Partial Identifiability in Inverse Reinforcement Learning for Agents with Non-Exponential Discounting

Joar Max Viktor Skalse, Alessandro Abate

Department of Computer Science, University of Oxford
joar.skalse@cs.ox.ac.uk, aabate@cs.ox.ac.uk

Abstract

The aim of inverse reinforcement learning (IRL) is to infer an agent's *preferences* from observing their *behaviour*. Usually, preferences are modelled as a reward function, R, and behaviour is modelled as a policy, π. One of the central difficulties in IRL is that multiple preferences may lead to the same observed behaviour. That is, R is typically underdetermined by π, which means that R is only *partially identifiable*. Recent work has characterised the extent of this partial identifiability for different types of agents, including *optimal* and *Boltzmann-rational* agents. However, work so far has only considered agents that discount future reward *exponentially*: this is a serious limitation, especially given that extensive work in the behavioural sciences suggests that humans are better modelled as discounting *hyperbolically*. In this work, we newly characterise partial identifiability in IRL for agents with non-exponential discounting: our results are in particular relevant for hyperbolical discounting, but they also more generally apply to agents that use other types of (non-exponential) discounting. We significantly show that generally IRL is unable to infer enough information about R to identify the correct optimal policy, which entails that IRL alone can be insufficient to adequately characterise the preferences of such agents.

Introduction

Inverse reinforcement learning (IRL) is a subfield of machine learning that aims to develop techniques for inferring an agent's *preferences* based on their *actions*. Preferences are typically modelled as a reward function, R, and behaviour is typically modelled as a policy, π. An IRL algorithm must additionally employ a *behavioural model* that describes how π is computed from R: by inverting this model, an IRL algorithm can then deduce R from π.

There are many motivations and applications underpinning IRL. For example, it can be used in *imitation learning* (e.g. Hussein et al. 2017) or as a tool for *preference elicitation* (e.g. Hadfield-Menell et al. 2016). In the former case it is not fundamentally important that the learnt reward function corresponds to the actual preferences of the observed agent, as long as it aids the imitation learning process. However, in the latter, it is instead fundamental that the learnt reward function captures the preferences of the observed agent as closely as possible. In this paper, we are primarily concerned with IRL in the context of preference elicitation, namely in settings where IRL is used to learn a representation of the preferences of an actual human subject, based on information about how that human behaves in some environment, and where we wish for the learnt reward function to capture these preferences as faithfully as possible.

One of the central challenges in IRL is that a given sequence of actions typically can be explained by many different goals. That is, there may be multiple reward functions that would produce the same policy under a given behavioural model. This means that the goals of an agent are ambiguous, or *partially identifiable*, even in the limit of infinite data. To clearly understand the impact of this partial identifiability, it is important that this ambiguity can be quantified and characterised. The ambiguity of the reward function in turn depends on the behavioural model. For some behavioural models, the partial identifiability has been studied (Ng and Russell 2000; Dvijotham and Todorov 2010; Cao, Cohen, and Szpruch 2021; Kim et al. 2021; Skalse et al. 2022; Schlaginhaufen and Kamgarpour 2023; Metelli, Lazzati, and Restelli 2023). However, this existing work has focused on a small number of behavioural models that are prevalent in the current IRL literature, whereas for other plausible or more general behavioural models, the issue of partial identifiability has largely not been studied.

One of the most important parts of a behavioural model is the choice of the *discount function*. In a sequential decision problem, different actions may lead the agent to receive more or less reward at different points in time. In these cases, it is common to let the agent discount future reward, so that reward which will be received sooner is given greater weight than reward which will be received later. Discounting can be done in many ways, but the two most prominent forms of discounting are *exponential discounting*, according to which reward received at time t is given weight γ^t; and *hyperbolic discounting*, according to which reward received at time t is given weight $1/(1 + kt)$. Here $\gamma \in (0, 1]$ and $k \in (0, \infty)$ are two parameters.[1] At the moment, most work on IRL assumes that the observed agent discounts exponen-

[1]For a more in-depth overview of discounting, see for example Frederick, Loewenstein, and O'Donoghue (2002).

tially. However, extensive work in the behavioural sciences suggests that humans (and other animals) are better modelled as using hyperbolic discounting (e.g., Thaler 1981; Mazur 1987; Green and Myerson 1996; Kirby 1997; Frederick, Loewenstein, and O'Donoghue 2002). It is therefore a significant limitation that current IRL work exclusively employs behavioural models with exponential discounting.

In this paper, we provide the first study of partial identifiability in IRL with non-exponential discounting. Specifically, we first introduce three new behavioural models for agents with general discounting. We then study the partial identifiability of the reward function under these models, and provide both an exact characterisation and a comparison between models. Notably, we show that IRL algorithms are unable to infer enough information about R to identify the correct optimal policy based on observations of an agent that discounts non-exponentially, which importantly suggests that IRL alone is insufficient to adequately characterise the preferences of such agents. All of our results apply to agents that use any general form of discounting, including the important hyperbolic discount rule. Our results thereby substantially extend the existing literature on partial identifiability in IRL, and have the potential to make it relevant to human decision making, since in particular hyperbolic discounting is thought to better fit human behaviour than exponential discounting.

Our analysis is mathematical, rather than empirical, to ensure that our results are exact and general. Moreover, we focus on behavioural models, rather than specific IRL algorithms, because we want to characterise what information is contained in certain types of data, and thereby discover limitations that apply to all IRL algorithms in this problem setting. This makes our results broadly applicable.

