Identifiability and Generalizability from Multiple Experts in Inverse Reinforcement Learning

Paul Rolland
LIONS, EPFL
Lausanne, Switzerland
paul.rolland@epfl.ch

Luca Viano
LIONS, EPFL
Lausanne, Switzerland
luca.viano@epfl.ch

Norman Schürhoff
SFI, UNIL
Lausanne, Switzerland
norman.schuerhoff@unil.ch

Boris Nikolov
SFI, UNIL
Lausanne, Switzerland
boris.nikolov@unil.ch

Volkan Cevher
LIONS, EPFL
Lausanne, Switzerland
volkan.cevher@epfl.ch

Abstract

While Reinforcement Learning (RL) aims to train an agent from a reward function in a given environment, Inverse Reinforcement Learning (IRL) seeks to recover the reward function from observing an expert's behavior. It is well known that, in general, various reward functions can lead to the same optimal policy, and hence, IRL is ill-defined. However, [1] showed that, if we observe two or more experts with different discount factors or acting in different environments, the reward function can under certain conditions be identified up to a constant. This work starts by showing an equivalent identifiability statement from multiple experts in tabular MDPs based on a rank condition, which is easily verifiable and is shown to be also necessary. We then extend our result to various different scenarios, i.e., we characterize reward identifiability in the case where the reward function can be represented as a linear combination of given features, making it more interpretable, or when we have access to approximate transition matrices. Even when the reward is not identifiable, we provide conditions characterizing when data on multiple experts in a given environment allows to generalize and train an optimal agent in a new environment. Our theoretical results on reward identifiability and generalizability are validated in various numerical experiments.

1 Introduction

Engineering a reward function in Reinforcement Learning can be troublesome in certain scenarios like driving [2], robotics [3], and economics/finance [4]. In economics and finance, the reward or objective/utility function of the agent are of fundamental importance but are not known a priori [5–8]. In such cases, it may be easier to get demonstrations from an expert policy. Therefore, multiple algorithms have been developed to learn from demonstrations, e.g., in inverse reinforcement learning (IRL) and imitation learning (IL).

In IRL, the goal is to recover the reward function maximized by the agent, while in IL the expert demonstrations are used solely to learn a nearly optimal policy. In economics/finance, inference on the reward function is the focus of a large literature on estimation, testing, and policy analysis of structural models [9–11]. However, the reward function is often highly parameterized and represented by a low-dimensional set of parameters, or the literature focuses on estimating reduced-form causal relationships but not the true reward function [12, 13]. The attractiveness of IRL relies on the fact that the reward function is the most "succinct" representation of a task [14]. Indeed, identifying the

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

reward function for each state-action pair allows generalizing the task to different transition dynamics and environments, which is not possible when using IL or highly parameterized structural models.

However, the IRL problem is unfortunately ill-posed since there always exist infinitely many reward functions for which the observed expert policy is optimal [15, 16]. The problem is known as reward shaping, and it is intuitively explained with the fact that, in the long term, the optimal policy is not affected by inflating the reward in the current period and decreasing the one in the next. This difficulty originated a long debate on advantages and disadvantages of IL and IRL [17–20].

When multiple experts are available, differing in the transition matrices of the environments they each act in, and/or their discount factors, IRL can in certain cases infer the true reward function, up to a constant [21–23, 1]. Inspired by [1], we derive an equivalent necessary and sufficient condition on the expert environments, which is easily verifiable, ensuring that the true reward can be identified up to a constant shift. When this identifiability condition holds, the state-action dependent rewards can be recovered from expert demonstrations. We then derive identifiability results in various alternative scenarios, e.g., when we only have access to approximate transition matrices and, alternatively, when the reward function is known to be a linear combination of given features [24, 25].

However, full reward identifiability remains a strong requirement, and we provide a negative result of non-identifiability from any number of experts, in the presence of exogenous variables in the MDP. Nonetheless, even when the identifiability condition does not hold, the recovered reward function could still be used to train an optimal expert for a different environment. To this end, we characterize situations where observing multiple experts in given environments allows to train an optimal agent in a new environment.

