Identifiability in Inverse Reinforcement Learning

Daniil Dzenhaliou Dmitriy Gorovoy Haocong Li
Nevò Mirzai Hamadani Sean Park Filippo Passerini

Team: BRICS

November 13, 2025

Abstract

Inverse Reinforcement Learning (IRL) is fundamentally ill-posed: multiple reward functions can explain the same observed behavior. This report offers a cohesive survey of three pivotal advances—Kim et al. [1], which characterizes identifiability finite-horizon MDPs via graph coverability and aperiodicity; Cao et al. [2], which generalizes to entropy-regularized, stochastic MDPs in both finite and infinite horizons through value-distinguishability and multi-environment observations; and Rolland et al. [3], which provides a practical, rank-based criterion for identifiability and generalizability from multiple experts, including structured reward classes and robustness to transition estimation error. Building on Rolland et al Theorem 8 [3], we derive an explicit uniform bound on the reward estimation error, as a function of the spectral deviation between true and empirical transition matrices. Our analysis shows how estimation noise affects IRL and provides sample complexity bounds for reliable reward recovery.

1 Introduction and Background

Inverse Reinforcement Learning (IRL) aims to recover the reward function that explains expert behavior in a Markov Decision Process (MDP), assuming the expert acts (near-)optimally. Unlike standard reinforcement learning, which optimizes a known reward, IRL tackles the inverse problem of inferring the reward from observed trajectories. Once identified, this reward can predict expert behavior in new contexts, support skill transfer, automate complex tasks, or just make sure the agent behaves optimally for safety reasons. Applications range from surgical robotics [4] and autonomous driving [5] to aerobatic flight control [6].

A key challenge in IRL is identifiability: multiple rewards can explain the same behavior, complicating generalization and safety. While classic IRL is under-identified, recent works [1, 2, 3] have introduced conditions under which reward recovery becomes unique. This work formalizes and analyzes the ideas from these papers, provides a generalization for one of the results in paper [3], and concludes with some open questions for future research.

The environment. We consider a simple Markov decision process (MDP). The MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{T}_0, r, \gamma)$ is described by: a finite state space $\mathcal{S}$; a finite set of actions $\mathcal{A}$; a (Markov) transition kernel $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to P(\mathcal{S})$, that is, a function $\mathcal{T}$ such that $\mathcal{T}(s, a)$ gives probabilities[1] of each value of S_{t+1}, given the state $S_t = s$ and action $A_t = a$ at time t; initial state distribution $\mathcal{T}_0 \in P(\mathcal{S})$; and a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with discount factor $\gamma \in [0, 1)$. By the MDP model, we understand the set of polices induced by all reward functions on the given MDP (instead of policies, we will also consider trajectory distributions, which we define later).

Next, we define the regularized reward function, which, as we will see later, having defined the IRL problem formally, is crucial for the problem being well-posed.

Entropy-regularized MDP. Given the lack of smoothness in the classical MDP, and to encourage exploration, a well-known variation on the classic MDP, introduces a regularization term based on the Shannon entropy. Given a policy π and regularization coefficient $\lambda \geq 0$, the entropy regularized optimizing agent will use a policy π^* which solves the following equation

$$\max_{\pi} \mathbb{E}_s^{\pi} \left[\sum_{t=0}^{T} \gamma^t \left(r(s_t, a_t) - \lambda \log \pi_t(a_t \mid s_t) \right) \right]. \tag{1}$$

[1] Here, and elsewhere, we write $P(X)$ for the set of all probability distributions on a set X.

where T is the time horizon, which is not necessarily finite. We call this setting the MaxEnt MDP $M_\lambda = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \lambda)$. policies $\pi : \mathcal{S} \to P(\mathcal{A})$. This way, we ensure that a given reward function induces a unique optimal policy.

From a simple analysis by [2], we have the following observations regarding the regularized MDP:

- For $T = \infty$, the optimal policy will select all actions in A with some positive probabilities.
- For $T = \infty$ if λ is increased, this has the effect of "flattening out" the choice of actions in the optimal policy, and the regularized problem degenerates to the classical MDP. Conversely, sending $\lambda \to 0$ will result in a true maximizer being chosen, and the regularized problem degenerates to the classical MDP.
- Adding a constant to the reward does not change the policy.

One way to look at IRL is inverting the trajectory distribution to find the reward function which is inspired by [7]. A trajectory of length k is a sequence of state-actions, i.e., $\tau = (x_t, a_t)_{t=0}^k$ where for all $0 \le t \le k$, $x_t \in \mathcal{X}, a_t \in \mathcal{A}$. The trajectory distribution of a policy π executed in our domain with time horizon T is defined by:

$$p(\tau; \pi, T) = P_0(x_0)\pi(a_0 \mid x_0) \prod_{t=1}^{T} \pi(a_t \mid x_t) P(x_t \mid x_{t-1}, a_{t-1})$$

We define the corresponding optimal trajectory distribution for an RL task as $p_r(\tau; T) := p(\tau; \pi_r^*, T)$ where π_r^* is the optimal policy induced by the reward r. Also, when T is assumed to be known, we just write p_r. Note that in the definition above, the formula also depends on our MDP and on the regularized reward function, though not stated explicitly. We also denote by $\mathcal{S}^0$ the set of feasible initial states, i.e., $x \in \mathcal{S}^0 \Rightarrow \mathcal{T}_0(s) > 0$, and by $\Omega[s, T]$ the set of feasible trajectories of length T in our domain starting from initial state $x \in \mathcal{S}^0$, so $\tau' \in \Omega[s, T]$ if $\tau'_0 = x$ and there exists a policy π that can sample it, i.e., $p(\tau'; \pi) > 0$.

In practice, the data-generating process for an MDP model, knowing the optimal policy, might be sampling trajectories τ of length T from the optimal policy. So, one way to define IRL in a finite-horizon setting is seeking to invert the map $r \to p_r(\tau)$ up to some equivalence class (for instance, addition of a constant).

Another approach assumes that we can approximate the optimal policy and transition dynamics directly. In other words, it assumes that we can sample trajectories of arbitrarily large length $\tau = (s_1, a_1, s_2, a_2, \ldots)$. Then, assuming each state $s \in \mathcal{S}$ appears infinitely often in the sequence τ, and the agent uses a randomized feedback control $\pi_\lambda(a|s)$, it is possible to infer this control. A simple, consistent estimator for the control is

$$(\pi_\lambda)_N(a|s) = \frac{\#\{a_t = a \text{ and } s_t = s; \quad t \le N\}}{\#\{s_t = s; \quad t \le N\}} \to \pi_\lambda(a|s) \quad \text{a.s. as } N \to \infty.$$

Similarly, assuming each state-action pair (s, a) appears infinitely often in τ, we can infer the controlled transition probabilities $\mathcal{T}(s'|s, a)$. A simple, consistent estimator is given by

$$\mathcal{T}_N(s'|s, a) = \frac{\#\{s_t = s,\ a_t = a \text{ and } s_{t+1} = s'; \quad t \le N\}}{\#\{s_t = s \text{ and } a_t = a; \quad t \le N\}} \to \mathcal{T}(s'|s, a) \quad \text{a.s. as } N \to \infty.$$

If our agent is known to follow a regularized optimal strategy, as in, and $\mathcal{T}(s' \mid s, a) > 0 \quad$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. then every state-action pair will occur infinitely often in the resulting trajectory, and the formula above is true. So another approach is to define the IRL problem as inverting the map $r \to \pi_r^*$ up to the addition of a constant. Having described the intuition, we define the problem more formally now.

Definition 1 (π-Identifiability and p-Identifiability)**.** *An MDP model with finite horizon T π-**identifiable** up to an equivalence relation $\cong$ if for all $r, \hat{r} \in R$,*

$$r \cong \hat{r} \iff \pi_r^* = \pi_{\hat{r}}^*.$$

*An MDP model with finite horizon T is p-**identifiable** up to an equivalence relation $\cong$ if for all $r, \hat{r} \in R$,*

$$r \cong \hat{r} \iff p_r = p_{\hat{r}}.$$

So the difference is that in the first case we invert the policy and in the second the trajectory distribution (the letter p stands for p_r in the p-identifiable). The latter definition is useful in cases when we can not directly estimate the policy, not being able to sample infinitely long trajectories, but we have access to the trajectory distributions of length T, which is a little more delicate than in the infinite-horizon case, as it is necessary to observe many finite-horizon state-action trajectories, rather than a single

infinite-horizon trajectory. To finish the formal definition, we need to define the equivalence classes we are interested in. There are two natural ways to define those equivalence classes as defined in Kim et al. [1]:

$$r \cong_\tau \hat{r} \iff \forall x \in \mathcal{X}^0,\, \tau', \tau'' \in \Omega[x, d, T], \quad \hat{r}(\tau') - r(\tau') = \hat{r}(\tau'') - r(\tau'')$$

where $r(\tau) = \sum_{t=0}^{T} \gamma^t r(x_t, a_t)$. In other words, two rewards are trajectory equivalent, i.e., $r \cong_\tau \hat{r}$, if for a fixed trajectory τ, they are equal up to a constant after discounted summing over state-action pairs in trajectories starting from the same initial vertex. So, the two rewards represent the same preferences over trajectories. The second equivalence relation is a stronger one:

$$r \cong_{s,a} \hat{r} \iff \forall (s', a'), (s'', a'') \in \mathcal{S} \times \mathcal{A}, \quad \hat{r}(s', a') - r(s', a') = \hat{r}(s'', a'') - r(x'', a'')$$

Definition 2 (Weak Identifiability)**.** *An MDP model with finite horizon T is **weakly p-identifiable** if it is p-identifiable up to $\cong_\tau$, i.e. trajectory equivalence.*

Note that the notion of weakly identifiable MDPs for infinite horizons is ill-posed, as we can not define a distribution on infinite trajectories. However, we can define another equivalence class that is stronger and works for both horizon settings.

Definition 3 (Strong Identifiability)**.** *An MDP model is **strongly π-identifiable** if it is identifiable up to rewards shifted by a constant, i.e. $\cong_{x,a}$. Similarly, an MDP model with finite horizon is **strongly p-identifiable** if it is identifiable up to rewards shifted by a constant, i.e. $\cong_{x,a}$.*

As there are no versions of weakly π-identifiable MDPs for simplicity, we will sometimes use just π-identifiable instead of strongly π-identifiable MDPs.

In other words, identifiable means we recover the reward from the policy directly; p-identifiable means we recover the reward from the distribution over trajectories (in case we do not know the policy); weak identifiability means that from the expert's behavior one can recover the true preferences over trajectories (the reward totals for trajectories, up to a constant), whereas strong identifiability means one can recover the reward at the level of individual state-action pairs (again up to the same global constant). Strong identifiability is a stricter condition than weak identifiability [1].

As was described by Russel [8] IRL is an ill-posed problem when $r \to p_r$ or $r \to \pi_r^*$ is not injective with respect to r, i.e., many underlying rewards induce the same optimal policy. For instance, any action would be optimal for the constant zero reward function. Thus, it is necessary to consider regularization in order to have a unique policy for the given reward function.

Next, we review known results and each of the three papers. [1] focuses on p-identifiable MDP-s and finds the relation between weakly p-identifiable and strongly p-identifiable MDP-s. Papers [2] and [3] focus purely on strongly π-identifiable MDP-s. [2] provides several sufficient conditions for some types of MDPs in both finite and infinite horizon settings for strong π-identifiability. Besides, it introduces the theory of multiple agents on the same set of states and actions. Lastly, [3] finishes the theory on multiple experts.

2 Relations Between Weak and Strong p-Identifiabilities

We start with Kim et al. [1], which focuses on p-identifiable MDPs and its main result is finding the if and only if relation between weakly p-identifiable and strongly p-identifiable MDPs. First, we show an intuitively obvious result: strong identifiability implies the weak one for the finite horizon setting.

Lemma 1. *The MaxEnt MDP model is strongly p-identifiable only if it is weakly p-identifiable.*

Theorem 1 (Deterministic dynamics yields weak p-identifiability [1])**.** *Consider a MaxEnt MDP model where the transition dynamics are deterministic and the initial state is fixed i.e. $\forall (s, a)$, $|\mathrm{supp}\big(P(\cdot \mid s, a)\big)| = 1$; and $|\mathrm{supp}(P_0)| = 1$. Then the model is weakly p-identifiable.*

The idea of the proof is that for deterministic case we can explicitly write trajectory distribution p_r and check that $\forall r, \hat{r} \in R,\ (p_r = p_{\hat{r}}) \Rightarrow (r \cong_\tau \hat{r})$. The other direction is simple (attached in the appendix). Next we focus on finding the criteria when weak identifiability implies strong. For this, we define graphs that correspond to MDPs.

Definition 4. *A **domain graph** for a domain $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{T}_0, \gamma)$ is a tuple $G_d := (V_d, E_d, V_d^0)$ where:*

1. *$V_d := \mathcal{S} \times \mathcal{A}$ are the vertices.*
2. *$V_d^0 := \{(s, a) \mid \mathcal{T}_0(s) > 0\}$ are the initial vertices.*
3. *$E_d := \{e := (v, v') = ((s, a), (s', a')) \mid v, v' \in V_d,\ \mathcal{T}(s' \mid s, a) > 0\}$ are the edges.*

In words, the domain graph has a vertex for each state-action pair and a directed edge between vertices if the corresponding transition occurs with positive probability under the domain dynamics. Note that we consider a directed graph. Next, we operate with standard graph theoretical terms. We denote by ζ a path of length $k \geq 0$ in the domain graph, which is a sequence of vertices $\zeta := (v_t)_{0 \leq t \leq k}$ such that $(v_t, v_{t+1}) \in E_d$. A domain graph G_d is *strongly connected* if there exists a directed path between any two $v, v' \in V_d$. A *cycle* C is a path that starts and ends at the same vertex, i.e., $v_0 = v_k$ (Note that vertices can repeat). We say that a domain graph G_d is *aperiodic* if there does not exist $n > 1$ that divides the length of every cycle in the graph, and *periodic* otherwise.

Definition 5. *The k^{th} layer of a vertex $v \in V_d$, denoted $L_k(v)$, is the set of all vertices reachable in exactly k-steps from v, i.e., $L_k(v) = \{v' \in V_d \mid \exists \zeta = (v_t)_{0 \leq t \leq k}$ such that $v_0 = v$, $v_k = v'\}$. We define $L_0(v) = \{v\}$ and for $V \subseteq V_d$, $L_k(V) = \bigcup_{v \in V} L_k(v)$.*

Intuitively, the size of the layers $|L_k(v)|$ should grow with k as vertices further away from the initial vertices in V_d^0 become reachable, although this is not strictly true, e.g., bipartite graphs where certain vertices can only be reached in an odd or even number of steps. An important family of domain graphs are those that are *coverable*.