Related Work

The issue of partial identifiability in IRL has been studied for many behavioural models. In particular, Ng and Russell (2000) study optimal policies with state-dependent reward functions, Dvijotham and Todorov (2010) study regularised MDPs with a particular type of dynamics, Cao, Cohen, and Szpruch (2021) study how the reward ambiguity can be reduced by combining information from multiple environments, Skalse et al. (2022) study three different behavioural models and introduce a framework for reasoning about partial identifiability in reward learning, Schlaginhaufen and Kamgarpour (2023) study ambiguity in constrained MDPs, and Metelli, Lazzati, and Restelli (2023) quantify sample complexities for optimal policies. However, all these papers assume exponential discounting.

Most IRL algorithms are designed for agents that discount exponentially, but some papers have considered hyperbolic discounting (Evans, Stuhlmueller, and Goodman 2015; Chan, Critch, and Dragan 2019; Schultheis, Rothkopf, and Koeppl 2022). However, these papers do not formally characterise the identifiability of R given their algorithms.

Preliminaries

In this section, we give a brief overview of all material that is required to understand this paper, together with our basic assumptions, and our choice of terminology.

Reinforcement Learning

In this paper, we take a *Markov decision processes* (MDP) to be a tuple $\langle \mathcal{S}, \mathcal{A}, \{s_\top\}, \tau, \mu_0, R, \gamma\rangle$ where $\mathcal{S}$ is a set of *states*, $\mathcal{A}$ is a set of *actions*, $\{s_\top\}$ is a *terminal state*, $\tau : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S} \cup \{s_\top\})$ is a *transition function*, $\mu_0 \in \Delta(\mathcal{S})$ is an *initial state distribution*, $R : \mathcal{S} \times \mathcal{A} \times (\mathcal{S} \cup \{s_\top\}) \to \mathbb{R}$ is a *reward function*, and $\gamma \in (0, 1]$ is a *discount rate*. We will also assume that $\mathcal{S}$ and $\mathcal{A}$ are finite. A *policy* is a function $\pi : (\mathcal{S}\times\mathcal{A})^\star \times \mathcal{S} \to \Delta(\mathcal{A})$. If a policy π can be expressed as a simpler function $\mathcal{S} \to \Delta(\mathcal{A})$, then we say that it is *stationary*. We use $\mathcal{R}$ to denote the set of all reward functions definable over $\mathcal{S}$ and $\mathcal{A}$, i.e. $\mathbb{R}^{\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$, and Π to denote the set of all (stationary and non-stationary) policies that can be defined over $\mathcal{S}$ and $\mathcal{A}$, i.e. $\Delta(A)^{(\mathcal{S}\times\mathcal{A})^\star\times\mathcal{S}}$.

A *trajectory* $\xi = \langle s_0, a_0, s_1 \dots\rangle$ is a (finite or infinite) sequence of states and actions that form a path in an MDP. If $s_\top \in \xi$, then we assume that ξ is finite, and that $s_\top$ is the last state in ξ.[2] The *return function* G gives the cumulative discounted reward of a trajectory, $G(\xi) = \sum_{t=0}^{|\xi|} \gamma^t R(s_t, a_t, s_{t+1})$. The *value function* $V^\pi : \mathcal{S} \to \mathbb{R}$ of a (stationary) policy π encodes the expected cumulative discounted reward from each state under policy π, and its related Q-function is $Q^\pi(s, a) = \mathbb{E}_{S'\sim\tau(s,a)} [R(s, a, S') + \gamma V^\pi(S')]$. If a policy π satisfies that $V^\pi(s) \geq V^{\pi'}(s)$ for all states s and all policies π', then we say that π is an *optimal policy*. $Q^\star$ denotes the Q-function of optimal policies. This function is unique, even when there are multiple optimal policies.

We say that an MDP is *episodic* if there is some $H \in \mathbb{N}$ such that any policy with probability 1 will enter the terminal state $s_\top$ after at most H steps, starting from any state. Note that $H \leq |\mathcal{S}|$ in any episodic MDP. We say that an MDP is *non-episodic* if $s_\top$ is unreachable from any $s \in \mathcal{S}$. Note that an MDP may be neither episodic or non-episodic. Since the transition function τ alone determines whether or not an MDP is episodic, non-episodic, or otherwise, we will also refer to episodic and non-episodic transition functions. In episodic MDPs, we refer to a trajectory that starts in some state $s_0 \in \mathrm{supp}(\mu_0)$ and ends in $s_\top$ as an *episode*.

When constructing examples of MDPs, it will sometimes be convenient to let the set of actions $\mathcal{A}$ vary between different states. In these cases, we may assume that each state has a "default action" that is chosen from the actions available in that state, and that all actions that are unavailable in that state simply are equivalent to the default action.

Inverse Reinforcement Learning

In IRL we wish to infer a reward function R based on a policy π that has been computed from R. To do this, we need a *behavioural model* that describes how π relates to R. One of the most common models is known as *Boltzmann Rationality* (e.g. Ramachandran and Amir 2007), and is given by $\mathbb{P}(\pi(s) = a) \propto \exp \beta Q^\star(s, a)$, where β is a temperature parameter, and $Q^\star$ is the optimal Q-function for exponential discounting of R with fixed discount parameter γ. In other

[2]More precisely, this means that a trajectory is an element of $(\mathcal{S} \times \mathcal{A})^\star \times (\mathcal{S} \cup \{s_\top\}) \cup (\mathcal{S} \times \mathcal{A})^\omega$.

words, a Boltzmann-rational policy is given by applying a *softmax function* to $Q^\star$. An IRL algorithm infers R from π by inverting a behavioural model. There are many algorithms for doing this (e.g. Ng and Russell 2000; Ramachandran and Amir 2007; Haarnoja et al. 2017, and many others), but for the purposes of this paper, it will not be important to be familiar with the details of these algorithms.