2 Related work

Since its introduction in [15, 16], the IRL problem has been known to be ill-posed, since the observed expert policy can be optimal with respect to various reward functions. The set of reward transformations that preserve policy optimality are studied in [26, 16, 1, 27, 28]. [29] studied the unidentifiability related to suboptimal experts.

In this paper, we assume access to the optimal entropy regularized policies of multiple experts. Significant progress has been made to construct heuristics that select a single reward function from the set of IRL solutions (often called the feasible set), such as feature-based matching [30], maximum margin IRL [31], maximum causal entropy IRL [32, 33], maximum relative entropy IRL [34], Bayesian IRL [35–37], first-order optimality conditions [38, 39] or second-order optimality conditions [40, 41]. Popular IL algorithms implicitly select a feasible reward function via a convex reward regularizer [19, 42, 43] or using preference/ranking based algorithms [44, 45]. However, none of these approaches guarantee the identification of the true reward function.

The problem of identifiability in IRL has been investigated first in [21, 22] that study a setting where the learner can actively select optimal experts in multiple environments. The main result in [21, 22] is that interactively querying environments outputs a reward within a constant shift from the true one. The multiple experts setting has also been studied in [46] but in the context of value alignment verification where the aim is not to recover the reward function but rather verify that the value function of the agent is close to a target value. IRL from multiple MDPs also appears in [23] where the authors consider the problem of learning a reward function compatible with a dataset of demonstrations collected by multiple experts. In addition, [47] study structural conditions on the MDP for reward identification in the finite horizon setting and [48] study identifiability in linearly solvable MDPs.

Our work is inspired by [1]. Our first identifiability result provides an equivalent statement as their *value distinguishability* condition, but can be easily checked in practice, and allows to derive other identifiability results in alternative scenarios. Finally, the motivation for IRL is often predicting the expert behavior under new transitions dynamics [49, 50, 20]. We show that for this goal, it is not necessary to identify the exact reward, hence we give a condition on the observed experts' environments and the test environment under which an optimal expert can be trained in the test environment. This perspective has also been taken in [51]. However, this work requires stronger assumptions on the transfer environment that we avoid in this paper, only requiring access to multiple experts. Moreover, our work contributes to AI safety [52–54] alleviating the *reward hacking* and *side effects* problems [53]. Indeed, by restricting the reward to linear combinations of a set of chosen

features, we can provably recover an interpretable reward function inducing the optimal behavior, which is particularly desirable in medical applications [55, 56].

An important consideration for IRL comes from [57] that formalizes the fact that there exist tasks that can not be induced by optimizing a reward function. In this work and in IRL in general, we bypass this difficulty assuming that the expert is optimizing a reward function.

2.1 Related works in the economics literature

The economics/finance literature differentiates between axiomatic and revealed preference theory. In axiomatic preference theory, the reward function is posited or derived from basic axioms. In empirical and experimental work, however, simple reward function specifications are often rejected and agents have been shown to exhibit behavioral biases and/or non-standard preferences.

Differently, our work relates to the literature on revealed preference. Revealed preference theory, initiated by [58, 59], provides an approach to analyze actions (e.g., consumer's demand or investors' trading) by assuming that observed choices provide information about the underlying preferences, or reward function. Revealed preference theory is, hence, similar in spirit to IRL. But IRL has not widely been used in revealed preference analysis. We refer to [60, 61] for excellent reviews of recent advances in revealed preference theory. The goal of revealed preference theory is to recover the agents' preferences. This task is important because knowledge of the reward function is required to conduct counterfactual policy analysis. Notice that for this task, knowing only the policy function is insufficient. In financial applications, for instance, the impact of a Tobin tax can be assessed only knowing investors' preferences for trading (see, e.g. [62]).

3 Preliminaries

A typical RL environment is characterised by a Markov Decision Process $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, T, \gamma, r, P_0\}$, where $\mathcal{S}, \mathcal{A}$ are the sets of states and actions respectively, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability, i.e., $T(s'|s, a)$ denotes the probability of arriving in state s' when taking action a in state s. $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the reward function, γ the discount factor and P_0 is the initial state distribution. At each time step t, an agent observes the current state $s_t \in \mathcal{S}$ and takes an action $a_t \sim \pi(\cdot|s_t)$ where π is the agent's policy which determines a distribution over all actions in $\mathcal{A}$ at every state. The agent gets a reward $r_t = r(s_t, a_t)$ and transitions to a new state s_{t+1} sampled according to the transition probability T.