Definition 6. *A vertex $v \in V_d$ is said to be* t-covering *for $t \geq 1$ if $L_t(v) = V_d$. We say that a domain graph G_d is* t-coverable *(or just coverable) if there exists an initial vertex $v \in V_d^0$ that is t-covering.*

Proposition 1. *Let G_d be strongly connected. Then, G_d is aperiodic if and only if it is coverable.*

The intuition on how one could come to this result is imagining a periodic graph; it is kind of like an m-partite graph. Thus, it is possible to cover only the vertices from one of the parts in layer $L_k(v)$. The proof relies on the fact that, on the other hand, if there are two coprime cycles, then we can loop over them any number of times and thus get any natural number bigger than some constant. This would show that the graph is coverable. On the other hand, if it is t coverable, then it can be shown to be $t+1, t+2, \ldots$ coverable, which creates loops of size $t, t+1, t+2$ as we can come from vertex v to vertex v in t steps. Now we state the main result of this section and the article [1].

Theorem 2 (Strong Identifiability Condition [1])**.** *Consider a MaxEnt MDP model whose domain graph G_d is strongly connected. Then:*

- *(Sufficiency) MDP is weakly p-identifiable, G_d is T_0-coverable, and $T \geq 2T_0 \Rightarrow$ MDP is strongly p-identifiable.*
- *(Necessity) MDP is strongly p-identifiable $\Rightarrow$ it is is weakly p-identifiable and G_d is coverable.*

The necessity is the hard part, and it gives intuition why the coverability is important. We leave the formal proof in the appendix and give a general idea here. Showing the contrapositive, MDP must be strongly (and therefore weakly) p-identifiable but not coverable. And the idea now is to construct two reward functions (assuming that the graph is not coverable) that are trajectory equivalent but not (state, action) equivalent, which would lead to a contradiction. The key remark here is that assuming the graph G_d is not coverable for any vertex v, the sequence $(L_k(v))_{k\geq}$ is periodic from some moment. Besides the period and the layers $L_k(v)$ from some moment are the same for any vertex v, and starting from one vertex v, the trajectories τ, τ' in the k-th layer will have points from the same periodic layers. So if we have periodic layers $(L_t)_{t \geq 0}$. Let $r, \hat{r}$ be two rewards such that $\forall v \notin L_0, \hat{r}(v) = r(v)$ and $\forall v \in \bar{L}_0, \hat{r}(v) = r(v) + c$ for some constant $c \in \mathbb{R}$. On the one hand, since the graph is not coverable, the rewards are obviously not (state, action) equivalent. On the other hand, using properties of uncoverable graphs, we deduce that they are trajectory equivalent. The hard part is to formally prove that layers of not coverable graphs behave in a nice way.

We could also restate Theorem 2, changing coverability with apereodicity as those are equivalent by Proposition 1. This inspired Kim et al. to propose algorithm **MDPIdTest**, which tests if a given weakly p-identifiable MDP model with strongly connected domain graph is strongly p-identifiable by checking if it is aperiodic, which can be done in polynomial time $O(|E_d|)$. However, it is possible to improve the result for not necessarily strongly connected domain graphs, which we state in the appendix. This result inspires another algorithm **MDPCoverTest**, which tests if weakly p-identifiable MDP (with not necessarily strongly connected domain graph) is also strongly p-identifiable. It runs with time complexity $O(|V_d|^3 \log |V_d|)$ and space complexity $O(|V_d|^2)$. The algorithm, correctness, and efficiency of both MDPIdTest and MDPCoverTest are addressed in Appendix A.4.

In summary, paper [1] introduces elegant and fundamental theory on weakly and strongly p-identifiable MDPs. It finds the exact relation between them, which can be described in graph-theoretical terms. Though it is stated under the assumption of strong connectivity, we find the assumption natural. However, the paper's main weakness is the failure to classify weakly identifiable MDPs. The only result was achieved for deterministic MDPs, which is a rather simple class of MDPs. This leads to the following open problem to consider:

Open problem 1. *Let $\mathcal{M}$ be the class of finite-horizon MDPs $M = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, T)$. Characterize families $\mathcal{C} \subset \mathcal{M}$ (beyond deterministic models) such that: M is p-identifiable up to $\cong_\tau$, i.e. trajectory equivalence $\forall M \in \mathcal{C}$.*

3 π-Identifiability

While the previous section characterized p-identifiability through properties of the trajectory distribution and the domain graph, it requires access to full trajectory data. However, in practice, we often only observe policies — for example, an expert's behavior — rather than complete trajectory distributions.

This motivates the introduction of π-*identifiability*, where we focus on the problem of recovering a reward function from a given policy. In this section, we establish conditions under which inverse reinforcement learning (IRL) admits a solution based on policies, specifically considering action-independent rewards. We also discuss identifiability, that is, under what conditions the recovered reward is unique up to a constant shift.

3.1 Action-Independent Rewards: Existence and Identifiability

We say that an IRL problem admits a solution if for a given policy π, there exists at least one reward function for which π is the optimal policy. We begin by characterizing when such a solution exists for action-independent rewards.

Theorem 3. *The IRL problem admits a solution with an action-independent reward $r : \mathcal{S} \to \mathbb{R}$ if and only if the system of equations*

$$\lambda(\log \bar{\pi}(a) - \log \bar{\pi}(a_0)) = \gamma(\mathcal{T}(a) - \mathcal{T}(a_0))v, \quad \forall a \in \mathcal{A}, \tag{2}$$

admits a solution $v \in \mathbb{R}^{|\mathcal{S}|}$ for a fixed $a_0 \in \mathcal{A}$.

Note that the system above is of $|\mathcal{A}| \times |\mathcal{S}|$ equations in $|\mathcal{S}|$ unknowns, so this is a non-trivial assumption.

We now move to the identifiability question: when is the recovered reward unique up to a constant shift?

Corollary 1 (Identifiability Condition)**.** *Assume $\gamma \in [0, 1)$. If the solution to (2) exists, the MDP is π-identifiable if and only if*

$$\{c\mathbf{1} : c \in \mathbb{R}\} = \bigcap_{a \in \mathcal{A} \setminus \{a_0\}} \mathcal{K}(a) = \bigcap_{a \in \mathcal{A}} \mathcal{K}(a),$$

where $\mathcal{K}(a)$ is the kernel of $\mathcal{T}(a) - \mathcal{T}(a_0)$ and $\mathbf{1}$ denotes the all-one vector in $\mathbb{R}^{|\mathcal{S}|}$.

3.2 Time-Homogeneous Rewards: Full Rank Conditions

We extend the analysis to time-homogeneous settings. Specifically, we study when the reward function can be identified from a finite-horizon, entropy-regularized optimal policy π_t^* and terminal reward g.

Theorem 4 (Time-Homogeneous IRL Identifiability)**.** *Suppose the MDP has full action rank and full access at horizon T from an initial state s_0. Then the time-homogeneous MDP model is π-identifiable: knowledge of $\{\pi_t^*\}_{t=0}^{T-1}$ and g uniquely determines the running reward r up to a constant shift.*

3.3 Extensions to Deterministic and Stochastic MDPs

We further characterize the conditions under which full action rank holds in practical settings. In particular, for deterministic MDPs satisfying irreducibility and aperiodicity conditions, identifiability follows with a sufficiently large horizon T. For stochastic MDPs, identifiability is guaranteed if, in addition to previous conditions, at each state, the transition probability vectors under different actions span the entire set of reachable next states.

The detailed conditions and proofs are deferred to Appendix 3.

Limitations and Open Questions. While the π-identifiability framework provides a more practical approach by relying on policy observations, it has important limitations:

- The results often rely on strong structural assumptions, such as full action rank and full access, which may not hold in real-world applications.

Open problem 2. *Consider a family of POMDPs $\mathcal{P}_i = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}_i, Z_i, \gamma_i)$, where $\mathcal{S}$ is a (possibly continuous) state space, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, $\mathcal{T}_i(s' \mid s, a)$ is the state transition kernel in environment i, $Z_i(o \mid s)$ is the observation kernel in environment i, and $\gamma_i \in [0, 1)$ is the discount factor. Let a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$ be given, and parameterize the reward by $R_\theta(s, a) = \theta^\top \phi(s, a), \theta \in \mathbb{R}^m$. An entropy-regularized expert policy in $\mathcal{P}_i$ is*

$$\pi_i^* = \arg\max_\pi J^{\text{ent}}_{\mathcal{T}_i, Z_i, \gamma_i}(\pi; R_\theta),$$

where the objective is taken over trajectories of hidden states and observations.

For what conditions on the feature map ϕ, the transition kernels $\{\mathcal{T}_i\}$, the observation kernels $\{Z_i\}$, and discounts $\{\gamma_i\}$ does the collection of optimal policies $\{\pi_i^*\}_{i=1}^n \mapsto \theta$ become (almost) injective, up to potential-based shaping*?*

4 Multiple Experts Theory for π-Identifiability

[2] shows that observing a single expert makes it hard to determine the reward up to a constant and starts the theory on multiple experts which turned out to be much easier for π-identifiability problem. [3] finishes the theory on multiple experts. Here, we provide the full multiple experts theory combined from both papers.

Definition 7 (Value-distinguishing environments [2, 3])**.** *Two MDPs $(\mathcal{T}^1, \gamma^1)$ and $(\mathcal{T}^2, \gamma^2)$ (sharing S, A) are called value-distinguishing if the only pair of value functions $v_1, v_2 : S \to \mathbb{R}$ that can satisfy*

$$v_1(s) - \gamma^1 \sum s' \mathcal{T}^1(s'|s,a) v_1(s') = v_2(s) - \gamma^2 \sum_{s'} \mathcal{T}^2(s'|s,a) v_2(s') \quad \forall s \in S, \forall a \in A \tag{3}$$

implies at least one of v_1 and v_2 is a constant function.

[2] states that the reward function can be identified up to a constant from two expert policies, given that they satisfy the above value-distinguishing assumption. [3] directly references and builds upon this value-distinguishing condition from [2] since it is practically difficult to verify. For instance, Theorem 7 below gives an equivalent linear-algebraic condition that is easy to verify in practice.

To derive this condition, [3] uses the entropy-regularized optimality equation for each expert $i = 1, 2$:

$$r(s,a) = \lambda \log \pi_i(a|s) - \gamma_i \sum_{s' \in \mathcal{S}} \mathcal{T}_i(s'|s,a) v^i(s') + v^i(s) \tag{4}$$

where v^i is the optimal value function in environment i. Setting the right-hand sides for $i = 1$ and $i = 2$ equal (they both equal $r(s,a)$) and rearranging yields:

$$\lambda \log \pi_2(a|s) - \lambda \log \pi_1(a|s) = \Big(v_1(s) - \gamma^1 \sum s' \mathcal{T}^1(s'|s,a) v_1(s')\Big) - \Big(v_2(s) - \gamma^2 \sum_{s'} \mathcal{T}^2(s'|s,a) v_2(s')\Big), \tag{5}$$

for all s, a. [3] then considers the π-identifiability of the reward function from observing two expert policies using the rank condition of the equivalent matrix-vector form of (5) shown in Theorem 7. The details are addressed in Appendix C. Furthermore, [3] extends these π-identifiability conditions beyond just two experts to scenarios involving $N \geq 2$ experts. For multiple experts, the reward function is π-identifiable up to a constant if a similar rank condition holds for a concatenated matrix of N terms. The paper also investigates π-identifiability when experts operate in MDPs that only differ in their discount rates, demonstrating that conditions for reward identification can still be established in such cases.

Non-identifiability with unobserved factors. A negative result from [3] shows that the reward identifiability is not possible when exogenous variables are present in the MDP, no matter how many expert observations there are. This is because such variables, whose dynamics are independent of the agent's actions, introduce uncertainties that prevent the unique recovery of the underlying reward function.

Linear reward functions. However, as demonstrated in section 5.3 of [3] through numerical validation in the Strebulaev-Whited experiment with an exogenous state variable, expressing the true reward function as a linear combination of carefully chosen features can lead to exact π-identifiability. In this case, the reward function is represented as $r(s,a) = w^T f(s,a)$, and π-identifiability can be verified using Theorem 7 from [3], which is an identifiability condition for a linear combination of features.

Generalizability to new environments. Apart from π-identifiability, [3] introduces the concept of *reward generalizability*: given that we have learned a reward r from experts in certain environments, r will also produce an optimal policy in a new, unseen environment with different $\mathcal{T}$ and γ. They provide a condition under which the learned reward performs optimally in a new MDP.

Definition 8 (Generalization [3])**.** *Consider three Markov decision problems on the same set of states and actions, but with different transition matrices $\mathcal{T}^1, \mathcal{T}^2, \mathcal{T}^3$ and discount factors $\gamma_1, \gamma_2, \gamma_3$. Suppose that we observe two optimal entropy-regularized experts with respect to the same reward function in environments 1 and 2. We say that $(\mathcal{T}^1, \gamma_1)$ and $(\mathcal{T}^2, \gamma_2)$ generalize to $(\mathcal{T}^3, \gamma_3)$ if any reward compatible with the two experts in environments 1 and 2 leads to an optimal expert in environment 3. The definition naturally extends to more than two observed experts.*

Subsequently, the generalizablity of $(\mathcal{T}^1, \gamma_1)$ and $(\mathcal{T}^2, \gamma_2)$ to $(\mathcal{T}^3, \gamma_3)$ is again equivalent to a rank condition test. The details are available in Appendix C, Theorem 8. The intuition is that the third environment should not differ too much from the two observed environments.

Interestingly, reward generalizability can help address the challenges posed by exogenous variables. For instance, in the Windy-Gridworld experiment, where wind acts as an exogenous variable preventing reward identifiability, [3] shows that even though

the reward function is not identifiable, generalizability to a new environment with an arbitrary wind distribution works by observing enough experts in environments with different wind distributions.

The recovery of the reward function is addressed through the algorithms proposed by [3]. The general approach begins with verifying whether the rank condition is satisfied. Once it is met, we can solve for v^1, v^2 from (17). The reward function $r(s, a)$ can then be recovered using Equation (3). Likewise, the procedure for a generalizable environment is similar. After confirming that the corresponding rank condition holds, the reward function can be recovered, and then an optimal policy can be trained using any reinforcement learning algorithm. Although the recovery of the reward function requires multiple experts, [3] suggests that this setting can be alternatively interpreted as a single expert adapting within a single environment that undergoes slight variations over time, while the underlying reward function remains unchanged.