Partial Identifiability

Following Skalse et al. (2022), we will characterise partial identifiability in terms of transformations and equivalence relations on $\mathcal{R}$. Let us first introduce a number of definitions:

Definition 1. A *behavioural model* is a function $\mathcal{R} \to \Pi$.

For example, we can consider the function $b_{\beta,\tau,\gamma}$ that, given a reward R, returns the Boltzmann-rational policy with temperature β in the MDP $\langle \mathcal{S}, \mathcal{A}, \{s_\top\}, \tau, \mu_0, R, \gamma \rangle$. Note that we consider the environment dynamics (i.e. the transition function, τ) to be part of the behavioural model. This makes it easier to reason about if and to what extent the identifiability of R depends on τ.

Definition 2. A *reward transformation* is a function $t : \mathcal{R} \to \mathcal{R}$. Given a behavioural model $f : \mathcal{R} \to \Pi$ and a set T of reward transformations, we say that f *determines* R *up to* T if $f(R_1) = f(R_2)$ if and only if $R_2 = t(R_1)$ for some $t \in T$.

Definition 2 states that the partial identifiability of the reward under a particular behavioural model can be fully characterised in terms of reward transformations. To see this, let us first build an abstract model of an IRL algorithm. Let $R^\star$ be the true reward function. We model the data source as a function $f : \mathcal{R} \to \Pi$, so that the learning algorithm observes the policy $f(R^\star)$. A reasonable learning algorithm should learn (or converge to) a reward function R_H that is compatible with the observed policy, i.e. a reward such that $f(R_H) = f(R^\star)$. This means that if f determines R up to T, then an IRL algorithm based on f is unable to distinguish between two reward functions R_1, R_2 exactly when R_1 and R_2 are related by some transformation in T. Hence, the partial identifiability of R under f can be characterised (Skalse et al. 2022).

The Non-Exponential Setting

In order to study partial identifiability in IRL with non-exponential discounting, we must first develop behavioural models for this setting. Since the Boltzmann-rational model is the most prominent behavioural model in the standard (exponentially discounted) setting, we will generalise the Boltzmann-rational behavioural model to work for general discount functions. However, before we can do this, we must first generalise the basic RL setting. We will allow a *discount function* to be any function $d : \mathbb{N} \to [0, 1]$ such that $d(0) = 1$. Some noteworthy examples of discount functions include *exponential discounting*, where $d(t) = \gamma^t$, *hyperbolic discounting*, where $d(t) = 1/(1 + k \cdot t)$, and *bounded planning*, where $d(t) = 1$ if $t \leq n$, else 0. Here γ, k, and n are parameters. In this paper, we are especially interested in hyperbolic discounting, since it is argued to be a good match to human behaviour. However, most of our results apply to arbitrary discount functions.[3]

Many of the basic definitions in RL can straightforwardly be extended to general discount functions. We consider an MDP to be a tuple $\langle \mathcal{S}, \mathcal{A}, \{s_\top\}, \tau, \mu_0, R, d \rangle$, where d may be any discount function. As usual, we define the trajectory return function as $G(\xi) = \sum_{t=0}^{|\xi|} d(t) \cdot R(\xi_t)$. We say that $V^\pi(\xi)$ is the expected future discounted reward if you start at the (finite) trajectory ξ and sample actions from π, and that $Q^\pi(\xi, a)$ is the expected future discounted reward if you start at trajectory ξ, take action a, and then sample all subsequent actions from π.[4] As usual, if π is stationary, then we let V^π and Q^π be parameterised by the current state, instead of the past trajectory.

It will be convenient to also use value- and Q-functions that start discounting from a different time than zero — we will indicate this with a superscript. Specifically, we let $V^{\pi,n}(\xi) = \mathbb{E}\left[\sum_{t=0}^{\infty} d(t + n) \cdot R(\zeta_t)\right]$, where the expectation is over a trajectory ζ given by starting with the (finite) trajectory ξ, and then sampling all subsequent actions from π. We also define $Q^{\pi,n}$ analogously. Note that $V^\pi = V^{\pi,0}$ and $Q^\pi = Q^{\pi,0}$. Intuitively speaking, $V^{\pi,n}(\xi)$ is the expected future discounted reward if you start at the (finite) trajectory ξ and then sample all subsequent actions from π, but discount as though you are starting at time n. Note that with exponential discounting, we have that $V^\pi \propto V^{\pi,n}$ for all n, but not with non-exponential discounting.

For exponential discounting where $\gamma < 1$, we have that $\sum_{t=0}^{\infty} \gamma^t < \infty$. This ensures that V^π always is strictly finite for any choice of R and τ. However, if $\sum_{t=0}^{\infty} d(t)$ diverges, then V^π will also diverge for some R and τ, which of course is problematic for policy selection. Therefore, it could be reasonable to impose the requirement that $\sum_{t=0}^{\infty} d(t) < \infty$ as a condition on d. Unfortunately, this would rule out the hyperbolic discount function. Since this discount function is of particular interest to us, we will instead impose conditions on the environment. In particular, if the MDP is *episodic* then V^π is always finite, regardless of which discount function is chosen. For this reason, most of our results will assume that the environment is episodic.

An important property of general discount functions is that they can lead to preferences that are *inconsistent over time*. To understand this, consider the following example:

Example 1. Let Gym be the MDP where $\mathcal{S} = \{s_0, s_1, s_2\}$, $\mathcal{A} = \{\text{buy}, \text{exercise}, \text{enjoy}, \text{go home}\}$, $\mu_0 = s_0$, and the transition function τ is the deterministic function given by the graph in Figure **??**. The discount function d is the hyperbolic discount function, $d(t) = 1/(1 + t)$, and R is the

[3]Note that average-reward reinforcement learning (Mahadevan 1996) is not covered by this setting. Averaging the rewards is not a form of discounting, but is instead an alternative to discounting. However, also note that if all possible episodes have the same length, then the average-reward objective is equivalent to using a constant discount function.