An agent acting optimally in $\mathcal{M}$ seeks to maximize its cumulative sum of rewards. In addition, we assume that the agent seeks to diversify its possible actions, and hence that it maximizes the following entropy regularized sum of discounted rewards:

$$V_\lambda^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{\infty} (\gamma^t (r(s_t, a_t) + \lambda \mathcal{H}(\pi(\cdot|s_t)))) \right], \tag{1}$$

where $\mathbb{E}_s^\pi$ denotes the expectation over trajectories $\{(s_t, a_t\}_{t\geq 0}$ starting from state $s_0 = s$ and following policy π and $\mathcal{H}(\pi) = -\sum_{a\in\mathcal{A}} \pi(a) \log \pi(a)$ is the entropy of π. The function V_λ^π is called the (entropy regularized) value function of π.

In Inverse RL, the reward function r is unknown, but we observe an agent acting optimally with respect to some reward function, and we wish to recover the reward function that the agent optimizes. We now recall some results from [1].

Theorem 1. *For a fixed policy $\pi(a|s) > 0$, discount factor $\gamma \in [0, 1)$, and an arbitrary choice of function $v : \mathcal{S} \to \mathbb{R}$, there is a unique corresponding reward function*

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

such that the MDP with reward r yields an entropy-regularized optimal policy $\pi_\lambda^* = \pi$ and $V_\lambda^\pi = v$.

By observing a single expert, it is hence possible to design a reward that yields any arbitrary value function, and there are hence $|\mathcal{S}|$ degrees of freedom remaining in the recovered reward function. An idea explored in [1] is to assume that we observe two experts in two different MDPs with different

transition dynamics and discount rates, but acting optimally with respect to the same reward function. The authors show that the reward can be identified up to a constant from observing the expert policies provided that the MDPs of the experts satisfy the following *value-distinguishing* assumption.

Definition 2. *Consider a pair of Markov decision problems on the same state and action spaces, but with respective discount rates γ_1, γ_2 and transition probabilities T^1, T^2. We say that this pair is value-distinguishing if, for any function $v^1, v^2 : \mathcal{S} \to \mathbb{R}$, the statement*

$$v^1(s) - \gamma_1 \sum_{s' \in \mathcal{S}} T^1(s'|s,a) v^1(s') = v^2(s) - \gamma_2 \sum_{s' \in \mathcal{S}} T^2(s'|s,a) v^2(s') \; for \; all \; a \in \mathcal{A}, s \in \mathcal{S} \quad (2)$$

implies at least one of v^1 and v^2 is a constant function.

The way this assumption is stated makes it difficult to verify in practice, and the authors of [1] do not attempt to verify it in their experiments.

4 Reward identification and generalization

In this section, we present our main theoretical results on reward identifiability and generalizability. In the first part, we show an equivalent condition to Definition 2 for reward identification from two experts (Theorem 3). The simplicity of our condition makes it easily verifiable and extendable to various scenarios, in particular to the cases where we observe more than two experts (Corollary 5), when the class of rewards is linearly parameterized with a set of given features (Theorem 7), or when we have access to approximated transition matrices (Theorem 8). We also provide a negative result on reward non-identifiability in MDPs with exogenous variables, which are common in many real world scenarios. In the second part, we analyse reward generalizability. Here, we provide a condition guaranteeing that a reward compatible with two experts leads to an optimal policy in a third environment (Theorem 11). The proofs of the results are all postponed to Appendix A.