Limitations and Open Questions. A key limitation is that expert policies are optimal in order to get an identifiability condition and require strongly convex regularizers [3]. Consequently, the identifiability conditions presented in the paper do not apply to other common types of policies, such as deterministic greedy policies or epsilon-greedy policies. This is because these policies only provide actions that yield the highest expected reward, but no information about other non-optimal actions. The following open problem will be an interesting one to consider for future works:

Open problem 3. *Let $\mathcal{M}_i = (\mathcal{S}, \mathcal{A}, P_i, \gamma_i), \quad i = 1, \ldots, n$, be a family of MDPs with continuous (or very large) state space $\mathcal{S} \subseteq \mathbb{R}^d$ and action space $\mathcal{A} \subseteq \mathbb{R}^k$. We can consider the following two common reward-approximation schemes:*

1. ***Kernel (RKHS) rewards.*** *Let $\mathcal{H}$ be an RKHS with feature map $\Phi : \mathcal{S} \times \mathcal{A} \to \mathcal{H}$ and kernel k. Parameterize*
$$r_f(s, a) = \langle f, \Phi(s, a) \rangle_{\mathcal{H}}, \quad f \in \mathcal{H}.$$
2. ***Neural-network rewards.*** *Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$ be a fixed embedding, and let $f_W : \mathbb{R}^m \to \mathbb{R}$ be a neural network with weights W:*
$$r_W(s, a) = f_W\big(\phi(s, a)\big).$$

In each environment $\mathcal{M}_i$, define the entropy-regularized expert policy $\pi_i^* = \arg\max_\pi J^{\text{ent}}_{P_i, \gamma_i}(\pi, R)$, and let

$$\mathcal{F}_i : \begin{cases} \mathcal{H} \to \Pi, & f \mapsto \pi_i^* \quad (\textit{RKHS case}) \\ \mathbb{R}^W \to \Pi, & W \mapsto \pi_i^* \quad (\textit{NN case}) \end{cases}$$

where Π is the space of (stochastic) policies.

Theorem 7 from [3] shows that when $\mathcal{S} \times \mathcal{A}$ is finite and $\{\phi(s_j, a_j)\}$ has full column-rank, the linear-in-features case is identifiable up to potential-based shaping. How can one generalize these rank-based identifiability conditions *to the infinite-dimensional RKHS reward class $\{r_f : f \in \mathcal{H}\}$, the nonlinear neural-network class $\{r_W : W \in \mathbb{R}^W\}$, and continuous state/action domains, where no finite "feature matrix" exists?*

5 Estimated Reward Bound

Throughout this section and its corresponding appendix (Appendix D), we will use $||\cdot||$ as the 2-norm for vectors and the spectral norm for matrices, unless specified to be a different norm.

Approximate transitions. [3] considers the scenario where the transition matrices are approximated, and it provides a π-identifiability condition for this case. Specifically, the work characterizes reward identifiability when only approximate transition matrices are available. This is relevant because exact transition matrices are often unknown and must be estimated from a sample, and verifying conditions on approximated matrices can be misleading due to small perturbations.

In this section, we build upon the *approximate π-identifiability* result of Theorem 8 in Rolland *et al.* (2022) [3]. The theorem will be shown as Theorem 9 in the appendix. While this theorem ensures that the *value functions* and hence the *reward function* are *formally* identifiable from the approximate dynamics, it does *not* quantify

$$\big\|v - \hat{v}\big\| \quad \text{or} \quad \big\|r - \hat{r}\big\|_\infty$$

where (v, r) solve the true linear system $A\,v = b$, and $(\hat{v}, \hat{r})$ solve its perturbed counterpart $\hat{A}\,\hat{v} = b$, with $\hat{A}$ built from $\hat{T}_a^i$. In other words, Theorem 9 guarantees solvability and uniqueness of (v, r) but offers no bound on the *accuracy* of $\hat{r}$. So we naturally have the following open problem:

Open problem 4. *Provide bounds for $\big\|v - \hat{v}\big\|$ and $\big\|r - \hat{r}\big\|_\infty$.*

For $||v - \hat{v}||$ we have the following result:

$$\|v - \hat{v}\| \leq \frac{\|\hat{A}^+\|^2 \|\Delta_A\|}{1 - \|\hat{A}^+\| \|\Delta_A\|} \|b\|. \tag{6}$$

For the detailed derivation, please see Appendix D.

For $||r - \hat{r}||_\infty$, we derive explicit, non-asymptotic bounds of the form

$$\|r - \hat{r}\|_\infty \leq \Phi\big(\epsilon, \|\Delta_A\|, \|\hat{A}^+\|, \|b\|, \gamma\big)$$

where ϵ is the bound on the transition estimation: $\|\mathcal{T}_a^i - \hat{\mathcal{T}}_a^i\|_2 \leq \epsilon$, $\Delta_A := A - \hat{A}$ and $||A - \hat{A}|| = \|\Delta_A\| \leq \sqrt{2|\mathcal{A}|} \max(\gamma_1, \gamma_2)$ [3], $\hat{A}^+$ is the pseudoinverse of $\hat{A}$. Also, $b \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ where $b(s,a) = \lambda \log \frac{\pi^1(a|s)}{\pi^2(a|s)}$, ordered by states first. We have the following result:

$$\|r - \hat{r}\|_\infty \leq \gamma\epsilon \|\hat{A}^+\| \|b\| + (1+\gamma) \frac{\|\hat{A}^+\|^2 \|\Delta_A\|}{1 - \|\hat{A}^+\| \|\Delta_A\|} \|b\|. \tag{7}$$

We will also refer to Appendix D for a detailed derivation of this result.

5.1 Implications of the Reward-Error Bound

Our result (7) provides several practical insights:

- **Sensitivity to transition error ε.** The term $\gamma\,\varepsilon\,\|\hat{v}\|$ grows linearly in ε: if our estimated dynamics are off by at most ε in operator norm, the *direct* impact on the recovered reward scales as $\gamma\,\|\hat{v}\|$. Thus for small ε (high-quality transition estimates), this contribution is negligible.
- **Coupling via the linear-solver error.** The second term collects all effects of solving $\hat{A}\,\hat{v} = b$ instead of $A\,v = b$. It grows as $(1+\gamma)$ times the relative perturbation $\big(\|\hat{A}^+\|^2\|\Delta_A\|\big)/(1 - \|\hat{A}^+\|\|\Delta_A\|)$. In particular:
  - If $\hat{A}$ is well conditioned (small $\|\hat{A}^+\|$), then even moderate $\|\Delta_A\|$ causes only mild growth.
  - As $\|\Delta_A\|$ approaches the threshold $1/\|\hat{A}^+\|$, this term blows up, signaling loss of stability.
- **Trade–off between model and solver quality.** Improving transition estimates ($\varepsilon \downarrow 0$) not only reduces the first term, but also shrinks $\|\Delta_A\|$ and thus the second term. Conversely, if $\hat{A}$ is nearly singular, the reward estimate can be highly sensitive even to tiny ε.

6 Conclusion

In this report, we traced the evolution of identifiability in Inverse Reinforcement Learning through three seminal works. Kim et al. [1] established a graph-theoretic foundation for deterministic, finite-horizon MDPs, showing that coverability (aperiodicity) is necessary and sufficient for reward recovery up to natural constant shifts. Building on this, Cao et al. [2] generalized identifiability to entropy-regularized, stochastic MDPs across both finite and infinite horizons via the notion of value-distinguishability, and introduced multi-environment conditions to break the inherent "value" degeneracy. Finally, Rolland et al. [3] provided a practical, rank-based criterion for identifiability (and, where full recovery fails, generalizability) from multiple experts, extending the theory to structured feature-parametrized rewards and quantifying robustness to transition-matrix estimation error. Our contribution complements these advances by deriving an explicit, non-asymptotic sup-norm bound on the reward estimation error; thereby quantifying how perturbations in the estimated transition kernels and in the expert matrix propagate through the IRL pipeline. In our work we also identified the following open questions: (1) Can we find more families of MDPs that are p-weakly identifiable? (2) In many real-world tasks, the learner only sees observations, not full states. What identifiability and generalizability guarantees can be obtained under partial observability (POMDP) assumptions? (3) The theorems and rank conditions rely on tabular MDPs. Can analogous necessary and sufficient identifiability conditions be derived for continuous state/action spaces via kernel or neural representations, or linear-function or nonlinear (deep) function approximation settings? (4) Can we provide more information regarding the distance between the estimated and true value and reward functions? For the open question (4), we provided some bounds. For the other open questions, we think that they are the next steps to consider and solve.

References

[1] Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. Reward identification in inverse reinforcement learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5496–5505. PMLR, 18–24 Jul 2021.

[2] Haoyang Cao, Samuel N. Cohen, and Łukasz Szpruch. Identifiability in inverse reinforcement learning, 2021.

[3] Paul Rolland, Luca Viano, Norman Schürhoff, Boris Nikolov, and Volkan Cevher. Identifiability and generalizability from multiple experts in inverse reinforcement learning, 2022.

[4] Kunpeng Li and Joel W. Burdick. A function approximation method for model-based high-dimensional inverse reinforcement learning, 2017.

[5] Tuan Phan-Minh, Stephan Zheng, Adrián García, Victoria Silva, Paul Lu, Sujit Gujar, Jane Wang, and Fei Fang. Driving in real life with inverse reinforcement learning, 2022.

[6] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 1–8. ACM, 2004.

[7] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1433–1438. AAAI Press, 2008.

[8] Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 101–103, New York, NY, USA, 1998. Association for Computing Machinery.

[9] Eric V. Denardo. Periodicity in markov chains. *SIAM Journal on Applied Mathematics*, 32(1):1–12, 1977.

[10] John P. Jarvis and Douglas R. Shier. Graph-theoretic analysis of finite markov chains. *Handbook of Markov Decision Processes*, pages 217–249, 1999.

[11] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 3rd edition, 2009.

A Appendix: Proofs for p-Identifiable MDPs - section 2

A.1 Weak p-Identifiability

When s is omitted, $\Omega[T] = \bigcup_{s\in\mathcal{S}^0} \Omega[s,T]$ denotes the set of all feasible trajectories. Note that Ω also depends on MDP, though not written explicitly.

Lemma 1. *The MaxEnt MDP model is strongly p-identifiable only if it is weakly p-identifiable.*

Proof. We assume the opposite: either ($r \cong_\tau \hat{r}$ and $p_r \neq p_{\hat{r}}$) or ($r \ncong_\tau \hat{r}$ and $p_r = p_{\hat{r}}$).

First, we show that two trajectory equivalent rewards induce the same trajectory distribution p_r, which would imply that ($r \cong_\tau \hat{r}$ and $p_r \neq p_{\hat{r}}$) is impossible. $\forall\tau \in \Omega[s_0,d,T]$, $\hat{r}(\tau) = r(\tau) + c_{s_0}$. It suffices to show that the optimal policies for r, $\hat{r}$ are the same. For any policy family Π,

$$
\begin{aligned}
\arg\max_{\pi\in\Pi} \mathbb{E}_{\tau\sim\pi}[\hat{r}(\tau)] + \mathcal{H}_\lambda(\pi) &= \arg\max_{\pi\in\Pi} \left(\sum_{s_0\in\mathcal{S}^0} \sum_{\tau\in\Omega[s_0,d,T]} p(\tau;\pi)\hat{r}(\tau) \right) + \mathcal{H}_\lambda(\pi) \\
&= \arg\max_{\pi\in\Pi} \left(\sum_{s_0\in\mathcal{S}^0} \sum_{\tau\in\Omega[s_0,d,T]} p(\tau;\pi)(r(\tau) + c_{s_0}) \right) + \mathcal{H}_\lambda(\pi) \\
&= \arg\max_{\pi\in\Pi} \left(\sum_{s_0\in\mathcal{S}^0} \sum_{\tau\in\Omega[s_0,d,T]} p(\tau;\pi) r(\tau) \right) + \left(\sum_{s_0\in\mathcal{S}^0} P_0(s_0) c_{s_0} \right) + \mathcal{H}_\lambda(\pi) \\
&= \arg\max_{\pi\in\Pi} \mathbb{E}_{\tau\sim\pi}[r(\tau)] + \mathbb{E}_{s_0\sim P_0}[c_{s_0}] + \mathcal{H}_\lambda(\pi) \\
&= \arg\max_{\pi\in\Pi} \mathbb{E}_{\tau\sim\pi}[r(\tau)] + \mathcal{H}_\lambda(\pi)
\end{aligned}
$$

where $\mathcal{H}_\lambda(\pi) := \mathbb{E}_\pi \left[-\lambda \sum_{t=0}^{T} \gamma^t \log \pi(a_t|s_t) \right]$ is the γ-discounted causal entropy. The last step holds since $\mathbb{E}_{s_0\sim P_0}[c_{s_0}]$ is constant with respect to π. Finally, we claim that the objective from equation (1) has a unique maximizer due to regularization. Thus, we have equal polices.

In the second case, it is enough to show that (state, action) equivalence of two rewards implies the trajectory equivalence. This would show that as by the assumption our MDP is strongly p-identifiable it holds that $p_r = p_{\hat{r}} \iff r \cong_{(s,a)} \hat{r} \implies r \cong_\tau \hat{r}$. To prove this, it is enough just to write $r(\tau)$ by definition and check that it changes by a constant if we add a constant at every (state, action) point.

□

Theorem 1. *Consider a MaxEnt MDP model where the transition dynamics are deterministic and the initial state is fixed i.e.* $\forall(x,a)$, $|\mathrm{supp}\big(P(\cdot \mid x,a)\big)| = 1$*; and* $|\mathrm{supp}(P_0)| = 1$. *Then the model is weakly p-identifiable.*

Proof. We are to show that $\forall r,\hat{r} \in R,\ (r \cong_\tau \hat{r}) \iff (p_r = p_{\hat{r}})$. Since $\mathcal{P}_{\text{MDP}}$ is a MaxEnt MDP model, by proof of Lemma 1 it holds that $(r \cong_\tau \hat{r}) \Rightarrow (p_r = p_{\hat{r}})$. We are left to prove that $\forall r,\hat{r} \in R,\ (p_r = p_{\hat{r}}) \Rightarrow (r \cong_\tau \hat{r})$. Which appears to be simple as deterministic MDPs have a known trajectory distribution for optimal policies.

