Y are incomparable, then $(X, Y) \prec X$ and $(X, Y) \prec Y$.*

Proof. Transformations that preserve (X, Y) necessarily preserve X, therefore $(X, Y) \preceq X$. But since X and Y are incomparable, there is some transformation that preserves X and not Y. This transformation does not preserve (X, Y). Therefore, $(X, Y) \prec X$. Similarly, $(X, Y) \prec Y$. □

We note that the above result is also an elementary consequence of the *lattice structure* of the partial order of partition refinement (Aigner, 1996, §I.2.B), since the combined data source corresponds to the *meet* of the original data sources.

Theorem 4.2. *Let $\mathcal{L} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be any function, R_1 any reward function, and τ_1, τ_2 any transition distributions. Then there exists a reward function R_2, produced from R_1 by S'-redistribution under τ_1, such that $\mathbb{E}_{S' \sim \tau_2(s,a)}\big[R_2(s, a, S')\big] = \mathcal{L}(s, a)$ for all s, a such that $\tau_1(s, a) \neq \tau_2(s, a)$.*

Proof. Per Definition 2.6, that R_2 is produced from R_1 by S'-redistribution under τ requires that, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\mathbb{E}_{S' \sim \tau(s,a)}\big[R_1(s, a, S')\big] = \mathbb{E}_{S' \sim \tau(s,a)}\big[R_2(s, a, S')\big]\,. \tag{12}$$

Let $s \in \mathcal{S}$ and $a \in \mathcal{A}$ be any state and action such that $\tau'(s, a) \neq \tau(s, a)$. Let $\vec{\tau}_{s,a}$ and $\vec{\tau'}_{s,a}$ be $\tau(s, a)$ and $\tau'(s, a)$ expressed as vectors, and let $\vec{R_1}_{s,a}$ be the vector where $\vec{R_1}^{(i)}_{s,a} = R_1(s, a, s_i)$. The question is then if there is an analogous vector $\vec{R_2}_{s,a}$ such that:

$$\begin{aligned} \vec{\tau}_{s,a} \cdot \vec{R_2}_{s,a} &= \vec{\tau}_{s,a} \cdot \vec{R_1}_{s,a}\,, \\ \vec{\tau'}_{s,a} \cdot \vec{R_2}_{s,a} &= \mathcal{L}(s, a). \end{aligned} \tag{13}$$

Since $\vec{\tau}_{s,a}$ and $\vec{\tau'}_{s,a}$ differ and are valid probability distributions, they are linearly independent. Therefore, the system of equations (13) always has a solution for $\vec{R_2}_{s,a}$. Form the required R_2 as R_1 modified to have the values of $\vec{R_2}_{s,a}$ in these states where the transition function is disturbed. □

E. Other Spaces of Reward Functions

Hitherto, we have assumed reward functions are members of $\mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$. That is, they are deterministic functions of transitions depending on the state, action, and successor state. In this appendix, we discuss several alternative spaces of reward functions and their implications for the invariance properties of various objects derived from the reward function.

E.1. Restricted-domain Reward Functions

It is common in both reinforcement learning and reward learning to consider less expressive spaces of reward functions. In particular, the domain of the reward function is often restricted to $\mathcal{S}$ or $\mathcal{S}\times\mathcal{A}$. When modelling a task, the choice of reward function domain is usually a formality: An MDP taking full advantage of the domain $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ has an "equivalent" MDP with a restricted domain and some added auxiliary states (Russell & Norvig, 2009, §17). Conversely, reward functions with restricted domains can be viewed as a special case of functions from $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ where the functions are constant in the final argument(s). Restricting the domain can be an appealing simplification when modelling a task, hence the popularity of these formulations.

When modelling a data source, this equivalence may not apply: We may not have access to data regarding auxiliary states, so assuming a restricted domain effectively assumes the latent reward is indeed constant with respect to the successor state (and possibly the action) of each transition. This assumption may or may not be warranted.

If a restricted domain of $\mathcal{S}$ or $\mathcal{S}\times\mathcal{A}$ is preferred, then our invariance results can be adapted in a straightforward manner. In general, since we are effectively considering a subspace of candidate reward functions for transformations, ambiguity can only decrease. In particular, these restrictions have two main consequences.