4.1 Reward identifiability

Consider two Markov decision problems on the same set of states and actions $\mathcal{S}$ and $\mathcal{A}$ respectively, but with different transition dynamics T^1, T^2 and discount factors γ_1, γ_2. Let $r \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ be the reward function common to the two environments, and let $v^1, v^2 \in \mathbb{R}^{|\mathcal{S}|}$ be the entropy regularized values functions associated expert policies π^1 and π^2 in each environment respectively. According to Theorem 1, we have that $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} r(s,a) &= \lambda \log \pi^1(a|s) - \gamma_1 \sum_{s' \in \mathcal{S}} T^1(s'|s,a) v^1(s') + v^1(s) \\ &= \lambda \log \pi^2(a|s) - \gamma_2 \sum_{s' \in \mathcal{S}} T^2(s'|s,a) v^2(s') + v^2(s). \end{aligned}$$

We hence deduce that $\forall a \in \mathcal{A}$,

$$\begin{pmatrix} I - \gamma_1 T^1_a & -(I - \gamma_2 T^2_a) \end{pmatrix} \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \lambda \log \pi^2(\cdot|a) - \lambda \log \pi^1(\cdot|a), \quad (3)$$

where $\forall a \in \mathcal{A}$, $T^i_a \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$ is the transition matrix for action a and expert $i = 1, 2$, i.e., $T^i_a(s,s') = T^i(s'|s,a)$. By including all available actions to the experts, we can write

$$\begin{pmatrix} I - \gamma_1 T^1_{a_1} & -(I - \gamma_2 T^2_{a_1}) \\ \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & -(I - \gamma_2 T^2_{a_{|\mathcal{A}|}}) \end{pmatrix} \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \begin{pmatrix} \lambda \log \pi^2(\cdot|a_1) - \lambda \log \pi^1(\cdot|a_1) \\ \vdots \\ \lambda \log \pi^2(\cdot|a_{|\mathcal{A}|}) - \lambda \log \pi^1(\cdot|a_{|\mathcal{A}|}) \end{pmatrix}. \quad (4)$$

In order to identify a unique reward function, we need to identify a unique associated value function. We hence want the linear system (4) to yield a unique solution, i.e., the $|\mathcal{A}||\mathcal{S}| \times 2|\mathcal{S}|$ matrix on the left hand side to be full rank, i.e., to have rank $2|\mathcal{S}|$. However, it is well known that, for any MDP, adding a constant to the reward would not change the associated optimal policy. Hence, there is an intrinsic

degree of freedom in reward identifiability which is impossible to get rid of from only observing expert policies. In order to identify the reward up to a constant, we need this degree of freedom to be the only one in the linear system (4), i.e., the associated matrix to have rank $2|\mathcal{S}|-1$. This result is summarized in the following theorem, and its complete proof can be found in Appendix A.1.

Theorem 3. *Consider two Markov decision problems on the same set of states and actions, but with different transition dynamics T_1, T_2 and discount factors γ_1, γ_2. Suppose that we observe two experts acting each in one of these environments, optimally with respect to the same reward function, in the sense that their policies maximize the entropy regularized reward in their respective environments. Then, the reward function can be recovered up to the addition of a constant if and only if*

$$\textit{rank}\begin{pmatrix} I-\gamma_1 T^1_{a_1} & I-\gamma_2 T^2_{a_1} \\ \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & I-\gamma_2 T^2_{a_{|\mathcal{A}|}} \end{pmatrix} = 2|\mathcal{S}|-1. \tag{5}$$

This condition turns out to be equivalent to Definition 2, as shown at the end of Appendix A.1, but is stated in a way that is easier to check in practice and allows us to further characterize identifiability in various scenarios. First of all, this result naturally extends to the case where we observe any number of experts. We provide hereafter the result in the case of three experts.

Corollary 4. *Consider three Markov decision problems on the same set of states and actions, but with different transition dynamics T_1, T_2, T_3 and discount factors $\gamma_1, \gamma_2, \gamma_3$. Suppose that we observe three experts acting each in one of these environments, optimally with respect to the same reward function. Then, the reward function can be recovered up to the addition of a constant if and only if*

$$\textit{rank}\begin{pmatrix} I-\gamma_1 T^1_{a_1} & I-\gamma_2 T^2_{a_1} & \boldsymbol{0} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & I-\gamma_2 T^2_{a_{|\mathcal{A}|}} & \boldsymbol{0} \\ I-\gamma_1 T^1_{a_1} & \boldsymbol{0} & I-\gamma_3 T^3_{a_1} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & \boldsymbol{0} & I-\gamma_3 T^3_{a_{|\mathcal{A}|}} \end{pmatrix} = 3|\mathcal{S}|-1. \tag{6}$$

An interesting scenario is the one where the two experts act in the same environment, and only the discount rate is varied.