From Ziebart et al. [7], for all MDPs with deterministic dynamics and a deterministic initial state, the trajectory distribution of the MaxEnt optimal policy is $p_r(\tau) = \frac{e^{r(\tau)}}{Z_r}$ where $Z_r = \int_{\Omega[T]} e^{r(\tau')} d\tau'$ is the partition function. Then, for all $\tau \in \Omega[T]$,

$$
\begin{aligned}
p_r(\tau) &= p_{\hat{r}}(\tau) \\
\log p_r(\tau) &= \log p_{\hat{r}}(\tau) \\
r(\tau) - \log Z_r &= \hat{r}(\tau) - \log Z_{\hat{r}} \\
r(\tau) &= \hat{r}(\tau) + \log \frac{Z_r}{Z_{\hat{r}}}.
\end{aligned}
$$

Since $\log \frac{Z_r}{Z_{\hat{r}}}$ is a constant with respect to τ, we have $r \cong_\tau \hat{r}$ as desired. □

A.2 Properties of Domain Graphs

Lemma 2. *Let $G_d = (V_d, E_d, V_d^0)$ be a domain graph.*

1. *(Commutative) For all $V \subseteq V_d$ and $t, t' \geq 0$, $L_{t'}(L_t(V)) = L_{t+t'}(V)$.*
2. *(Monotonic) For all $V, V' \subseteq V_d$ such that $V \subseteq V'$ and $t \geq 0$, $L_t(V) \subseteq L_t(V')$.*

Proof. (Commutative) We first show that $L_{t'}(L_t(V)) \subseteq L_{t+t'}(V)$. Let $v \in L_{t'}(L_t(V))$. Then by definition, there exists a path of length t' such that it starts in $L_t(V)$ and ends in v, and there exists another path of length t from V to v'_0. Then we can concatenate these two paths to obtain a wanted path. Now, we prove $L_{t+t'}(V) \subseteq L_{t'}(L_t(V))$. Suppose $v \in L_{t+t'}(V)$. Then there exists a path of length $t + t'$ from V to v. We can split this path into two subpaths: the first one of length t (the first t edges). Then by definition of $L_t(V)$ we get the the first part ends exactly in $L_t(V)$

Monotonicity is obvious. □

Lemma 3. *Let G_d be a domain graph and $v \in V_d$ be t-covering. Then for all $t' \geq t$, $L_{t'}(v) = V_d$.*

Proof. We prove by induction. Base case: $t' = t$ trivially holds since $L_t(v) = V_d$ by definition of a covering vertex. Inductive step: assume $L_{t'}(v) = V_d$ for some $t' \geq t$. Then, $L_{t'+1}(v) = L_1(L_{t'}(v)) = L_1(V_d)$. We know that for every vertex there is at least one ingoing edge as it is coverable. Then $L_1(V_d) = V_d$ and the result follows by induction.

□

Proposition 1. *Let G_d be strongly connected. Then, G_d is aperiodic if and only if it is coverable.*

Proof. (aperiodic ⇒ coverable) If G_d is aperiodic, there exist two cycles of coprime lengths k, k'. The idea here is that we use strong connectivity and the fact that any number greater than $(k-1)(k'-1)$ can be represented as $ak + bk'$ for $a, b \geq 0$ integers. So for an arbitrary vertex v, we first go to the vertex in the first cycle. Then we make a loops around the first cycle. Then we go to a vertex from the second cycle and make b loops around it. Finally, we go to a vertex u. The path from v to the first cycle, from the first cycle to the second, and from the second to u exists as the graph is strongly connected (let their lengths be p_1, p_2, and p_3). Then for any natural number $N \geq (k-1)(k'-1) + p_1 + p_2 + p_3$ there exists a path from v to u. Thus, we have shown that for any vertex v, it can cover all vertices from V_d at some moment. Besides, the latest moment it can happen is at time $(k-1)(k'-1) + 3|V_d|$ as $p_i \leq |V_d|$.

(coverable ⇒ aperiodic) If G_d is coverable, there exists $v \in V_d^0$ and $t \geq 0$ such that $L_t(v) = V_d$. By the previous lemma, $L_{t+1}(v) = V_d$. Since $v \in L_t(v)$ and $v \in L_{t+1}(v)$, there exists cycles of coprime lengths t, $t+1$ that start and end at v. Repeating this argument, we will get cycles of length k and $k+1$ for $k \geq 2$ which are coprime. □

A.3 Strong p-Identifiability Criteria

Lemma 4. *Let the MDP model be weakly p-identifiable. Then, it is strongly p-identifiable if and only if for all $r, \hat{r} \in R$, $(r \cong_\tau \hat{r}) \Rightarrow (r \cong_{s,a} \hat{r})$. In other words, $\forall r \in R, [r]_\tau \subseteq [r]_{s,a}$.*

Proof. Let our MDP be weakly p-identifiable.

(Sufficiency) By assumption $(r \cong_\tau \hat{r}) \Rightarrow (r \cong_{s,a} \hat{r})$. By proof of Lemma 1, $(r \cong_{x,a} \hat{r}) \Rightarrow (r \cong_\tau \hat{r})$. So we get that $p_r = p_{\hat{r}} \iff r \cong_\tau \hat{r} \iff r \cong_{(s,a)} \hat{r}$ where the first follows by the assumption of weak p-identifiability.

(Necessity) Suppose there exists $r, \hat{r} \in R$ such that $r \cong_\tau \hat{r}$ but $r \ncong_{s,a} \hat{r}$. By weak p-identifiability, $(r \cong_\tau \hat{r}) \Rightarrow (p_r = p_{\hat{r}})$, so $(p_r = p_{\hat{r}}) \neq (r \cong_{s,a} \hat{r})$. Thus, MDP is not strongly p-identifiable.

□

Theorem 2 (Strong Identifiability Condition [1])**.** *Consider a MaxEnt MDP model whose domain graph G_d is strongly connected. Then:*

- *(Sufficiency) MDP is weakly p-identifiable, G_d is T_0-coverable, and $T \geq 2T_0 \Rightarrow$ MDP is strongly p-identifiable.*
- *(Necessity) MDP is strongly p-identifiable ⇒ it is is weakly p-identifiable and G_d is coverable.*

Proof. (Sufficiency) By lemma 4 we need to show that $r, \hat{r} \in R$, $(r \cong_\tau \hat{r}) \Rightarrow (r \cong_{s,a} \hat{r})$

In the language of domain graphs, we can translate weak p-identifiability using $r(\zeta)$ for some path ζ which is equivalent to discount trajectory reward along some τ. Let $r, \hat{r}$ be any two rewards such that $\hat{r}(\zeta) - \hat{r}(\zeta') = r(\zeta) - r(\zeta')$ for all paths ζ, ζ' that start from the same vertex from V_d^0.

Now by the assumption there exists v_0 that is T_0-covering which is $L_{T_0}(v_0) = V_d$. Now we inductively prove that for any two vertices in $L_t(v_0)$, the reward functions $r, \hat{r}$ on them differ by a constant. That is let H_t be the statement:

$$\forall v, v' \in L_t(v_0^*),\ \hat{r}(v') - \hat{r}(v) = r(v') - r(v).$$

- Base of the induction: H_0 holds trivially as $L_0(v_0) = \{v_0\}$.
- Step of the induction: let H_l holds for every $l < t \leq T_0$.

  By domain graphs properties $v_0 \in L_{T-t}(v_0)$ as $T - t \geq T_o$ and $L_{T_0}(v_0) = V_d$. So we fix some path ξ of length $T - t$ from v_0 to v_0. Now we consider all paths of length T that start with ξ. So first $T - t$ steps we go along the path ξ and then we go arbitrarily. Consider two such paths ζ, ζ'. For any function r we have $r(\zeta) - r(\zeta') = r(\xi) - r(\xi) + \gamma^{T-t}(r(\zeta_{T-t:T-1}) - r(\zeta'_{T-t:T-1})) + \gamma^T(r(v_T) - r(v'_T))$ where $\zeta_{T-t:T-1}$ denotes the edges from $T-t$ to $T-1$. So we split the path into three parts: the common part ξ; the $t-1$ edges, and the last vertex. Now we consider function $\hat{r}$ and get a similar expression $\hat{r}(\zeta) - \hat{r}(\zeta') = \gamma^{T-t}(\hat{r}(\zeta_{T-t:T-1}) - \hat{r}(\zeta'_{T-t:T-1})) + \gamma^T(\hat{r}(v_T) - \hat{r}(v'_T))$.

  But by induction hypothesis we know that $(r(\zeta_{T-t:T-1}) - r(\zeta'_{T-t:T-1}) = (\hat{r}(\zeta_{T-t:T-1}) - \hat{r}(\zeta'_{T-t:T-1})$. And by weak p-identifiability we have $r(\zeta) - r(\zeta') = \hat{r}(\zeta) - \hat{r}(\zeta') = r(v_T) - r(v'_T) = \hat{r}(v_T) - \hat{r}(v'_T)$.

Shortly what we've done: applied that $T \geq 2T_0$ by saying that there exist path from v_0 to v_0 of length $T - t$; then we split the left path of length t into a path of first $t - 1$ edges (along which the rewards coincide by induction H_l); and then we conclude H_t as by weak p-identifiability as the rewards are the same up to constant on the trajectories of length T and on the first $T - 1$ edges they are also the same by induction and the fact that first $T - t$ edges are identically the same.

(Necessity) The idea is that we want to show the contrapositive: MDP must be strongly (and therefore weakly) p-identifiable but not coverable. And the idea now is to construct two reward functions (assuming that the graph is not coverable) that are trajectory equivalent but not (state, action) equivalent, which would lead to a contradiction. The key remark here is that assuming the graph G_d is not coverable for any vertex v, the sequence $(L_k(v))_{k\geq}$ is periodic from some moment. Besides the period and the layers $L_k(v)$ from some moment are the same for any vertex v, and starting from one vertex v, the trajectories τ, τ' in the k-th layer will have points from the same periodic layers. So if we have periodic layers $(L_t)_{t\geq 0}$. Let $r, \hat{r}$ be two rewards such that $\forall v \notin L_0, \hat{r}(v) = r(v)$ and $\forall v \in \bar{L}_0, \hat{r}(v) = r(v) + c$ for some constant $c \in \mathbb{R}$. On the one hand, since the graph is not coverable, the rewards are obviously not (state, action) equivalent. On the other hand, using properties of uncoverable graphs, we deduce that they are trajectory equivalent. The hard part is to formally prove that layers of not coverable graphs behave in a nice way.

Lemma 5. *Let G_d be strongly connected. Then for all $v, v' \in V_d$ and $T \geq 0$, there exists $t \geq T$ such that $v' \in L_t(v)$.*

Proof of this lemma is a direct consequence of the fact that G_d is strongly connected. In particular, for every v there exists such T_v such that $v \in L_{T_v}(v)$.

Lemma 6. *Let G_d be connected. Then, for all $v \in V_d$, the sequence $\{L_t(v)\}_{t\geq 0}$ is* eventually periodic*, i.e., for all $v \in V_d$, there exist $T_v \geq 0$, $\delta_v \geq 1$ such that, for all $t \geq \bar{T}_v$, $L_t(v) = L_{t+\delta_v}(v)$.*

Proof. We first show that $L_{T_v n}(v)$ converges. For this, it is enough to prove that $(L_{nT_v}(v))_{n\geq 0}$ is "growing", i.e. $L_{nT_v}(v) \subseteq L_{(n+1)T_v}(v)$ for all $n \geq 0$ by induction. As it is bounded from above by V_d it would converge to $\bar{L}(v) \subseteq V_d$. We prove that it is growing by induction. The base case, when $n = 0$, holds trivially. Now assume for induction step that $L_{nT_v}(v) \subseteq L_{(n+1)T_v}(v)$. Then, $L_{(n+1)T_v}(v) = L_{T_v}(L_{nT_v}(v)) \subseteq L_{T_v}(L_{(n+1)T_v}(v)) = L_{(n+2)T_v}(v)$

So for all $v \in V_d$, there exists $\bar{n}_v \geq 0$ such that, for all $n \geq \bar{n}_v$, $L_{nT_v}(v) = \bar{L}(v)$. Then we see that for all $t \geq \bar{n}_v T_v$, it holds that

$$\begin{aligned}
L_{t+T_v}(v) &= L_{(t-\bar{n}_v T_v)+\bar{n}_v T_v+T_v}(v) \\
&= L_{(t-\bar{n}_v T_v)+(\bar{n}_v+1)T_v}(v) && (8)\\
&= L_{(t-\bar{n}_v T_v)}\big(L_{(\bar{n}_v+1)T_v}(v)\big) && (9)\\
&= L_{(t-\bar{n}_v T_v)}(L_{\bar{n}_v T_v}(v)) && (10)\\
&= L_{t-\bar{n}_v T_v+\bar{n}_v T_v}(v) \\
&= L_t(v).
\end{aligned}$$

where (8) $\to$ (9) holds since $(\bar{n}_v + 1)T_v \geq 0$ and $t - \bar{n}_v T_v \geq 0$ (apply commutative property of domain graphs). Furthermore, (9) $\to$ (10) holds by the convergence of $L_{nT_v}(v)$ since $L_{(\bar{n}_v+1)T_v}(v) = \bar{L}(v) = L_{\bar{n}_v T_v}(v)$. □

In other words, Lemma 6 states that the layers induced by starting at any vertex always converge to a periodic sequence.

Definition 9. *Let $(a_t)_{t\geq 0}$ be a sequence. We say that a sequence $(b_t)_{t\geq 0}$ is a* tail *of the sequence $(a_t)_{t\geq 0}$ if and only if there exists an index $N \geq 0$ such that $b_t = a_{t+N}$. Let $(a_t)_{t\geq 0}$ be an eventually periodic sequence. We say that a sequence $(b_t)_{t\geq 0}$ is a* periodic tail *of the sequence $(a_t)_{t\geq 0}$ if and only if $(b_t)_{t\geq 0}$ is a periodic sequence and a tail of $(a_t)_{t\geq 0}$.*

Lemma 7. *Let G_d be strongly connected. Let us denote $\bar{L}_t(v) := L_t(\bar{L}(v))$. Then, the sequence $(\bar{L}_t(v))_{t\geq 0}$ is a periodic tail of the sequence $\{L_t(v)\}_{t\geq 0}$. Let $\delta_v \geq 1$ denote the period of the tail sequence $(\bar{L}_t(v))_{t\geq 0}$ so that $\bar{L}_t(v) = \bar{L}_{t'}(v)$ for $0 \leq t < t'$ if and only if $(t' - t) \bmod \delta_v = 0$. Then $T_v \bmod \delta_v = 0$.*

Proof. From Lemma 6, $(L_{nT_v}(v))_{n\geq 0}$ converges to $\bar{L}_0$, so there exists $\bar{n}_v$ such that $L_{\bar{n}_v T_v}(v) = \bar{L}_0(v)$. Therefore, $L_{t+\bar{n}_v T_v}(v) = \bar{L}_t(v)$ and $(\bar{L}_t(v))_{t\geq 0}$ is a tail of the sequence $\{L_t(v)\}_{t\geq 0}$, besides, is periodic as $\big(\bar{L}_0(v) = \bar{L}_{T_v}(v)\big) \Rightarrow \big(\forall t \geq 0,\ \bar{L}_t(v) = \bar{L}_{T_v+t}(v)\big)$. The latter equality is true since $L_{(\bar{n}_v+1)T_v}(v) = L_{\bar{n}_v T_v}(v)$.