[4]As we will soon see, we will have to consider non-stationary policies in the setting with non-exponential discounting. This is why the Q-function and value function must be parameterised by the entire past trajectory, instead of just the current state.

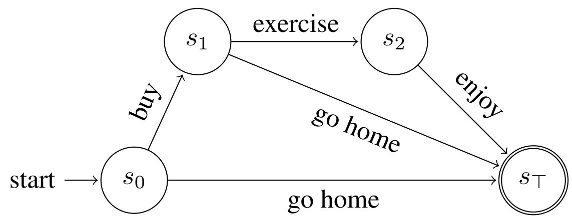

Figure 1: The MDP `Gym`, described in Example 1.

reward function given by $R(buy) = -1$, $R(exercise) = -16$, $R(enjoy) = 30$, and $R(go\ home) = 0$. □

This is a deterministic, episodic environment with three states $\{s_0, s_1, s_2\}$, where s_0 is initial. In state s_0, the agent can choose between either buying a gym membership, or going home. If it buys the gym membership, then it gets to choose between exercising at the gym, or going home. If it exercises, then it gets to enjoy the benefits of exercise, after which the episode ends. Similarly, if the agent ever goes home, the episode also ends.

We can calculate the value of each trajectory from the initial state s_0; $G(\text{go home}) = 0$, $G(\text{buy, go home}) = -1$, and $G(\text{buy, exercise, enjoy}) = 1$. This means that the most valuable trajectory from s_0 involves buying a gym membership, and then exercising. However, if we calculate the value of each trajectory from state s_1, we (paradoxically) find that $G(\text{go home}) = 0$ and $G(\text{exercise, enjoy}) = -1$. This means that the agent at state s_0 would prefer to buy a gym membership, and then exercising. However, after having bought the gym membership, the agent now prefers to go home instead of exercising. In other words, the agent has preferences that are inconsistent over time. We can formalise this as follows.

Definition 3. A discount function d is *temporally consistent* if for all sequences $\{x_t\}_{t=0}^{\infty}$, $\{y_t\}_{t=0}^{\infty}$, $\sum_{t=0}^{\infty} d(t) \cdot x_t < \sum_{t=0}^{\infty} d(t) \cdot y_t$ implies that $\sum_{t=0}^{\infty} d(t+n) \cdot x_t < \sum_{t=0}^{\infty} d(t+n) \cdot y_t$ for all n.

Intuitively, if a discount function d is temporally consistent, and at some time n it prefers a sequence of rewards $\{x_t\}_{t=0}^{\infty}$ over another sequence $\{y_t\}_{t=0}^{\infty}$, then this is also true at every other time n. On the other hand, if d is *not* temporally consistent, then it may change its preference as time passes, as in the MDP from Example 1. It is easy to show that exponential discounting *is* temporally consistent, and Example 1 demonstrates that hyperbolic discounting is *not* temporally consistent. What about other discount functions? As it turns out, exponential discounting is the *only* form of discounting that is temporally consistent. This means that *all other discount functions* can lead to preferences that are not consistent over time.

Proposition 1. *d is temporally consistent if and only if $d(t) \propto \gamma^t$ for some $\gamma \in [0, 1]$.*

For a proof of Proposition 1, see Strotz (1955) or Lattimore and Hutter (2014). Note that this temporal inconsistency is an important reason for why hyperbolic discounting is considered to be a good fit for human data — under experimental conditions, humans can exhibit *preference reversals* in a way that is consistent with hyperbolic discounting (see e.g. Frederick, Loewenstein, and O'Donoghue 2002). However, temporal inconsistency also implies that there no longer is an unamibiguous notion of what it means for a policy to be "better" than another policy in this setting. For instance, in Example 1, should the "best" policy choose to exercise at s_1, or should it choose to go home? There are multiple ways to answer this question, which in turn means that there are multiple ways to formalise what it means for an agent to "use" hyperbolic discounting (or other non-exponential discount functions). As such, modelling agents that discount non-exponentially (such as humans) involves some subtle modelling choices that are not present for exponentially discounting agents. In the next section, we explore several ways of dealing with this issue.

New Behavioural Models for General Discounting

We wish to construct behavioural models that are analogous to Boltzmann-rationality for non-exponential discounting. Recall that in the exponentially discounted setting, the Bolzmann-rational policy is given by applying a softmax function to the optimal Q-function. We must therefore first decide what it means for a policy to be "optimal" in this setting. Because of temporal inconsistency, the ordinary notion of optimality does not automatically apply, and there are multiple ways to extend the concept. Accordingly, we introduce three new definitions:

Definition 4. A policy π is is *resolute* if there is no π' or ξ such that $V^{\pi,|\xi|}(\xi) < V^{\pi',|\xi|}(\xi)$.

A resolute policy maximises expected reward as calculated from the initial state. In other words, it effectively ignores the fact that its preferences might be changing over time, and instead always sticks to the preferences that it had at the start. In Example 1, a resolute policy would buy a gym membership, and then exercise.

Definition 5. A policy π is *naïve* if for each trajectory ξ, if $a \in \text{supp}(\pi(\xi))$, then there is a policy $\pi^\star$ such that $\pi^\star$ maximises $V^{\pi^\star,0}(\xi)$ and $a \in \text{supp}(\pi^\star(\xi))$.

A naïve policy ignores the fact that its preferences may not be temporally consistent. Rather, in each state, it computes a policy that is resolute from that state, and then takes an action that this policy would have taken, without taking into account that it may not actually follow this policy later. In Example 1, a naïve policy would buy a gym membership, but then go home without exercising.