Corollary 5. *Consider two Markov decision problems on the same set of states and actions, with the same transition matrix T and reward function but different discount factors $\gamma_1 \neq \gamma_2$. Then, the reward function is identifiable up to a constant by observing two experts in $(T, \gamma_1), (T, \gamma_2)$ iff*

$$\textit{rank}\begin{pmatrix} T_{a_1} - T_{a_2} \\ \vdots \\ T_{a_1} - T_{a_{|\mathcal{A}|}} \end{pmatrix} = |\mathcal{S}|-1. \tag{7}$$

Remark 1. *Interestingly, condition* (7) *is equivalent to the condition for identification of a action-independent reward from a single expert, assuming such a reward exists ([1], Corollary 3).*

Next, we provide a negative result concerning MDPs with exogenous variables, i.e., a variable whose dynamics are independent of the agent's action. This MDP class is common in economics/finance and has been studied in many real world scenarios including inventory control problems [63], variable weather conditions and customer demands [64], wildfire management [65], and stock market fluctuations [66]. We also provide examples involving such variables in the experimental section.

Corollary 6. *Suppose that the state space is constructed as a set of variables each taking a finite number of values, i.e., $\mathcal{S} = \{s \in \mathbb{R}^d : s_i \in \mathcal{S}_i\}$. The transition matrices for each action a can be defined by specifying the evolution of each state variable s_i^{t+1} depending on (s^t, a). Suppose that there exists a state variable whose evolution only depends on its previous value, but neither on the other state variables nor the action taken: such a variable is called an **exogenous** variable. Note that this variable can still affect the evolution of all other variables, and its evolution can vary across the environment of the observed experts. Then, the reward function is **not** identifiable (even up to a constant) using any number of experts.*

Such a negative result motivates the search for milder requirements than arbitrary reward identification, which is too hard of a goal to achieve in certain scenarios.

A possible way to improve reward identifiability is to restrict the class of possible rewards, e.g., by constraining it to be a linear combination of a set of chosen features. This is known as Feature matching IRL [49, 67–70]. The smaller the set of features, the easier to identify the reward, as described in the following theorem. This method also allows to recover a more interpretable reward function, since the recovered parameters are associated with specific features.

Theorem 7. *Suppose that we restrict the class of possible reward functions to the one parameterized as $r_w(s,a) = w^T f_{s,a}\ \forall a \in \mathcal{A}, s \in \mathcal{S}$ where $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a given feature function, and $w \in \mathbb{R}^d$ denotes the reward parameters. Suppose that the d chosen features are linearly independent, i.e., that $f_{s,a}^T v = 0\ \forall s,a \Rightarrow v = 0$. Then, if $\mathbf{1} \in Im \begin{pmatrix} f_{a_1} \\ \vdots \\ f_{a_{|\mathcal{A}|}} \end{pmatrix}$, the reward is identifiable up to constant by observing experts acting in $(T^1, \gamma_1), (T^2, \gamma_2)$ if and only if*

$$rank \begin{pmatrix} I - \gamma_1 T^1_{a_1} & I - \gamma_2 T^2_{a_1} & \boldsymbol{0} \\ \vdots & \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & I - \gamma_2 T^2_{a_{|\mathcal{A}|}} & \boldsymbol{0} \\ I - \gamma_1 T^1_{a_1} & \boldsymbol{0} & f_{a_1} \\ \vdots & \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & \boldsymbol{0} & f_{a_{|\mathcal{A}|}} \end{pmatrix} = 2|\mathcal{S}| + d - 1. \tag{8}$$

where $f_a = (f_{s_1,a} \dots f_{s_{|\mathcal{S}|},a})^T \in \mathbb{R}^{|\mathcal{S}| \times d}$. On the other hand, if $\mathbf{1} \notin Im \begin{pmatrix} f_{a_1} \\ \vdots \\ f_{a_{|\mathcal{A}|}} \end{pmatrix}$, then the reward can be exactly recovered provided that the rank of the matrix on the left hand side of equation (8)*, which augments equation* (5) *by the features being matched, is $2|\mathcal{S}| + d$.*

Finally, it usually happens that the exact transition matrices $\{T_a\}_{a \in \mathcal{A}}$ are not known exactly and must be estimated, e.g., from samples. Verifying condition (5) on the approximated matrices may be misleading since the rank is very sensitive to small perturbations. Hence, we provide hereafter an identifiability condition in the case where we only have access to approximated transition matrices.