Now, for showing $T_v \bmod \delta_v = 0$ we first know that $T_v \geq \delta_v$ trivially holds since $\bar{L}_0(v) = \bar{L}_{T_v}(v)$. But then we have $\bar{L}_{q\delta_v}(v) = \bar{L}_{T_v}(v)$ so T_v must be divisible by δ_v by the assumption. □

Lemma 8. *Let G_d be strongly connected. Let $\delta_v \geq 1$ denote the period of the tail sequence $(\bar{L}_t(v))_{t\geq 0}$. Then for all $v \in V_d$ and $t \geq 0$, if $t \bmod \delta_v \neq 0$, then $v \notin \bar{L}_t(v)$.*

Proof. Due to periodicity, it suffices to prove that for all $v \in V_d$ and $0 < t < \delta_v$, $v \notin \bar{L}_t(v)$. Assume for contradiction that there exist $v \in V_d$ and $0 < t < \delta_v$ such that $v \in \bar{L}_t(v)$.

We then claim that $\bar{L}_0(v) \subseteq \bar{L}_t(v)$. If it is true, then we can easily finish the proof. In first case, if $\bar{L}_0(v) = \bar{L}_t(v)$ we get a shorter period t instead of δ_t.

If $\bar{L}_0(v) \subset \bar{L}_t(v)$, then for all $n \geq 1$, $\bar{L}_{nt}(v) \subseteq \bar{L}_{(n+1)t}(v)$ by simple argument with monotonicity and commutativity. This implies $\bar{L}_{nt}(v) \subset \bar{L}_{n't}(v)$. Choosing $n = 1$ and $n' = \delta_v$, we have $\bar{L}_0(v) = \bar{L}_{\delta_v}(v) \subset \bar{L}_{\delta_v t}(v)$ and so $\bar{L}_0(v) \neq \bar{L}_{\delta_v t}(v)$ which contradicts pereodicity.

Now we prove that $\bar{L}_0(v) \subseteq \bar{L}_t(v)$. By the previous lemma, we know that $T_v = q\delta_v$ where T_v is the smallest horizon after which v returns to itself. Then for all $n \geq 0$ $L_{nT_v}(v) \subseteq L_{nT_v}(\bar{L}_t(v)) = L_{nT_v}(L_t(v)) = L_{t+nT_v}(L_0(v)) = L_{t+nT_v}(v) = \bar{L}_{t+nq\delta_v}(v)$ which holds by the monotonicity and commutativity properties of domain graphs. By Lemma 6, there exists $\tilde{n}_v \geq 0$ such that, for all $n \geq \tilde{n}_v$, $L_{nT_v}(v) = \bar{L}_0(v)$. Combining this result with obtained inclusion, there exists $\tilde{n}_v \geq 0$ such that, for all $n \geq \tilde{n}_v$, $L_{nT_v}(v) = \bar{L}_0(v) \subseteq \bar{L}_{t+nq\delta_v}(v)$ which together with the assumption of periodicity gives the result: assume the contrary that inclusion does not hold and get a contradiction.

□

Lemma 9. *Let G_d be strongly connected. Let $\delta_v \geq 1$ denote the period of the tail sequence $(\bar{L}_t(v))_{t\geq 0}$. Then, for all $v \in V_d$ and $0 < t < t'$ such that $t' - t \bmod \delta_v \neq 0$, $\bar{L}_t(v) \cap \bar{L}_{t'}(v) = \emptyset$, i.e. limiting layers within a period are all disjoint sets regardless of the starting vertex.*

Proof. Assume for contradiction that there exists $0 \leq t < t' \leq \delta_v$ such that $0 < t' - t < \delta_v$, and a shared vertex $v_{t,t'} \in V_d$ such that $v_{t,t'} \in \bar{L}_t(v)$ and $v_{t,t'} \in \bar{L}_{t'}(v)$. Since G_d is strongly connected $v_{t,t'}$ can reach v and so there exists a l such that $v \in \bar{L}_{t+l}(v)$ and $v \in \bar{L}_{t'+l}(v)$. We now enumerate all cases for the value of $t + l$. At least one of the numbers $t + l$ and $t' + l$ is not congruent to zero modulo δ_v as $t, t' < \delta_v$ and are distinct. This, in turn, contradicts the previous lemma.

□

Lemma 10. *Let G_d be strongly connected. Then, for all $v, v' \in V_d$, the sequence $(\bar{L}_t(v))_{t\geq 0}$ is a periodic tail of the sequence $(L_t(v'))_{t\geq 0}$, i.e., vertex layers all converge to the same periodic sequence regardless of the starting vertex.*

Proof. Pick any $v, v' \in V_d$ and consider their corresponding periodic tails $(\bar{L}_t(v))_{t\geq 0}, (\bar{L}_t(v'))_{t\geq 0}$ (which exist by Lemma 8). Without loss of generality, we will let the first layer of the periodic tails be those containing the initial vertex, i.e., $v \in \bar{L}_0(v), v' \in \bar{L}_0(v')$. Such layers exist by previous lemmas.

Let $t_v, t_{v'} \geq 0$ denote the horizons at which $v' \in \bar{L}_{t_{v'}}(v), v \in \bar{L}_{t_v}(v')$ which exist by Lemma 5. Then, we claim $\bar{L}_0(v) \subseteq \bar{L}_{t_v}(v')$. To see this, first note that the sequence $(L_{nT_v}(v))_{n\geq 0}$, converges to $\bar{L}_0(v)$, where $T_v \geq 1$ is the shortest time horizon at which $v \in L_{T_v}(v)$. Furthermore, $(\bar{L}_{t_v+nT_v}(v'))_{n\geq 0} = (\bar{L}_{t_v}(v'))_{n\geq 0}$ since $(v \in \bar{L}_{t_v}(v'))$ by definition of t_v and $v \in \bar{L}_{t_v+nT_v}(v'))$ (as we can spend t_v steps for coming from v' to v and then nT_v steps for coming from v to v) we deduce by Lemma 9 that $\bar{L}_{t_v+nT_v}(v') = \bar{L}_{t_v}(v')$. Since $\{v\} \subseteq L_{T_v}(v)$, it follows from monotonicity that $L_{nT_v}(v) \subseteq L_{nT_v}(\bar{L}_{t_v}(v')) = \bar{L}_{t_v+nT_v}(v') = \bar{L}_{t_v}(v')$ for all $n \geq 0$ where we just apply previous remarks. Since $L_{nT_v}(v) = \bar{L}_0(v)$, we thus have $\bar{L}_0(v) \subseteq \bar{L}_t(v')$. The same argument can be applied to obtain $\bar{L}_0(v') \subseteq \bar{L}_t(v)$.

Assume that $\bar{L}_0(v') = \bar{L}_0(v)$, then it trivially follows that the sequences $(\bar{L}_t(v))_{t\geq 0}, (\bar{L}_t(v'))_{t\geq 0}$ are the same. So, let us assume further $\bar{L}_0(v') \neq \bar{L}_0(v)$, then $\bar{L}_t(v) = L_t(\bar{L}_0(v)) \subseteq L_t(\bar{L}_{t_v}(v')) = \bar{L}_{t_v+t}(v')$ for all $t \geq 0$ and $\bar{L}_t(v') = L_t(\bar{L}_0(v')) \subseteq L_t(\bar{L}_{t'_v}(v)) = \bar{L}_{t'_v+t}(v)$ for all $t \geq 0$. Thus $\bar{L}_0(v) \subseteq \bar{L}_{t_v}(v') \subseteq \bar{L}_{t_v+t'_v}(v)$. Then, $v \in \bar{L}_0(v) \Rightarrow v \in \bar{L}_{t_v+t'_v}(v)$ and it follows that $\bar{L}_0(v) = \bar{L}_{t_v+t'_v}(v)$ by by Lemma 9. Thus, $\bar{L}_0(v) = \bar{L}_{t_v}(v')$ which implies that $\bar{L}_t(v) = \bar{L}_{t_v+t}(v')$ for all $t \geq 0$ and so $(\bar{L}_t(v))_{t\geq 0}$ is a tail of $(\bar{L}_t(v'))_{t\geq 0}$.

□

From Lemma 10, we see that the layer sequence converges to the same periodic tail sequence regardless of the starting vertex. Thus, we shall henceforth denote a periodic tail of G_d as $(\bar{L}_t)_{t\geq 0}$, dropping the dependence on initial vertex.

Lemma 11. *Let G_d be strongly connected and let $(\bar{L}_t)_{t\geq 0}$ be a periodic tail of the layer sequences in G_d. For all $v \in V_d$ and $t, t' \geq 0$, $(L_t(v) \cap \bar{L}_{t'} \neq \emptyset) \Rightarrow (L_t(v) \subseteq \bar{L}_{t'})$*

Proof. Suppose for contradiction that there exists $v \in V_d$ and $t, t' \geq 0$ such that $(L_t(v) \cap \bar{L}_{t'} \neq \emptyset)$, but $(L_t(v) \nsubseteq \bar{L}_{t'})$. Let $v^- \in L_t(v) - \bar{L}_{t'}$ and $v^\cap \in L_t(v) \cap \bar{L}_{t'}$. The idea is to construct a limiting sequence that is not disjoint and is distinct from $\bar{L}_{t'}$, which would contradict Lemma 9.

Let $T_v \geq 1$ denote the smallest positive horizon such that $v \in L_{T_v}(v)$. Then, $L_{nT_v+t}(v) = L_t(L_{nT_v}(v)) \subseteq L_t(L_{(n+1)T_v}(v)) = L_{(n+1)T_v+t}(v)$ for all $n \geq 0$ by monotonicity since $L_{nT_v}(v) \subseteq L_{(n+1)T_v}(v)$ from the proof of Lemma 6. Thus, the sequence $(L_{nT_v+t}(v))_{n\geq 0}$ must converge to some fixed set $\bar{L}_{t^*}$ since the sequence is growing and bounded above. Thus, $\bar{L}_{t^*}$ is an element of the tail $(\bar{L}_t)_{t\geq 0}$. Since $v^-, v^\cap \in L_t(v)$, we have $v^-, v^\cap \in \bar{L}_{t^*}$. This contradicts Lemma 9 since $\bar{L}_{t'}, \bar{L}_{t^*}$ are two tail layers that are not the same but also not disjoint. □

We now prove the necessary direction of the main theorem. As already stated, it is enough to show that a strongly p-identifiable MDP (which is automatically also weakly identifiable) is not coverable, then we will get a contradiction.

Let $(\bar{L}_t)_{t\geq 0}$ be a periodic tail of the layer sequences in G_d. Let $r, \hat{r}$ be two rewards such that $\forall v \notin \bar{L}_0$, $\hat{r}(v) = r(v)$ and $\forall v \in \bar{L}_0$, $\hat{r}(v) = r(v) + c$ for some constant $c \in \mathbb{R}$. Since there does not exist a covering initial state, clearly, $\bar{L}_0 \subset V_d$ and thus $r \not\cong_{x,a} \hat{r}$. We will show now that $r \cong_\tau \hat{r}$.

For all $v \in V_d^0$ and for all paths $\zeta = (v_t)_{0\leq t\leq T}$, $\zeta' = (v'_t)_{0\leq t\leq T}$ starting from v, we claim that $\hat{r}(v_t) - \hat{r}(v'_t) = r(v_t) - r(v'_t)$ for all $0 \leq t \leq T$. To see this, first note that $v_t, v'_t \in L_t(v)$ for all $t \geq 0$. We consider two cases:

(1) If $v_t \in \bar{L}_0$, then $v'_t \in \bar{L}_0$ since $v_t, v'_t \in L_t(v)$ and, by Lemma 11, $(L_t(v) \cap \bar{L}_{t'}) \neq \emptyset \Rightarrow (L_t(v) \subseteq \bar{L}_{t'})$. Thus, $\hat{r}(v_t) - \hat{r}(v'_t) = r(v_t) + c - r(v'_t) - c = r(v_t) - r(v'_t)$.

(2) If $v_t \notin \bar{L}_0$, then $v'_t \notin \bar{L}_0$ by Lemma 11. Thus, $\hat{r}(v_t) - \hat{r}(v'_t) = r(v_t) - r(v'_t)$.

Then, $r(\zeta') - r(\zeta) = \sum_{t=0}^{T} \gamma^t \big(r(v'_t) - r(v_t)\big) = \sum_{t=0}^{T} \gamma^t \big(\hat{r}(v'_t) - \hat{r}(v_t)\big) = \hat{r}(\zeta') - \hat{r}(\zeta)$ which leads to contradiction. □

A.4 Strong p-Identifiability Test Algorithms

Before presenting the algorithms, we will first state some useful corollaries here that will be helpful in understanding the algorithms.

Corollary 2 (Strong Identification Condition)**.** *Consider a MaxEnt MDP model. If MDP is weakly p-identifiable, G_d is T_0-coverable, and $T \geq 2T_0 \Rightarrow$ MDP is strongly p-identifiable.*

Proof. The proof follows directly from the proof of sufficiency part from Theorem 2 where we did not use the fact the domain graph is strongly connected. □

We hereby present the MDPIdTest algorithm. The correctness and efficiency of the MDPIdTest are addressed in the following

Algorithm 1 Strong Identifiability Test for MDP models with Strongly Connected Domain Graphs

procedure MDPIDTEST(MDP model $\mathcal{M}$)
    Construct a domain graph $G_d = (V_d, E_d, V_d^0)$ from d.
    Set $gcd =$ PeriodFinder(V_d, E_d) [9].
    return $gcd == 1$
end procedure

theorem:

Theorem 3 ([1])**.** *Assume the MDP model is weakly p-identifiable, and G_d is strongly connected. Then, MDPIdTest returns True if and only if there exists some horizon T for which the MDP model is strongly p-identifiable. Moreover, it runs in time and space polynomial in $|S|$ and $|E|$ (number of states and transitions).*

Proof. For correctness, the procedure `MDPIdTest` returns *True* if and only if the domain graph G_d is aperiodic, as shown in [9, 10] using standard number theoretic and theoretical computer scientific techniques. Since, by assumption, the MDP model is weakly p-identifiable and G_d is strongly connected, the aperiodicity of G_d guarantees the existence of a transformation T such that the MDP model is strongly identifiable (by Theorem 2). For efficiency, graph aperiodicity testing can be performed in $\mathcal{O}(|E_d|)$ time and space, as discussed in [9, 10], also under standard theoretical computer science techniques. □

We also present the MDPCoverTest algorithm here. The correctness and efficiency of the MDPCoverTest are addressed in the

Algorithm 2 Strong Identifiability Sufficiency Test for General MDP models

procedure MDPCOVERTEST(MDP model $\mathcal{M}$)
    Construct a transition matrix M from d. $M_{ij} = \tilde{\mathcal{T}}(v^{(j)}|v^{(i)})$ where $\tilde{\mathcal{T}}(s', a'|s, a) = \mathcal{T}(s'|s, a)$.
    Compute $M^{|V_d|^2}$
    if *The rows for the initial vertices in $M^{|V_d|^2}$ contain only non-zero entries* **then**
        return 1
    else
        return 0
    end if
end procedure

following theorem.