Definition 6. A policy π is *sophisticated* if $\text{supp}(\pi(\xi)) \subseteq \text{argmax} Q^{\pi,0}(\xi, a)$ for all trajectories ξ.

A sophisticated policy is aware that its preferences are temporally inconsistent, and acts accordingly. Specifically, π is sophisticated if it only takes actions that are optimal *given that all subsequent actions are sampled from π*. In Example 1, a sophisticated policy would choose to not exercise in state s_1. Hence, in state s_0, it would realise that in s_1 it would go home, instead of exercising. Thus, in s_0 it prefers to go home over buying a gym membership and then going home, and it chooses to go home without buying a membership.

For consistency, if $d(t) = \gamma^t$ for some $\gamma \in (0,1)$, then Definitions 4-6 reduce to optimality. Formally:

Theorem 1. *In an MDP with exponential discounting, the following are equivalent: (1) π is optimal, (2) π is resolute, (3) π is naïve, and (4) π is sophisticated.*

All proofs are provided in the appendix. Notice that, while Definitions 4-6 are all equivalent under exponential discounting, they can be quite different if other forms of discounting are used, as already exemplified by Example 1. As such, each of these definitions give us a reasonable way to extend the notion of an "optimal" policy to the setting with general discount functions.

We next focus on the issue of *existence*, showing that *each* of these three types of policies are guaranteed to exist in *any episodic* MDP, regardless of what discount function is employed. Moreover, in each case, we show that this still holds if we restrict our attention to *deterministic* policies. We also show that naïve and sophisticated policies both are guaranteed to exist if we restrict our attention to *stationary* policies, but that there are episodic MDPs in which there are no stationary resolute policies.

Theorem 2. *In any episodic MDP, there exists a deterministic resolute policy.*

Theorem 3. *In any episodic MDP, there exists a stationary, deterministic, naïve policy.*

Theorem 4. *In any episodic MDP, there exists a stationary, deterministic sophisticated policy.*

Proposition 2. *There are episodic MDPs with no stationary resolute policies.*

Note that Proposition 2 is a consequence of the fact that non-exponential discounting can lead to preferences that are not temporally consistent, and that the agent may reach a given state at various time steps. Thus, the action taken by a resolute agent in that state may depend on the time at which it reaches that state. On the other hand, naïve and sophisticated agents have the same preferences regardless of what has happened in the past. Also note that a given reward function may allow for *several* policies that are resolute, naïve, or sophisticated: for example, all policies are both resolute, naïve, and sophisticated for a trivial reward function that is 0 everywhere. This lack of *uniqueness* should not be surprising.

To work out behavioural models that are analogous to Boltzmann-rationality, we must next develop analogies to the optimal Q-function for the non-exponential setting. For resolute and naïve policies, this is straightforward:

Definition 7. Given an episodic MDP, the *resolute Q-function* $Q^{\mathrm{R}} : \mathcal{S} \times \mathbb{N} \times \mathcal{A} \to \mathbb{R}$ is defined as

$$Q^{\mathrm{R}}(s,t,a) := \mathbb{E}_{S' \sim \tau(s,a)} \left[R(s,a,S') + \max_{\pi} V^{\pi,t+1}(S') \right].$$

Definition 8. Given an episodic MDP, the *naïve Q-function* $Q^{\mathrm{N}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as

$$Q^{\mathrm{N}}(s,a) := \mathbb{E}_{S' \sim \tau(s,a)} \left[R(s,a,S') + \max_{\pi} V^{\pi,1}(S') \right].$$

Note that the resolute Q-function must depend on the current time, since resolute policies may have to be non-stationary. Also note that $Q^{\mathrm{N}}(s,a) = Q^{\mathrm{R}}(s,0,a)$. We next show that these Q-functions are guaranteed to exist and to be unique in any episodic MDP. This means that we can talk about "the" resolute Q-function and "the" naïve Q-function for each given (episodic) MDP:

Proposition 3. *In any episodic MDP, the resolute Q-function Q^{R} exists and is unique.*

Proposition 4. *In any episodic MDP, the naïve Q-function Q^{N} exists and is unique.*

A policy π is resolute if and only if it only takes actions that maximise Q^{R}, and naïve if and only if it only takes actions that maximise Q^{N}. These Q-functions thus provide the appropriate generalisations of the optimal Q-function, $Q^\star$, corresponding to resolute and naïve policies respectively.

For sophisticated policies, the situation is more complicated. This is a consequence of the following fact:

Proposition 5. *There are episodic MDPs M with hyperbolic discounting and policies π_1, π_2 such that both π_1 and π_2 are sophisticated in M, but such that $Q^{\pi_1} \neq Q^{\pi_2}$, and such that the policy π_3 given by*

$$\mathbb{P}(\pi_3(s) = a) = \frac{\mathbb{P}(\pi_1(s) = a) + \mathbb{P}(\pi_2(s) = a)}{2}$$

is not sophisticated in M.

This implies that there need not be a unique "sophisticated Q-function" in a given episodic MDP. Moreover, unlike resolute and naïve policies (and optimal policies in exponentially discounted MDPs), the set of sophisticated policies is not (in general) convex. This means that we cannot create a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that a policy is sophisticated if and only if it only takes actions that maximise f. Nonetheless, we can still create a reasonable "canonical" sophisticated Q-function, though this will require a bit more work:

Definition 9. In an episodic MDP, a stationary sophisticated policy π is *canonical* if in all s we have that

$$\mathbb{P}(\pi(s) = a_1) = \mathbb{P}(\pi(s) = a_2)$$

for all $\{a_1, a_2\} \in \mathrm{argmax} Q^{\pi,0}(s,a)$. We the define the *sophisticated Q-function* Q^{S} as

$$Q^{\mathrm{S}} = Q^{\pi,0},$$

where π is the canonical sophisticated policy, provided that this policy exists and is unique.