Theorem 8. *Suppose that we approximate the transition matrices $\{T^i_a\}_{a \in \mathcal{A}}$ as $\{\hat{T}^i_a\}_{a \in \mathcal{A}}$ such that $\|T^i_a - \hat{T}^i_a\|_2 \leq \epsilon\ \forall a \in \mathcal{A}$, $i = 1, 2$. Suppose that we verify condition* (5) *using the approximated matrices, i.e., we compute the second smallest eigenvalue σ of the following matrix:*

$$\begin{pmatrix} I - \gamma_1 \hat{T}^1_{a_1} & I - \gamma_2 \hat{T}^2_{a_1} \\ \vdots & \vdots \\ I - \gamma_1 \hat{T}^1_{a_{|\mathcal{A}|}} & I - \gamma_2 \hat{T}^2_{a_{|\mathcal{A}|}} \end{pmatrix}. \tag{9}$$

Then, condition (5) *on the true transition matrices $\{T_a\}_{a \in \mathcal{A}}$ holds provided that*

$$\sigma > \epsilon \sqrt{2|\mathcal{A}|} \max(\gamma_1, \gamma_2). \tag{10}$$

Remark 2. *The matrix estimator $\hat{T}_a$ can be obtained from samples. For example, [71][Lemma 5] shows that a high probability bound on the max norm $\|T_a - \hat{T}_a\|_{\max} \leq \epsilon$ requires $\mathcal{O}(\epsilon^{-4})$ samples from a generative model [72]. This would imply the following bound on the spectral norm: $\|T_a - \hat{T}_a\|_2 \leq |\mathcal{S}| \|T_a - \hat{T}_a\|_{\max} \leq |\mathcal{S}|\epsilon$. However, the dependence on ϵ can be improved as we show next applying the matrix Bernstein bound [73, 74].*

Theorem 9. *Let $\hat{T}_a$ be the empirical estimator for T_a. Then with probability greater than $1 - \delta$,*

$$\|T_a - \hat{T}_a\|_2 \leq |\mathcal{S}| \sqrt{\frac{\log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}}{2N}} + \frac{2(|\mathcal{S}| + 1) \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}}{3N} \quad \forall a \in \mathcal{A}. \tag{11}$$

Therefore, we can obtain $\|T_a - \hat{T}_a\|_2 \leq \epsilon$ with $\mathcal{O}(\epsilon^{-2})$ samples.

4.2 Generalization to unknown environments

We now focus on reward generalizability, i.e., the ability to recover a reward function that would allow us to train an optimal policy in a new environment. Suppose that we recover a reward function that is compatible with two experts acting in two MDPs $\mathcal{M}_1, \mathcal{M}_2$, and that we use this reward to train an expert in a third environment $\mathcal{M}_3$, assuming all environments share the same true reward function but possibly different transition dynamics and discount factors. What condition guarantees that the trained expert will be optimal in $\mathcal{M}_3$?

This generalization requirement is milder than full reward identification. Indeed, being able to identify the reward (even up to a constant) naturally allows to train an optimal policy in any other environment sharing the same reward. However, even in the presence of non-trivial degrees of freedom, it may be the case that any recovered reward suffices to train an optimal policy in a given other environment.

Intuitively, the third training environment should not vary too much from the observed environments $\mathcal{M}_1, \mathcal{M}_2$. More precisely, if observing a third expert in environment 3 does not provide any further identification of the reward than with environments 1 and 2, then any reward compatible with environments 1 and 2 leads to an optimal policy in environment 3. The condition is made precise in the following theorem.