Theorem 4 ([1])**.** *If the model is weakly p-identifiable and strongly connected, MDPCoverTest returns True if and only if the model is strongly p-identifiable (i.e., if G_d is T_0-coverable and $T \geq 2T_0$ for the given horizon T). It runs in $O(|S|^3 \log |S|)$ time.*

Proof. Let M be the transition matrix, i.e., $M_{ij} = \tilde{\mathcal{T}}(v^{(j)} \mid v^{(i)})$ where $\tilde{\mathcal{T}}(s', a' \mid s, a) = \mathcal{T}(s' \mid s, a)$. It follows that $M_{ij}^{|V_d|^2} \neq 0$ if and only if $v^{(j)} \in L_{|V_d|^2}(v^{(i)})$, meaning $v^{(j)}$ is reachable from $v^{(i)}$ within at most $|V_d|^2$ steps. If MDPCoverTest returns 1 (True), then there exists a node $v^{(i)} \in V_d^0$ such that the corresponding row in $M^{|V_d|^2}$ is fully non-zero, i.e., $L_{|V_d|^2}(v^{(i)}) = V_d$. Thus, G_d is $|V_d|^2$-coverable from $v^{(i)}$. Let $T_0 = 2|V_d|^2$. From Theorem 2, this implies that MDPCoverTest returns 1 if and only if G_d is $|V_d|^2$-coverable.

For efficiency, it is well known that computing powers of an $n \times n$ matrix A^m can be done in $O(n^3 \log m)$ time and $O(n^2)$ space [11]. Since M is of size $|V_d| \times |V_d|$, computing $M^{|V_d|^2}$ has time complexity

$$O\left(|V_d|^3 \log |V_d|^2\right) = O\left(|V_d|^3 \log |V_d|\right)$$

and space complexity $O(|V_d|^2)$. A naive method for checking whether a row is fully non-zero involves scanning all its entries, which can be done in $O(|V_d|^2)$ time and $O(1)$ space. Hence, the result follows. □

B Appendix: Proofs for π-identifiables MDPs - section 3

Theorem 5. *The IRL problem admits a solution with action-independent reward $r : \mathcal{S} \to \mathbb{R}$ if and only if the system of equations*

$$\lambda(\log \bar{\pi}(a) - \log \bar{\pi}(a_0)) = \gamma(\mathcal{T}(a) - \mathcal{T}(a_0))v, \quad \forall a \in \mathcal{A}, \tag{11}$$

admits a solution $v \in \mathbb{R}^{|\mathcal{S}|}$ for fixed $a_0 \in \mathcal{A}$.

Proof. **Necessity.** Suppose the IRL problem admits an action-independent solution $r : \mathcal{S} \to \mathbb{R}$. Then for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$r(s) = \lambda \log \bar{\pi}(a|s) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s, a)v(s') + v(s),$$

where v is the corresponding value function. Notice that for any $a \in \mathcal{A}$, for all $s \in \mathcal{S}$,

$$\begin{aligned} r(s) &= \lambda \log \bar{\pi}(a|s) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s, a)v(s') + v(s) \\ &= \lambda \log \bar{\pi}(a_0|s) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s, a_0)v(s') + v(s). \end{aligned}$$

Therefore, taking v to be the vector with components $v(s)$, we have a solution to the system of equations 11.

Sufficiency. Let v be a solution to the system of equations 11. By abuse of notation, we may write $v(s)$ for the components of v. Then for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\lambda \log \bar{\pi}(a|s) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s, a)v(s') = \lambda \log \bar{\pi}(a_0|s) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s, a_0)v(s').$$

Therefore, the quantity

$$\hat{r}(s) := \lambda \log \bar{\pi}(a_0|s) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s, a_0)v(s') + v(s)$$

is independent of a. From Theorem 1[2], we conclude that $\hat{r}$ is a solution to the IRL problem. □

Corollary 3. *Suppose $\gamma \in [0, 1)$. Assuming a solution to (11) exists, the IRL problem is identifiable (*i.e.*, the true action-independent reward function can be inferred up to a constant shift) if and only if, writing $\mathcal{K}(a)$ for the kernel of $(\mathcal{T}(a) - \mathcal{T}(a_0))$, we know that*

$$\{c\mathbf{1} : c \in \mathbb{R}\} = \bigcap_{a \in \mathcal{A} \setminus \{a_0\}} \mathcal{K}(a) = \bigcap_{a \in \mathcal{A}} \mathcal{K}(a),$$

where $\mathbf{1}$ denotes the all-one vector in $\mathbb{R}^{|\mathcal{S}|}$.

Note that $\{c\mathbf{1} : c \in \mathbb{R}\} \subset \mathcal{K}(a)$ for any $a \in \mathcal{A} \setminus \{a_0\}$ and $\mathcal{T}(a_0) = 0$ implies $\mathcal{K}(a) = \mathbb{R}^{|\mathcal{S}|}$.

Proof Sketch.

- By the Fredholm alternative, the solution set to (11) is the affine space $v_0 + \text{span}\left(\bigcap_{a \in \mathcal{A}} \mathcal{K}(a)\right)$ for any particular solution v_0.
- Identifiability requires that any two solutions differ only by a constant vector.
- If $\bigcap_a \mathcal{K}(a)$ contains only constant vectors, identifiability holds.

[1]The analogous results for finite-horizon problems with time-inhomogeneous rewards (and general discount factor) can be obtained through the same method.

- Otherwise, the presence of a non-constant vector in the intersection implies the existence of non-constant shifts, contradicting identifiability.

Proof. Let v_0 be a solution to (15), which is assumed to exist. By the Fredholm alternative the solution set $\mathbb{Y}_{\mathcal{S}}$ for 11 is given by

$$\mathbb{Y}_{\mathcal{S}} = \left\{ v_0 + \kappa : \kappa \in \text{span}\left(\bigcap_{a \in \mathcal{A}} \mathcal{K}(a) \right) \right\}.$$

From Theorem 5, the set of action-independent solutions for the IRL is given by

$$\mathbb{F}_{\mathcal{S}} = \left\{ r : r(s) = \lambda \log \bar{\pi}(a_0|s) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s, a_0) v(s') + v(s); \quad \text{for } v \in \mathbb{Y}_{\mathcal{S}},\ s \in \mathcal{S} \right\}.$$

We then observe that the stated condition is sufficient — if constant vectors are the only valid choices for κ, then v and hence $r \in \mathbb{F}_{\mathcal{S}}$ will only vary by constants.

To show necessity, denote by r_0 the solution corresponding to v_0. Suppose there exists a vector

$$\hat{v} \in \left(\bigcap_{a \in \mathcal{A}} \mathcal{K}(a) \right) \setminus \{ \mathbf{c1} : c \in \mathbb{R} \}.$$

Define

$$\Delta(s) = \hat{v}(s) - \gamma \sum_{s' \in \mathcal{S}} T(s'|s, a_0) \hat{v}(s'), \quad \forall s \in \mathcal{S}.$$

It follows that $f_0 + \Delta \in \mathbb{F}_{\mathcal{S}}$; if Δ is not a constant, we see that the reward is not uniquely identifiable.

To show Δ is not a constant, let

$$\overline{v} = \max_{s \in \mathcal{S}} \hat{v}(s), \quad \overline{s} \in \arg\max_{s \in \mathcal{S}} \hat{v}(s), \quad \underline{v} = \min_{s \in \mathcal{S}} \hat{v}(s), \quad \underline{s} \in \arg\min_{s \in \mathcal{S}} \hat{v}(s), \quad \tilde{v} = \frac{\sum_{s \in \mathcal{S}} \hat{v}(s)}{|\mathcal{S}|}.$$

Then $\underline{v} = \hat{v}(\underline{s}) < \tilde{v} < \hat{v}(\overline{s}) = \overline{v}$. We have

$$\begin{aligned} \Delta(\overline{s}) - (1-\gamma)\tilde{v} &= \overline{v} - \gamma \sum_{s \in \mathcal{S}} \mathcal{T}(s|\overline{s}, a_0)[\hat{v}(s) - \tilde{v}] \\ &\geq (1-\gamma)(\overline{v} - \tilde{v}) > 0; \\ \Delta(\underline{s}) - (1-\gamma)\tilde{v} &= \underline{v} - \gamma \sum_{s \in \mathcal{S}} \mathcal{T}(s|\underline{s}, a_0)[\hat{v}(s) - \tilde{v}] \\ &\leq (1-\gamma)(\underline{v} - \tilde{v}) < 0. \end{aligned}$$

Therefore, Δ is not a constant. It follows that our condition is necessary in order to have an identifiable action-independent reward. □

Theorem 6. *Suppose our MDP has full action rank and full access, at horizon $\mathcal{T}$, from an initial state s_0. Then the time-homogeneous IRL problem is well posed, that is, knowledge of the (time-dependent) entropy-regularized optimal strategy $\pi_t^*(a|s)$, and the terminal reward g, is sufficient to uniquely determine a time-homogeneous running reward r, if it exists, up to a constant.*

Conversely, if our MDP has full access but not full action rank at horizon $\mathcal{T}$, from the state s_0, the IRL problem remains ill-posed.

Proof Sketch.

- Use the entropy-regularized Bellman equations to express the running reward $r(s, a)$ in terms of $v_t(s)$ and $\pi_t^*(a|s)$.
- Set up a linear system for the value functions v_t based on observed optimal policies.
- Under the full action rank and full access assumptions, this linear system is invertible up to a constant shift, ensuring identifiability of f.

Proof. We first prove the sufficiency statement. The optimal policy satisfies

$$\lambda \log \pi_t^*(a|s) = Q_t^*(s,a) - V_t^*(s) = r(s,a) + \gamma \left(\sum_{s'} \mathcal{T}(s'|s,a) V_{t+1}^*(s') \right) - V_t^*(s).$$

We write (for notational simplicity), $v(s) = V_{\mathcal{T}-1}^*(s)$, and hence, given $V_T^* \equiv g$ by assumption,

$$r(s,a) = v(s) + \lambda \log \pi_{\mathcal{T}-1}^*(a|s) - \gamma \left(\sum_{s'} \mathcal{T}(s'|s,a) g(s') \right). \tag{12}$$

This shows that r is completely determined (if it exists) by the function v.

We also observe that for every t we have the recurrence relation

$$\begin{aligned} V_t^*(s) &= -\lambda \log \pi_t^*(a|s) + r(s,a) + \gamma \left(\sum_{s'} \mathcal{T}(s'|s,a) V_{t+1}^*(s') \right) \\ &= \lambda \log \frac{\pi_{t-1}^*(a|s)}{\pi_t^*(a|s)} + v(s) + \gamma \left(\sum_{s'} \mathcal{T}(s'|s,a) (V_{t+1}^*(s') - g(s')) \right). \end{aligned}$$

This holds for any choice of action a (unlike the usual dynamic programming relation, which only involves the optimal policy). Writing $\mathbf{V}_t$ for the vector with components $\{V_t^*(s)\}_{s \in \mathcal{S}}$ we have the recurrence relation

$$\mathbf{V}_t = \Upsilon_t(a) + v + \gamma \mathcal{T}(a) \mathbf{V}_{t+1}, \quad \mathbf{V}_{T-1} = v, \tag{13}$$

where Υ_t is a known vector-valued function, with components

$$[\Upsilon_t(a)]_s = \lambda \log \frac{\pi_{t-1}^*(a|s)}{\pi_t^*(a|s)} - \gamma \sum_{s'} \mathcal{T}(s'|s,a) g(s').$$

Solving the recurrence relation, we have, for any sequence of actions $a_0, \ldots, a_{\mathcal{T}-1}$,

$$\mathbf{V}_0 = \left(\sum_{t=0}^{\mathcal{T}-1} \left[\gamma^t \left(\prod_{t'=0}^{t-1} \mathcal{T}(a_{t'}) \right) \Upsilon_t(a_t) \right] \right) + \left(\sum_{t=0}^{\mathcal{T}-1} \gamma^t \left(\prod_{t'=0}^{t-1} \mathcal{T}(a_{t'}) \right) v \right) + \gamma^{\mathcal{T}} \left(\prod_{t'=0}^{\mathcal{T}-1} \mathcal{T}(a_{t'}) \right) g.$$

From this linear system, we can extract the single row corresponding to the fixed initial state s_0. Assuming this is the row indicated by the e_i basis vector, we have

$$V_0^*(s_0) = e_i^\top \left(\sum_{t=0}^{\mathcal{T}-1} \gamma^t \left(\prod_{t'=0}^{t-1} \mathcal{T}(a_{t'}) \right) \right) v + G(a_0, \ldots, a_{\mathcal{T}-1}), \tag{14}$$

for a known function G, expressible in terms of γ, g, and $\{\pi_t^*\}_{t=0}^{\mathcal{T}-1}$.

Now that the MDP has full action rank, the system of equations,

$$-G(a_0, \ldots, a_{\mathcal{T}-1}) = e_i^\top \left(\sum_{t=0}^{T-1} \gamma^t \prod_{t'=0}^{t-1} \mathcal{T}(a_{t'}) \right) v, \quad \forall a_0, \ldots, a_{\mathcal{T}-1},$$

admits at most one solution, denoted by $\bar{v}$. Substituting into 14, we have a unique solution to the equation $V_0^*(s_0) = 0$. However, we need to consider all possible values of $V_0^*(s_0)$.

For any choice of actions $\{a_t\}_{t=0}^{T-1}$,

$$e_i^\top \left[\sum_{t=0}^{T-1} \gamma^t \prod_{t'=0}^{t-1} \mathbb{T}(a_{t'}) \right] \mathbf{1} = \begin{cases} \frac{1-\gamma^T}{1-\gamma}, & \gamma \in (0,1), \\ T, & \gamma = 1, \end{cases}$$

where $\mathbf{1}$ denotes the all-one vector in $\mathbb{R}^N$. Therefore, the set of all possible $(V_0^*(s_0), v)$ pairs is given by

$$\begin{cases} \left\{ (c, \bar{v} + c\frac{(1-\gamma)}{1-\gamma}) : \forall c \in \mathbb{R} \right\}, & \gamma \in (0,1), \\ \left\{ (c, \bar{v} + c\mathbf{1}) : \forall c \in \mathbb{R} \right\}, & \gamma = 1, \end{cases}$$

From 12, we conclude that r can be identified up to a constant.