Thus, a sophisticated policy is "canonical" if it is stationary, and if it mixes uniformly between all actions that have equal value. We next show that any episodic MDP always has a unique canonical sophisticated policy, which in turn means that Q^{S} is well-defined.[5]

Proposition 6. *In any episodic MDP, the sophisticated Q-function Q^{S} exists and is unique.*

[5]However, unlike what is the case for Q^{R} and Q^{N}, it is not the case that a policy is sophisticated if and only if it only takes actions that maximise Q^{S}. This is exemplified by the environment that is used in the proof of Proposition 5.

Given the above results, we are now finally equipped to define three behavioural models that generalise Boltzmann-rationality to non-exponential discounting:

Definition 10. Given an episodic transition function τ, discount d, and temperature $\beta \in (0, \infty)$, the *Boltzmann-resolute* behavioural model is the function $r_{\tau,d,\beta} : \mathcal{R} \to \Pi$ for which $r_{\tau,d,\beta}(R)$ is the policy π such that

$$\mathbb{P}(\pi(\xi) = a) \propto \exp \beta Q^{\mathrm{R}}(s, |\xi|, a),$$

where s is the last state in ξ.

Definition 11. Given an episodic transition function τ, discount d, and temperature $\beta \in (0, \infty)$, the *Boltzmann-naïve* behavioural model is the function $n_{\tau,d,\beta} : \mathcal{R} \to \Pi$ for which $r_{\tau,d,\beta}(R)$ is the stationary policy π such that

$$\mathbb{P}(\pi(s) = a) \propto \exp \beta Q^{\mathrm{N}}(s, a).$$

Definition 12. Given an episodic transition function τ, discount d, and temperature $\beta \in (0, \infty)$, the *Boltzmann-sophisticated* behavioural model is the function $s_{\tau,d,\beta} : \mathcal{R} \to \Pi$ for which $r_{\tau,d,\beta}(R)$ is the stationary policy π such that

$$\mathbb{P}(\pi(s) = a) \propto \exp \beta Q^{\mathrm{S}}(s, a).$$

In summary, we have devised three behavioural models that generalise the Boltzmann-rational one to general discount functions. It is of course an important question which of these behavioural models might provide the best fit to human behaviour: we consider this issue to be out of scope for this paper, and will instead analyse each of the behavioural models.

Partial Identifiability for General Discounting

Having laid the groundwork necessary to generalise the Boltzmann-rational behavioural model to non-exponential discounting, we are now able to characterise how ambiguous the reward function is for these new behavioural models. We will first provide an exact characterisation of this ambiguity, expressed in terms of necessary and sufficient conditions. Moreover, in order to interpret intuitively these results, we will also provide a number of results with more intuitive takeaways.

Exact Characterisation

If two reward functions R_1, R_2 have the property that $Q^\star_2(s, a) = Q^\star_1(s, a) - \Phi(s)$ for some function $\Phi : \mathcal{S} \to \mathbb{R}$, where $Q^\star_1$ and $Q^\star_2$ are the optimal Q-functions for R_1 and R_2, then those reward functions are said to differ by *potential shaping* (Ng, Harada, and Russell 1999).[6] It can be shown that Boltzmann-rational policies (with exponential discounting) determine R up to potential shaping (Skalse et al. 2022). We will show that this result generalises to the setting with non-exponential discounting, if the definition of potential shaping is adjusted appropriately. To state this result properly, we must first introduce two new definitions:

[6]Note that the definition in Ng, Harada, and Russell (1999) technically differs from this, but it can be shown to be equivalent.

Definition 13. Given an episodic MDP, we say that two reward functions R_1, R_2 differ by *sophisticated potential shaping* if there is a *potential function* $\Phi : \mathcal{S} \to \mathbb{R}$ such that

$$Q^{\mathrm{S}}_2(s, a) = Q^{\mathrm{S}}_1(s, a) - \Phi(s)$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, where Q^{S}_1 and Q^{S}_2 are computed relative to τ and d (for R_1 and R_2 respectively).

Definition 14. Given an episodic MDP, we say that two reward functions R_1, R_2 differ by *naïve potential shaping* if there is a *potential function* $\Phi : \mathcal{S} \to \mathbb{R}$ such that

$$Q^{\mathrm{N}}_2(s, a) = Q^{\mathrm{N}}_1(s, a) - \Phi(s)$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, where Q^{N}_1 and Q^{N}_2 are computed relative to τ and d (for R_1 and R_2 respectively).

Note that these two definitions do not state if such reward functions actually exist, nor do they state how to compute them. Our next result therefore shows that we always can find a reward R_2 that differs from R_1 by naïve or sophisticated potential shaping with Φ, for any R_1 and any Φ. The key insight is that this R_2 can be computed from R_1 and Φ via backwards induction, provided that the MDP is episodic.

Theorem 5. *For any episodic MDP with reward R_1 and for any potential function $\Phi : \mathcal{S} \to \mathbb{R}$, there exists a reward function R_2 that differs from R_1 by naïve potential shaping with Φ, and a reward function R_3 that differs from R_1 by sophisticated potential shaping with Φ.*

Using this, we can now provide an exact characterisation of the ambiguity of the Boltzmann-sophisticated and the Boltzmann-naïve behavioural model, which is applicable for any discount function. This is a core result:

Theorem 6. *In any episodic MDP, the Boltzmann-sophisticated policy determines R up to sophisticated potential shaping.*

Theorem 7. *In any episodic MDP, the Boltzmann-naïve policy determines R up to naïve potential shaping.*

It would be desirable to generalise these results to the Boltzmann-resolute behavioural model next. However, this presents a number of challenges, primarily stemming from the fact that the Boltzmann-resolute policy may be time-dependent. We detail and discuss these issues in the appendix.