Definition 10. *Consider three Markov decision problems on the same set of states and actions, but with different transition matrices T_1, T_2, 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 $(T^1, \gamma_1), (T^2, \gamma_2)$* ***generalize to*** *$(T^3, \gamma_3)$ 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.*

Theorem 11. *$(T^1, \gamma_1), (T^2, \gamma_2)$ generalize to (T^3, γ_3) if and only if*

$$\textit{rank}\begin{pmatrix} I-\gamma_1 T^1_{a_1} & I-\gamma_2 T^2_{a_1} \\ \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & I-\gamma_2 T^2_{a_{|\mathcal{A}|}} \end{pmatrix} = \textit{rank}\begin{pmatrix} I-\gamma_1 T^1_{a_1} & I-\gamma_2 T^2_{a_1} & \boldsymbol{0} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & I-\gamma_2 T^2_{a_{|\mathcal{A}|}} & \boldsymbol{0} \\ I-\gamma_1 T^1_{a_1} & \boldsymbol{0} & I-\gamma_3 T^3_{a_1} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & \boldsymbol{0} & I-\gamma_3 T^3_{a_{|\mathcal{A}|}} \end{pmatrix} - |\mathcal{S}|. \tag{12}$$

This condition is also necessary, in the sense that, if it does not hold, then there exists a reward function compatible with experts 1 and 2 but which leads to a sub-optimal policy in environment 3.

One interesting question is whether observing two experts in the same environment with different discount factors allows to generalize to any other expert with arbitrary discount factor. It turns out to be the case under some commutativity constraint on the transition matrices.

Corollary 12. *Consider a single environment with transitions T. Suppose that there exists an action $a_0 \in \mathcal{A}$ such that T_{a_0} commutes with T_a for all $a \in \mathcal{A}$. Then for any $0 < \gamma_1, \gamma_2, \gamma_3 < 1$ with $\gamma_1 \neq \gamma_2$, $(T, \gamma_1), (T, \gamma_2)$ generalize to (T, γ_3).*

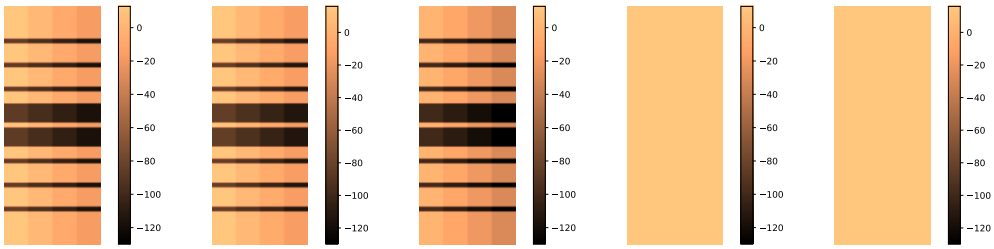
Remark 3. *The commutativity condition cannot simply be removed. Indeed, we provide in Appendix A.9 an example with two actions with non-commutative transition matrices for which condition (12) is not satisfied.*

5 Experiments

We now present empirical validations of our claims[1]. In particular, we verify the identifiability requirement given by Theorem 3 in the context of randomly generated transition matrices and different gridworlds with uniform additive noise in the dynamics.

In addition, we study a `Windy-Gridworld` and a financial model that we term `Strebulaev-Whited` both involving exogenous variables in their state spaces. In agreement with Corollary 6, the reward

[1]Code available at the following link `https://github.com/lviano/Identifiability_IRL`



(a) r_α (b) r_γ (c) r_{true} (d) $r_\alpha - r_{\text{true}}$ (e) $r_\gamma - r_{\text{true}}$

Figure 1: Comparison between true and recovered reward in `Gridworld` with an action dependent reward, $|\mathcal{S}| = 100$. It can be noticed that the reward function r_γ recovered changing discount factors is within a constant shift from the true reward (subplots (b),(e)). The same conclusion holds for r_α recovered from different α(see subplots (a),(d)).

function is not identifiable in these environments, highlighting the necessity of imposing milder requirements than full reward recovery. For example, in `Windy-Gridworld`, we show that by observing multiple experts acting in environments with different wind distributions, we can generalize, i.e., train an optimal expert in environments with arbitrary other wind distribution, in accordance with Theorem 11. On the other hand, in `Strebulaev-Whited`, given the additional information that the reward function can be represented as a linear combination of some known features, we can identify the reward, validating the condition of Theorem 7. The algorithms are described in Appendix B.