To show necessity, we observe from the above that, if the system is not full action-rank, then there exists a linear subspace of choices of v, which do not differ only by constants, such that we can construct the same value vectors $\mathbf{V}_t$ for all t, satisfying 13 and hence 12. It follows that we have a nontrivial manifold of rewards r which generate the same optimal policies, that is, the rewards are not identifiable. □

Corollary 4. *Suppose $\gamma \neq 0$ and our MDP is deterministic, that is $\mathcal{T}(s'|s,a) \in \{0,1\}$, and one of the following holds:*

1. *the underlying Markov chain is irreducible and aperiodic (i.e. with randomly chosen actions, the underlying Markov chain is irreducible and aperiodic)*
2. *the initial state $s_0 = e_i$ admits a self-loop (i.e. it is possible to transition from this state to itself), and all states can be accessed from the initial state in at most d transitions*
3. *there exist cycles[2] starting at the initial state $s_0 = e_i$ with lengths R, R', such that $\gcd(R, R') = 1$, and all states can be accessed from the initial state in at most d transitions.*

Then there exists a horizon $\mathcal{T}$ such that the time-homogeneous IRL problem is well posed (as in Theorem 4). In particular, in case (ii), it is sufficient to take any finite $\mathcal{T} \geq d+1$; in case (iii), it is sufficient to take any finite $\mathcal{T} \geq d + RR'$.

Proof. We first observe that it is a classical result on Markov chains (see, for example, [Seneta, 2006, Theorem 1.5]) that the conditions of case (i) guarantee those of case (iii), for some choice of $R, R' > 0$. The conditions of case (ii) also guarantee those of case (iii), with both the cycles being the self-loop. It is therefore sufficient to consider case (iii).

To show that the MDP has full action rank, we observe that for every possible path of states, there exists a corresponding sequence of actions, and vice versa. We will therefore use these different perspectives interchangeably. We also observe that, as our MDP is deterministic, $e_i^\top \prod_{t=0}^{\mathcal{T}-1} \mathcal{T}(a_t)$ is a vector indicating the current state at time t, when started in state e_i. Therefore,

$$\mathbb{O}_{\mathcal{T}-1}(\{a_t\}_{t\geq 0}) := e_i^\top \left(\sum_{t=0}^{\mathcal{T}-1} \gamma^t \prod_{t'=0}^{t-1} \mathcal{T}(a_{t'}) \right)$$

is a row vector, containing a time-weighted occupation density. In particular, if $\gamma = 1$, it simply counts the number of times we have entered each state. Our aim, therefore, is to construct a collection of paths that gives a full-rank system of occupation densities.

Starting in state $s_0 = e_i$, consider a shortest path (i.e. a path with the fewest number of transitions) to each state s'. Denote these paths $r_{s'} = \{s_0 \to \cdots \to s'\}$, and the corresponding sequence of actions $a^{s'}$. These paths have lengths $|r_{s'}|$ and time-weighted occupation densities $\mathbb{O}_{|r_{s'}|}(\{a_t^{s'}\}_{t\geq 0})$ which are linearly independent. This gives us $N = |\mathcal{S}|$ paths.

We now consider prefixing our paths with cycles, in order to make them the same length. Fix an arbitrary integer value $T' \geq \max_{s'} |r_{s'}| + |Q||Q'| - 1$. By Lemma 1[2], for all states s, there exist nonnegative integers λ_s, μ_s such that $\mathcal{T}' = \lambda_s|Q| + \mu_s|Q'| + |r_{s'}|$. Therefore, taking the concatenated path consisting of λ_s repeats of cycle Q, then μ_s repeats of cycle Q', then our shortest path $r_{s'}$, gives us a path from s_0 to s of length $\mathcal{T}'$. Denote each of these paths P_s.

The concatenation of paths has an elegant effect on the occupation densities. If Q is a cycle and r a path, then:

$$\mathbb{O}_{|Q*r|}(\{a_t^{Q*r}\}_{t\geq 0}) = \mathbb{O}_{|Q|-1}(\{a_t^Q\}) + \gamma^{|Q|}\mathbb{O}_{|r|}(\{a_t^r\}). \tag{15}$$

We now observe that for the initial state, the shortest path is of length zero (i.e. has no transitions). From Lemma 1[2], as $\mathcal{T}' \geq |Q||Q'|$, we know that there are multiple choices of λ, μ satisfying the stated construction, and therefore there are at least two possible paths P_{s_0} and $\tilde{P}_{s_0}$ with the desired length, from the initial state to itself, using distinct numbers of cycles $(\lambda_{s_0}, \mu_{s_0})$ and $(\tilde{\lambda}_{s_0}, \tilde{\mu}_{s_0})$.

[2] If the cycles are both a self-loop, then this becomes degenerate, but in the following step, the final column and row of the matrix M can be omitted, and the remainder of the argument follows in essentially the same way.

This construction yields a collection of paths with full rank occupation densities. To verify this explicitly, extract the rows corresponding to the paths $\{R_s\}_{s\in\mathcal{S}}$ and $\tilde{R}_{s_0}$, we use 15 to see that

$$\begin{bmatrix} \mathbb{O}_{\mathcal{T}-1}(\{a_t^{P_{s_0}}\}_{t\geq 0}) \\ \mathbb{O}_{\mathcal{T}-1}(\{a_t^{P_{s_1}}\}_{t\geq 0}) \\ \vdots \\ \mathbb{O}_{\mathcal{T}-1}(\{a_t^{P_{s_{N-1}}}\}_{t\geq 0}) \\ \mathbb{O}_{\mathcal{T}-1}(\{a_t^{\tilde{P}_{s_0}}\}_{t\geq 0}) \end{bmatrix} = M \begin{bmatrix} \mathbb{O}_{|r_{s_0}|}(\{a_t^{r_{s_0}}\}_{t\geq 0}) \\ \mathbb{O}_{|r_{s_1}|}(\{a_t^{r_{s_1}}\}_{t\geq 0}) \\ \vdots \\ \mathbb{O}_{|r_{s_N}|}(\{a_t^{r_{s_N}}\}_{t\geq 0}) \\ \mathbb{O}_{|Q|-1}(\{a_t^{Q}\}_{t\geq 0}) \\ \mathbb{O}_{|Q'|-1}(\{a_t^{Q'}\}_{t\geq 0}) \end{bmatrix}, \tag{16}$$

where

$$\Gamma(\lambda, Q) := \begin{cases} \frac{(1-\gamma^{|\lambda Q|})}{(1-\gamma^{|Q|})}, & \gamma \neq 1, \\ \lambda, & \gamma = 1, \end{cases}$$

$$M = \begin{bmatrix} \gamma^{T'} & 0 & \cdots & 0 & \Gamma(\lambda_{s_0}, Q) & \gamma^{\lambda_{s_0}|Q|}\Gamma(\mu_{s_0}, Q') \\ 0 & \gamma^{T'-|r_{s_1}|} & \cdots & 0 & \Gamma(\lambda_{s_1}, Q) & \gamma^{\lambda_{s_1}|Q|}\Gamma(\mu_{s_1}, Q') \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma^{T'-|r_{s_N}|} & \Gamma(\lambda_{s_N}, Q) & \gamma^{\lambda_{s_N}|Q|}\Gamma(\mu_{s_N}, Q') \\ \gamma^{T'} & 0 & \cdots & 0 & \Gamma(\lambda_{s_0}, Q) & \gamma^{\tilde{\lambda}_{s_0}|Q|}\Gamma(\tilde{\mu_{s_0}}, Q') \end{bmatrix}.$$

After subtracting the first from the last row of M, as $\lambda_{s_0} \neq \tilde{\lambda}_{s_0}$, we see that M has a simple structure, in particular it is a full-rank matrix with $N+1$ rows and $N+2$ columns. As the final matrix on the right-hand side of 16 is of rank N, this implies that the left-hand side of 16 is also of rank N (by Sylvester's rank inequality). As the left-hand side of 16 is a selection of rows from the matrix considered in Definition 3[2], we conclude that our MDP must be of full action rank.

Our collection of paths also shows that our system has full access at horizon $\mathcal{T} = \mathcal{T}' + 1$, and therefore the identification result follows from Theorem 6. By varying $\mathcal{T}'$, we see this result holds for any choice of $\mathcal{T} \geq |Q||Q'| + \max_s |r_s|$, as desired. □

Corollary 5. *Suppose $\gamma \neq 0$, and our MDP is stochastic and satisfies one of the sets of assumptions ((i), (ii) or (iii)) of Corollary 1 and that from every state, we have at least as many actions (with linearly independent resulting transition probabilities) as we have possible future states, that is,*

$$\textit{rank}\,\{\mathcal{T}(\cdot|s,a) : a \in \mathcal{A}\} = \#\,\{s' : \mathcal{T}(s'|s,a) > 0 \textit{ for some } a \in \mathcal{A}\}\,.$$

Then for any initial state s_0, there exists a horizon T such that the time-homogeneous IRL problem is well posed (as in Theorem 6). The sufficient bounds on $\mathcal{T}$ from Corollary 3 also apply.

Proof. For a given state s, consider the space spanned by the basis vector corresponding to the possible future states. Given we have as many actions as possible future states, and the rank-nullity theorem, we know that this space must be the same as the space spanned by the vectors $\{\mathcal{T}(\cdot|s,a) : a \in \mathcal{A}\}$. In particular, there exists a set of weights c_a over actions (which do not need to sum to one or be nonnegative) such that $\sum_{a\in\mathcal{A}} c_a \mathcal{T}(a)$ is the basis vector corresponding to any possible transition.

In other words, there is no difference between the linear span generated by these stochastic transitions and deterministic transitions. As actions at every time can be varied independently, and the requirement that an MDP has full action rank depends only on the space spanned by transition matrices, the problem reduces to the setting of Corollary 3. □

C Appendix: Details Behind Multiple Experts Identifiability

C.1 Identifiability from Two Experts

In vector-matrix form, denote by T_a^i the transition matrix for action a in environment i, and by v_i the value vector. Then (5) can be written as:

$$\begin{pmatrix} I - \gamma_1 \mathcal{T}_{a_1}^1 & -(I - \gamma_2 \mathcal{T}_{a_1}^2) \\ \vdots & \vdots \\ I - \gamma_1 \mathcal{T}_{a_{|\mathcal{A}|}}^1 & -(I - \gamma_2 \mathcal{T}_{a_{|\mathcal{A}|}}^2) \end{pmatrix} \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \begin{pmatrix} \lambda \log \pi^2(\cdot|a_1) - \lambda \log \pi^1(\cdot|a_1) \\ \vdots \\ \lambda \log \pi^2(\cdot|a_{|\mathcal{A}|}) - \lambda \log \pi^1(\cdot|a_{|\mathcal{A}|}) \end{pmatrix}, \tag{17}$$

This linear system is used to solve for v^1 and v^2 given the observed policies. Theorem 3 from [3] then states that the reward function is identifiable up to a constant if the rank of the matrix associated with this form of the equivalence condition satisfies:

Theorem 7 (Reward Identifiability with Two Experts [3]). *The reward function is identifiable up to a constant from observing two expert policies if and only if:*

$$\textit{rank}\begin{pmatrix} I-\gamma_1\mathcal{T}^1_{a_1} & I-\gamma_2\mathcal{T}^2_{a_1} \\ \vdots & \vdots \\ I-\gamma_1\mathcal{T}^1_{a_{|\mathcal{A}|}} & I-\gamma_2\mathcal{T}^2_{a_{|\mathcal{A}|}} \end{pmatrix} = 2|\mathcal{S}|-1. \tag{18}$$

Proof Sketch. Denote A as the left-hand side matrix of (17). One checks that

$$A\begin{pmatrix}\mathbf{1}/(1-\gamma_1)\\ \mathbf{1}/(1-\gamma_2)\end{pmatrix}=0,$$

so $\begin{pmatrix}\mathbf{1}/(1-\gamma_1)\\ \mathbf{1}/(1-\gamma_2)\end{pmatrix}$ is in the kernel of A. This means that $\dim\ker A \geq 1$. Then, if $\text{rank}A = 2|\mathcal{S}|-1$ using the rank-nullity theorem, there are no other linearly independent vectors in $\ker A$. Therefore, $\dim\ker A = 1$, leading to a solution up to a constant shift. If $\text{rank}A < 2|\mathcal{S}|-1$, then $\dim\ker A > 1$, leading to more complex non-identifiabilities beyond a constant shift.

C.2 Generalizability

Below is a rank condition that checks generalizability:

Theorem 8 (Reward generalizability condition [3]). *$(\mathcal{T}^1,\gamma_1)$ and $(\mathcal{T}^2,\gamma_2)$ generalize to $(\mathcal{T}^3,\gamma_3)$ if and only if:*

$$\begin{pmatrix} I-\gamma_1\mathcal{T}^1_{a_1} & I-\gamma_2\mathcal{T}^2_{a_1} \\ \vdots & \vdots \\ I-\gamma_1\mathcal{T}^1_{a_{|\mathcal{A}|}} & I-\gamma_2\mathcal{T}^2_{a_{|\mathcal{A}|}} \end{pmatrix} = \text{rank}\begin{pmatrix} I-\gamma_1\mathcal{T}^1_{a_1} & I-\gamma_2\mathcal{T}^2_{a_1} & 0 \\ \vdots & \vdots & \vdots \\ I-\gamma_1\mathcal{T}^1_{a_{|\mathcal{A}|}} & I-\gamma_2\mathcal{T}^2_{a_{|\mathcal{A}|}} & 0 \\ I-\gamma_1\mathcal{T}^1_{a_1} & 0 & I-\gamma_3\mathcal{T}^3_{a_1} \\ \vdots & \vdots & \vdots \\ I-\gamma_1\mathcal{T}^1_{|\mathcal{A}|} & 0 & I-\gamma_3\mathcal{T}^3_{|\mathcal{A}|} \end{pmatrix} - |\mathcal{S}| \tag{19}$$

Otherwise, the reward function from environments 1 and 2 leads to a sub-optimal policy for environment 3.

Proof Sketch. Denote A_1 as the matrix on the left side of (19) and A_2 as the matrix on the right side of (19).