It is important to note that whether or not two reward functions R_1 and R_2 differ by naïve or sophisticated potential shaping is relative to a given transition function, and that there is no simple closed-form expression for computing R_2 based on R_1 and Φ (unlike what is the case for potential shaping for exponential discounting, as introduced by Ng, Harada, and Russell 1999). The reason for this is that the optimal Q-function $Q^\star$ in the exponentially discounted case can be expressed by a local recursive equation (namely the Bellman optimality equation), but this is in general not possible for Q^{N} and Q^{S} with non-exponential discounting.

Qualitative Characterisation

We have provided an exact characterisation of the ambiguity of the underlying reward R given both naïve and sophisticated policies. However, these necessary and sufficient conditions can appear to be technically sophisticated. For this reason, we next provide a result that is easier to interpret qualitatively:

Theorem 8. *Let $f_{\tau,d,\beta}, g_{\tau,d,\beta} \in \{r_{\tau,d,\beta}, n_{\tau,d,\beta}, s_{\tau,d,\beta}\}$ be two behavioural models. Let d_1 and d_2 be any two discount functions, and let $\beta_1, \beta_2 \in (0, \infty)$ be any two temperature parameters. Then unless $d_1(t) = d_2(t)$ for all $t \leq |\mathcal{S}| - 1$, there exists an episodic transition function τ such that for any reward R_1 there exists a reward R_2 such that*

$$f_{\tau,d_1,\beta_1}(R_1) = f_{\tau,d_1,\beta_1}(R_2),$$

but such that

$$g_{\tau,d_2,\beta_2}(R_1) \neq g_{\tau,d_2,\beta_2}(R_2).$$

Moreover, unless $d(t) = \alpha \cdot \gamma^t$ for some $\alpha, \gamma \in [0, 1]$ and all $t \leq |\mathcal{S}| - 1$, we also have that, for any reward R_1 there exists a reward R_2, such that $f_{\tau,d_1,\beta_1}(R_1) = f_{\tau,d_1,\beta_1}(R_2)$, but such that R_1 and R_2 have different optimal policies under exponential discounting with γ.

Let us briefly unpack this result. Suppose R_1 is the true reward function, and that the training data for a given IRL algorithm is generated via f_{τ,d_1,β_1}. Suppose also that we want to use the learnt reward function to compute the output of a different behavioural model g_{τ,d_2,β_2}. If $f_{\tau,d_1,\beta_1}(R_1) = f_{\tau,d_1,\beta_1}(R_2)$, then the IRL algorithm may converge to R_2 instead of R_1, since they have identical f_{τ,d_1,β_1}-policies. However, if $g_{\tau,d_2,\beta_2}(R_1) \neq g_{\tau,d_2,\beta_2}(R_2)$, then R_1 and R_2 have different policies under g_{τ,d_2,β_2}. In other words, unless d_1 and d_2 are exactly equal over all time horizons that are possible in a given state space,[7] then the reward function is too ambiguous under f_{τ,d_1,β_1} to infer the correct value of g_{τ,d_2,β_1}. For example, this means that the Boltzmann-sophisticated policy for the hyperbolic discount function, or the Boltzmann-naïve policy for the bounded planning discount function, both leave the underlying reward too ambiguous to infer the Boltzmann-rational policy under exponential discounting, and so on. This suggests that the ambiguity of the reward can be problematic if we want to use the learnt reward to compute a policy using a discount function that is different from that used by the observed agent.

Note that Theorem 8 says that there exists *some* transition function τ for which this issue can occur. This does, by itself, not rule out the possibility that the ambiguity of R may be more modest for "typical" transition functions. Therefore, our next result applies to a very wide range of transition functions. We say that a non-terminal state s' is *controllable* if there is a state s and actions a_1, a_2 such that $\mathbb{P}(\tau(s, a_1) = s') \neq \mathbb{P}(\tau(s, a_2) = s')$, and that τ is *non-trivial* if it has at least one controllable state.

Theorem 9. *Let $f_{\tau,d,\beta}, g_{\tau,d,\beta} \in \{r_{\tau,d,\beta}, n_{\tau,d,\beta}, s_{\tau,d,\beta}\}$ be two behavioural models. Let d_1 and d_2 be any two discount functions, and let $\beta_1, \beta_2 \in (0, \infty)$ be any two temperature parameters. Let τ be any non-trivial episodic transition function. Then unless $d_1(1)/d_1(0) = d_2(1)/d_2(0)$, we have that for any reward R_1 there exists a reward R_2 such that*

$$f_{\tau,d_1,\beta_1}(R_1) = f_{\tau,d_1,\beta_1}(R_2),$$

but such that

$$g_{\tau,d_2,\beta_2}(R_1) \neq g_{\tau,d_2,\beta_2}(R_2).$$

Moreover, unless $d_1(1)/d_1(0) = \gamma$, we also have that there for any reward R_1 exists a reward R_2 such that $f_{\tau,d_1,\beta_1}(R_1) = f_{\tau,d_1,\beta_1}(R_2)$, but such that R_1 and R_2 have different optimal policies under exponential discounting with γ.

Nearly all transition functions are non-trivial, so Theorem 9 applies very broadly. Note that Theorem 8 makes weaker assumptions about the discount function but stronger assumptions about the transition function, whereas Theorem 9 makes stronger assumptions about the discount function but weaker assumptions about the transition function.