Firstly, the reward function transformation of S'-redistribution vanishes to the identity transformation, since it allows variation only in the successor state argument of the reward function, which is now impossible. This reduces the effective ambiguity of the Q-function and all derivative objects. Notably, the Q-function uniquely identifies the reward function, and Boltzmann policies have the same invariances as Boltzmann comparisons between trajectories. Restricting the domain to $\mathcal{S}$ means the (state) value function for an arbitrary known policy also uniquely identifies the reward function but doesn't otherwise alter the invariances we have explored.

Secondly, for most MDPs, the available potential-shaping transformations are restricted, but not eliminated. The function added in a potential-shaping transformation $(\gamma \cdot \Phi(s') - \Phi(s))$ nominally depends on the successor state of the transition. Some transformed reward functions may rely on this dependence, falling outside of the restricted domain. However, some non-zero transformations will usually remain. For example, in a discounted MDP without terminal states, a non-zero constant potential function $\Phi(s) = k$ does not *effectively* depend on s, and the reward transformation of adding $\gamma \cdot \Phi(s') - \Phi(s) = (\gamma - 1) \cdot k$ to a reward function does not introduce a dependence on s'. In general, the set of remaining potential-shaping transformations will depend on the network structure of the MDP. At the extreme, in a deterministic MDP with state-action rewards, all potential-shaping transformations are permitted, since a dependence on s' can be satisfied by a.

E.2. Stochastic Reward Functions

Certain tasks are naturally modelled as providing rewards drawn stochastically from some distribution upon each transition. An even more expressive space of reward functions than we consider is the space of *transition-conditional reward distributions*.[3] Identifying the reward function in this case is more challenging in general because the latent parameter contains a full distribution of information for each input, rather than a single point. In the spirit of this paper, we sketch a characterisation of this additional ambiguity.

A deterministic reward function can be viewed as the conditional expectation of a reward distribution function. Taking the expectation of the reward distribution for each transition introduces invariance, since the expectation operation is not injective (except in certain restricted cases such as for parametric families of distributions that can be parametrised by their mean). The invariance introduced is akin to S'-redistribution, but with an expectation over the support of the reward distribution rather than the successor state of each transition.

In the extension of the RL formalism to account for stochastic rewards, this expectation is effectively the first step in the

[3] Of course, it's also possible to consider reward to be distributed conditionally on only the state or state-action components of a transition and not the full transition.

derivation of each of the objects we have studied. Therefore, all of these objects inherit this new invariance.

As a consequence, all data sources are effectively more ambiguous with respect to this new latent parameter. For example, if optimal comparisons between trajectories are understood to be performed based on the pairwise comparison of the expected return of each individual trajectory, then these comparisons are also invariant to transformations of the reward distributions that preserve their means.

Fortunately, much of reinforcement learning also focuses on expected return and reward *in application*. Accordingly, most downstream tasks are tolerant to any ambiguity in the exact distribution of stochastic rewards, beyond identifying the mean. Since this is the same kind of ambiguity that is introduced by considering the latent parameter of reward learning as a conditional distribution rather than a deterministic function, our results are still informative for these situations.

E.3. Further Spaces and Future Work

For certain applications, including *risk-sensitive RL* where non-mean objectives are pursued (Morimura et al., 2010a;b; Dabney et al., 2018), the distribution of stochastic rewards can be consequential. Moreover, the introduction of stochastic rewards suggests considering data sources based on samples rather than expectations, such as a data source of trajectory comparisons based on sampled trajectory returns. Characterising the invariances of these objectives to transformations of the reward distribution, and thereby their *ambiguity tolerance*, is left to future work.

In future extensions of this work to handle continuous MDPs, there will be an opportunity to study the effect of restricting to various parametrised spaces of reward functions. For example, it is common in reinforcement learning and reward learning to study MDPs with reward functions that are linear in a *feature vector* associated with each transition. This kind of restriction may reduce the available reward transformations compared to those available to a non-parametric reward function in a similar manner to restricting the domain of a finite reward function as discussed above.

The relaxation of the Markovian assumption also introduces a broader space of reward functions and with it new dimensions for transformations and invariance. As one example related to potential shaping, the non-Markovian additive transformations studied by Wiewiora et al. (2003) will amount to new invariances of the optimal policy and other related objects.