Before doing a proof sketch, we first introduce Lemma 13 from [3]:

Lemma 12. (19) *holds if and only if $\forall v^1, v^2$ satisfying $(I-\gamma_1\mathcal{T}^1_a)v^1 = (I-\gamma_2\mathcal{T}^2_a)v^2, \forall a$, there exists v^3 such that $(I-\gamma_3\mathcal{T}^3_a)v^3 = (I-\gamma_1\mathcal{T}^1_a)v^1, \forall a$.*

Sufficiency: Let r^* be the true reward and r any reward compatible with experts 1 and 2. We need to show that π^3 is optimal with respect to r^*. In this case:

$$r(\cdot,a) = \lambda\log\pi^1(a|\cdot) + (I-\gamma_1 T^1_a)v^1 = \lambda\log\pi^2(a|\cdot) + (I-\gamma_2 T^2_a)v^2 \tag{20}$$

$$r^*(\cdot,a) = \lambda\log\pi^1(a|\cdot) + (I-\gamma_1 T^1_a)v^1_* = \lambda\log\pi^2(a|\cdot) + (I-\gamma_2 T^2_a)v^2_* \tag{21}$$

Subtracting the above two equations:

$$r^*(\cdot,a) - r(\cdot,a) = (I-\gamma_1 T^1_a)v^1_* = (I-\gamma_2 T^2_a)v^2_* \tag{22}$$

Also,

$$r^*(\cdot,a) = \lambda\log\pi^1(a|\cdot) + (I-\gamma_1 T^1_a)v^1_* \tag{23}$$

$$= r(\cdot,a) - (I-\gamma_1 T^1_a)v^1 + (I-\gamma_1 T^1_a)v^1_* \tag{24}$$

$$= \lambda\log\pi^3(a|\cdot) + (I-\gamma_3 T^3_a)v^3 + (I-\gamma_1 T^1_a)(v^1 - v^1_*) \tag{25}$$

Using Lemma 13, there is $\tilde{v}^3$ solving $(I-\gamma_3 T^3_a)\tilde{v}^3 = (I-\gamma_1 T^1_a)(v^1 - v^1_*)$

Setting $v^3_* = v^3 + \tilde{v}^3$, we can derive

$$r^*(\cdot,a) = \lambda\log\pi^3(a|\cdot) + (I-\gamma_3 T^3_a)v^3_* \tag{26}$$

In other words, π^3 is also optimal under r^*.

Necessity: Intuitively, there is some reward compatible with experts 1, 2 that cannot be "explained" by any value vector in environment 3, so its induced policy in 3 must be suboptimal.

D Appendix : Derivation of the reward bound - section 5

We restate Theorem 8 in Rolland *et al.* (2022) [3] here.

Theorem 9 (Approximate Identifiability)**.** *Suppose the estimated transition matrices satisfy* $\|T_a^i - \hat{T}_a^i\|_2 \leq \epsilon$ *for all* $a \in \mathcal{A}, i = 1, 2$ *and let* $\sigma_2(\hat{A})$ *be the second smallest singular value of* $\hat{A}$*. where* $\hat{A}$ *is the approximate block matrix and* A *is the true block matrix:*

$$\hat{A} = \begin{pmatrix} I - \gamma_1 \hat{\mathcal{T}}_{a_1}^1 & I - \gamma_2 \hat{\mathcal{T}}_{a_1}^2 \\ \vdots & \vdots \\ I - \gamma_1 \hat{\mathcal{T}}_{a_{|\mathcal{A}|}}^1 & I - \gamma_2 \hat{\mathcal{T}}_{a_{|\mathcal{A}|}}^2 \end{pmatrix} \qquad A = \begin{pmatrix} I - \gamma_1 \mathcal{T}_{a_1}^1 & I - \gamma_2 \mathcal{T}_{a_1}^2 \\ \vdots & \vdots \\ I - \gamma_1 \mathcal{T}_{a_{|\mathcal{A}|}}^1 & I - \gamma_2 \mathcal{T}_{a_{|\mathcal{A}|}}^2 \end{pmatrix}$$

Then, when the second smallest singular value $\sigma_2(\hat{A}) > \epsilon\sqrt{2\,|\mathcal{A}|}\max(\gamma_1, \gamma_2)$*, A has rank* $2|\mathcal{S}| - 1$*. Consequently, 7 is valid for matrix A, meaning that the also the true reward* r *is* π-identifiable up to an additive constant *despite using only the approximate* $\hat{T}_a^i$*.*

D.1 Bound on Value-function estimate

We now derive a fully explicit bound on:

$$||v - \hat{v}||$$

In this derivation, we need to create another condition, explicitly we need :

$$\|\hat{A}^+ \Delta_A\| < 1$$

so that the following Neumann-series arguments apply.

D.1.1 Step 1: Express the solution via pseudoinverses

After accounting for the one and only degree of freedom (constant shift) e.g. by setting $v^1(s_0) = 0$, and given the condition of 9, both A and $\hat{A}$ are full rank matrices so that :

$$v = A^+ b, \qquad \hat{v} = \hat{A}^+ b,$$

and hence

$$v - \hat{v} = (A^+ - \hat{A}^+)\, b \implies \|v - \hat{v}\| \leq \|A^+ - \hat{A}^+\|\, \|b\|.$$

D.1.2 Step 2: Relate A^+ to $\hat{A}^+$

Write

$$A = \hat{A} - \Delta_A = \hat{A}\,(I - \hat{A}^+ \Delta_A),$$

and note $\hat{A}^+\hat{A} = I$ on the column-space. Since $\|\hat{A}^+ \Delta_A\| < 1$, the factor $I - \hat{A}^+ \Delta_A$ is invertible. A standard pseudoinverse identity then gives

$$A^+ = \big(\hat{A}\,(I - \hat{A}^+ \Delta_A)\big)^+ = (I - \hat{A}^+ \Delta_A)^{-1}\,\hat{A}^+.$$

Thus

$$A^+ - \hat{A}^+ = \big[(I - \hat{A}^+ \Delta_A)^{-1} - I\big]\,\hat{A}^+.$$

D.1.3 Step 3: Neumann series expansion

Since $\|\hat{A}^+ \Delta_A\| < 1$, we have

$$(I - \hat{A}^+ \Delta_A)^{-1} = \sum_{k=0}^{\infty} (\hat{A}^+ \Delta_A)^k,$$

so subtracting the $k = 0$ term,

$$(I - \hat{A}^+ \Delta_A)^{-1} - I = \sum_{k=1}^{\infty} (\hat{A}^+ \Delta_A)^k.$$

Hence

$$A^+ - \hat{A}^+ = \sum_{k=1}^{\infty} (\hat{A}^+ \Delta_A)^k\,\hat{A}^+.$$

D.1.4 Step 4: Operator-norm bound

Using the triangle inequality and sub-multiplicativity,

$$\begin{aligned}\|A^+ - \hat{A}^+\| \;&\le\; \sum_{k=1}^{\infty} \big\| (\hat{A}^+ \,\Delta_A)^k \,\hat{A}^+ \big\| \;\le\; \sum_{k=1}^{\infty} \|\hat{A}^+ \,\Delta_A\|^k \,\|\hat{A}^+\| \\ &=\; \|\hat{A}^+\| \sum_{k=1}^{\infty} \big(\|\hat{A}^+\|\,\|\Delta_A\|\big)^k \;=\; \|\hat{A}^+\| \,\frac{\|\hat{A}^+\|\,\|\Delta_A\|}{1 - \|\hat{A}^+\|\,\|\Delta_A\|} \\ &=\; \frac{\|\hat{A}^+\|^2\,\|\Delta_A\|}{1 - \|\hat{A}^+\|\,\|\Delta_A\|}.\end{aligned}$$

D.1.5 Step 5: Final bound on $\|v - \hat{v}\|$

Combining with $\|v - \hat{v}\| \le \|A^+ - \hat{A}^+\|\,\|b\|$ gives the result:

$$\boxed{\|v - \hat{v}\| \;\le\; \frac{\|\hat{A}^+\|^2\,\|\Delta_A\|}{1 - \|\hat{A}^+\|\,\|\Delta_A\|}\;\|b\|.}$$

D.2 Bound on the recovered reward

We now bound the sup-norm error $\|r - \hat{r}\|_\infty$ in terms of the transition error ε and the value-solver error $\|v - \hat{v}\|$.

D.2.1 Step 1: Pointwise reward error

Recall the Bellman-recovery formula for any (s, a):

$$r(s,a) = \lambda \log \pi^1(a \mid s) \;+\; \gamma \sum_{s'} T_a^1(s')\, v^1(s') \;-\; v^1(s),$$

where we can use $v^1 \vee v^2$ arbitrary. for this derivation we will just use v^1 in order to distinguishit from $v = (v^1\ v^2)^T$ and similarly for $\hat{r}, \hat{T}^1, \hat{v}^1$. Subtracting gives

$$\Delta r(s,a) \;=\; r(s,a) - \hat{r}(s,a) = \gamma\Big[\sum_{s'} T_a^1(s')\, v^1(s') - \sum_{s'} \hat{T}_a^1(s')\, \hat{v}^1(s')\Big] - \big[v^1(s) - \hat{v}^1(s)\big].$$

D.2.2 Step 2: Bound the transition–value term

Define

$$(*) = \sum_{s'} T_a^1(s')\, v^1(s') - \sum_{s'} \hat{T}_a^1(s')\, \hat{v}^1(s') = \sum_{s'} \big[(T_a^1 - \hat{T}_a^1)(s')\, \hat{v}^1(s')\big] + \sum_{s'} T_a^1(s') \big[v^1(s') - \hat{v}^1(s')\big].$$

Using $\|T_a^1 - \hat{T}_a^1\| \le \varepsilon$ and $\|T_a^1\| \le 1$, we get

$$\big|(*)\big| \;\le\; \|T_a^1 - \hat{T}_a^1\|\,\|\hat{v}^1\| \;+\; \|T_a^1\|\,\|v^1 - \hat{v}^1\| \;\le\; \varepsilon\,\|\hat{v}^1\| \;+\; \|v^1 - \hat{v}^1\|.$$

Hence

$$\big|\Delta r(s,a)\big| \;\le\; \gamma\big[\varepsilon\,\|\hat{v}^1\| + \|v^1 - \hat{v}^1\|\big] \;+\; \|v^1 - \hat{v}^1\| = \;\gamma\varepsilon\,\|\hat{v}^1\| \;+\; (1+\gamma)\,\|v^1 - \hat{v}^1\|.$$

D.2.3 Step 3: Substitute the value-error bound

From Section 3, we have

$$\|v - \hat{v}\| \;\le\; \delta_v \;=\; \frac{\|\hat{A}^+\|^2\,\|\Delta_A\|}{1 - \|\hat{A}^+\|\,\|\Delta_A\|}\;\|b\|.$$

Furthermore we have that : $\|v^1 - \hat{v}^1\| \le \|v - \hat{v}\|$.

Since $\hat{v} = \hat{A}^+ b$ is known and $\|\hat{v}^1\| \le \|\hat{v}\| \le \|\hat{A}^+\|\|b\|$, the final bound is

$$\boxed{\|r - \hat{r}\|_\infty \;\le\; \gamma\,\varepsilon\,\|\hat{A}^+\|\|b\| \;+\; (1+\gamma)\,\frac{\|\hat{A}^+\|^2\,\|\Delta_A\|}{1 - \|\hat{A}^+\|\,\|\Delta_A\|}\;\|b\|.}$$

All quantities on the right are directly computable from $\hat{A}$, b, ε and $\|\Delta_A\|$.

D.3 Implications of the Reward-Error Bound

Our main result

$$\|r - \hat{r}\|_\infty \;\le\; \gamma\,\varepsilon\,\|\hat{A}^+\|\|b\| \;+\; (1+\gamma)\,\frac{\|\hat{A}^+\|^2\,\|\Delta_A\|}{1 - \|\hat{A}^+\|\|\Delta_A\|}\,\|b\|$$

provides several practical insights:

- **Sensitivity to transition error ε.** The term $\gamma\,\varepsilon\,\|\hat{v}\|$ grows linearly in ε: if our estimated dynamics are off by at most ε in operator norm, the *direct* impact on the recovered reward scales as $\gamma\,\|\hat{v}\|$. Thus, for small ε (high-quality transition estimates), this contribution is negligible.
- **Coupling via the linear-solver error.** The second term collects all effects of solving $\hat{A}\,\hat{v} = b$ instead of $A\,v = b$. It grows as $(1+\gamma)$ times the relative perturbation $\big(\|\hat{A}^+\|^2\|\Delta_A\|\big)/(1 - \|\hat{A}^+\|\|\Delta_A\|)$. In particular:
  - If $\hat{A}$ is well conditioned (small $\|\hat{A}^+\|$), then even moderate $\|\Delta_A\|$ causes only mild growth.
  - As $\|\Delta_A\|$ approaches the threshold $1/\|\hat{A}^+\|$, this term blows up, signaling loss of stability.
- **Trade–off between model and solver quality.** Improving transition estimates ($\varepsilon \downarrow 0$) not only reduces the first term, but also shrinks $\|\Delta_A\|$ and thus the second term. Conversely, if $\hat{A}$ is nearly singular, the reward estimate can be highly sensitive even to tiny ε.

D.4 On the Invertibility Condition $\|\hat{A}^+\,\Delta_A\| < 1$

The requirement

$$\|\hat{A}^+\,\Delta_A\| \;<\; 1$$

ensures that $\hat{A}\,(I - \hat{A}^+\Delta_A) = A$ remains of full column–rank and that the Neumann series $(I - \hat{A}^+\Delta_A)^{-1} = \sum_{k=0}^{\infty}(\hat{A}^+\Delta_A)^k$ converges.

- **Interpretation as a relative-perturbation bound.** Since $\|\hat{A}^+\| = 1/\sigma_{\min}(\hat{A})$, the condition is equivalent to $\|\Delta_A\| < \sigma_{\min}(\hat{A})$. In other words, the additive error in $\hat{A}$ must be strictly smaller than its smallest singular value.
- **Practical relevance.**
  - If $\hat{A}$ is *well conditioned* ($\sigma_{\min}$ not too small), then typical estimation errors $\|\Delta_A\|$ will satisfy the condition almost automatically.
  - If $\hat{A}$ is *ill conditioned* (large $\|\hat{A}^+\|$), then even small perturbations may violate $\|\hat{A}^+\Delta_A\| < 1$, leading to rank loss or large solver error.
- **Strength of the assumption.** It is *standard* in matrix-perturbation theory and essentially *necessary* to guarantee continuity of the pseudoinverse and bounded error. In well-posed IRL setups, where the feature matrices (and thus A) are designed to be full-rank and not near-singular, the assumption holds readily. In borderline or under-determined problems, one must regularize or otherwise enforce a margin $\sigma_{\min} \gg \|\Delta_A\|$ to restore stability.