Discussion and Further Work

We have analysed partial identifiability in IRL with non-exponential discounting, including (but not limited to) hyperbolic discounting. To this end, we have introduced three types of policies (resolute policies, naïve policies, and sophisticated policies) that generalise the standard notion of optimality to non-exponential discount functions, and shown that these policies always exist in any episodic MDP. We have used these policies to generalise the Boltzmann-rational model to non-exponential discounting in three ways, and analysed the identifiability of the reward function under these models. We have demonstrated that the Boltzmann-naïve and the Boltzmann-sophisticated policies let us identify the true reward function up to naïve and sophisticated potential shaping, and shown that each of the three models in general is too ambiguous (even in the limit of infinite data) to compute the correct policy for a different form of discounting. We have thus made an important contribution to the study of partial identifiability in IRL, by extending existing results to the setting with non-exponential discounting. This is of particular importance, since hyperbolic discounting is considered to be a good fit to human behaviour.

There are several ways that our work can be extended. Improving our understanding of identifiability in IRL is of crucial importance, if we want to use IRL (and similar techniques) as a tool for preference elicitation. This analysis should consider behavioural models that are actually realistic. We have considered hyperbolic discounting, since this is widespread in the behavioural sciences, but there are many other ways to make our models more psychologically plausible. For example, it would be interesting to incorporate models of human risk-aversion, such as prospect theory (Kahneman and Tversky 1979). Moreover, our analysis is primarily restricted to *episodic* environments with a bounded horizon — it would be interesting to generalise it to broader classes of environments. This issue is further discussed in the appendix. It would also be interesting to exactly characterise the partial identifiability of the Boltzmann-resolute model — this issue is also discussed in the appendix. Finally, it would be interesting to study the case where the discount function is *misspecified*, i.e., where the IRL algorithm assumes that the observed agent discounts using some function d_1, but where it in fact discounts using some other function d_2 (see, e.g., Skalse and Abate 2023, 2024).

[7] Note that in an episodic MDP, any episode has length at most $|\mathcal{S}| - 1$. In other words, the horizon cannot exceed $|\mathcal{S}| - 1$.

References

Cao, H.; Cohen, S. N.; and Szpruch, L. 2021. Identifiability in Inverse Reinforcement Learning. *arXiv preprint*, arXiv:2106.03498 [cs.LG].

Chan, L.; Critch, A.; and Dragan, A. 2019. Irrationality can help reward inference.

Dvijotham, K.; and Todorov, E. 2010. Inverse Optimal Control with Linearly-Solvable MDPs. In *Proceedings of the 27th International Conference on Machine Learning*, 335–342. Haifa, Israel: Omnipress, Madison, Wisconsin, USA.

Evans, O.; Stuhlmueller, A.; and Goodman, N. D. 2015. Learning the Preferences of Ignorant, Inconsistent Agents. arXiv:1512.05832.

Frederick, S.; Loewenstein, G.; and O'Donoghue, T. 2002. Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, 40(2): 351–401.

Green, L.; and Myerson, J. 1996. Exponential versus hyperbolic discounting of delayed outcomes: Risk and waiting time. *American Zoologist*, 36(4): 496–505.

Haarnoja, T.; Tang, H.; Abbeel, P.; and Levine, S. 2017. 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*, 1352–1361. Sydney, Australia: PMLR.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative Inverse Reinforcement Learning. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Hussein, A.; Gaber, M. M.; Elyan, E.; and Jayne, C. 2017. Imitation Learning: A Survey of Learning Methods. *ACM Comput. Surv.*, 50(2).

Kahneman, D.; and Tversky, A. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2): 263–291.

Kim, K.; Garg, S.; Shiragur, K.; and Ermon, S. 2021. Reward Identification in Inverse Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5496–5505. Virtual: PMLR.

Kirby, K. 1997. Bidding on the Future: Evidence Against Normative Discounting of Delayed Rewards. *Journal of Experimental Psychology: General*, 126: 54–70.

Lattimore, T.; and Hutter, M. 2014. General time consistent discounting. *Theoretical Computer Science*, 519: 140–154. Algorithmic Learning Theory.

Mahadevan, S. 1996. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22(1): 159–195.

Mazur, J. 1987. An adjusting procedure for studying delayed reinforcement (Vol. 5). *Quant Anal Behav*, 55–73.

Metelli, A. M.; Lazzati, F.; and Restelli, M. 2023. Towards Theoretical Understanding of Inverse Reinforcement Learning.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 278–287. Bled, Slovenia: Morgan Kaufmann Publishers Inc.

Ng, A. Y.; and Russell, S. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, volume 1, 663–670. Stanford, California, USA: Morgan Kaufmann Publishers Inc.

Ramachandran, D.; and Amir, E. 2007. Bayesian Inverse Reinforcement Learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, 2586–2591. Hyderabad, India: Morgan Kaufmann Publishers Inc.

Schlaginhaufen, A.; and Kamgarpour, M. 2023. Identifiability and Generalizability in Constrained Inverse Reinforcement Learning. arXiv:2306.00629.

Schultheis, M.; Rothkopf, C. A.; and Koeppl, H. 2022. Reinforcement Learning with Non-Exponential Discounting. arXiv:2209.13413.

Skalse, J.; and Abate, A. 2023. Misspecification in Inverse Reinforcement Learning. arXiv:2212.03201.

Skalse, J.; and Abate, A. 2024. Quantifying the Sensitivity of Inverse Reinforcement Learning to Misspecification. arXiv:2403.06854.

Skalse, J.; Farrugia-Roberts, M.; Russell, S.; Abate, A.; and Gleave, A. 2022. Invariance in Policy Optimisation and Partial Identifiability in Reward Learning. *arXiv preprint arXiv:2203.07475*.

Strotz, R. H. 1955. Myopia and Inconsistency in Dynamic Utility Maximization. *The Review of Economic Studies*, 23(3): 165–180.

Thaler, R. 1981. Some empirical evidence on dynamic inconsistency. *Economics Letters*, 8(3): 201–207.