5.1 Identifiability experiments

Experiments on `Random-Matrices` The first experiment involves randomly generated transition matrices and reward function with $|\mathcal{S}| = 18, |\mathcal{A}| = 5$. This setting matches the numerical evidence in [1]. Their algorithm recovers the reward function but the connection with their theoretical contribution is not highlighted. On the contrary, we have no theory practice mismatch, since we verify exactly the condition in Theorem 3. In particular, for the 100 random seed we tried the rank of the matrix A is $2|\mathcal{S}| - 1 = 35$, then invoking Theorem 3 we can conclude that the reward function is identifiable up to a constant shift. We provide a visual example of the recovered reward in Figure 4 in Appendix C.

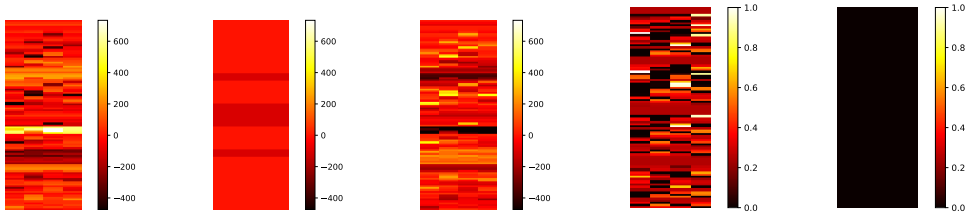
Experiments on `Gridworld` As a second example of identifiability, we consider `Gridworld`, where the state space is a squared grid with 100 states while the action set is given by $\mathcal{A} = \{\text{up}, \text{down}, \text{left}, \text{right}\}$ with dynamics given by $T_\alpha(s'|s,a) = (1-\alpha)T_{\text{det}}(s'|s,a) + \alpha U(s'|s,a)$ where $T_{\text{det}}(s'|s,a)$ represents deterministic transition dynamics where for example the action right leads to the state on the right with probability 1. If an action would lead outside the grid, then the agent stays in the current state with probability 1. The dynamics $U(s'|s,a)$ are instead uniform over the states that are first adjacent to the current state. In other words, $U(\cdot|s,a) = \text{Unif}(\mathcal{N}(s)) \quad \forall a \in \mathcal{A}$ where $\mathcal{N}(s)$ denotes the set of first neighbors of the state s.

We generate two different environments changing the value of α, choosing $\alpha^1 = 0.4$ and $\alpha^2 = 0.2$. We notice that, even using the same discount factor $\gamma = 0.9$, the condition of Theorem 3 holds. When α is kept fixed, we also notice that the condition of Corollary 5 holds, and hence the reward can be recovered by just varying the discount factor γ of the experts. We numerically verify that the reward can indeed be identified up to a constant shift in these two settings (see Figure 1).

5.2 Generalizability experiments

In this section, we present cases where identifiability is not possible due to the presence of exogenous variables. However, we notice that the generalizability condition in Theorem 11 is often satisfied, even for a test environment with parameters rather different than the environments of the observed experts. We start briefly describing the environments to later comment on the results.

Experiments on `WindyGridworld` The `WindyGridworld` environment augments the `Gridworld` state representation by including a wind direction. The wind impacts the position transitions by making the agent move one step in the direction of the wind in addition to the action taken. The wind

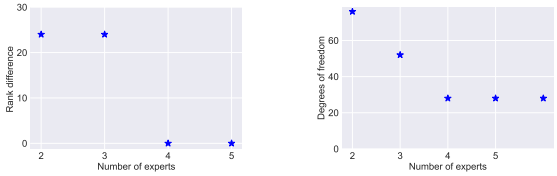


(a) r (b) r_{true} (c) $r - r_{\text{true}}$ (d) $\pi_{r_{\text{true}}}^{T_{\text{test}}}$ (e) $\pi_{r_{\text{true}}}^{T_{\text{test}}} - \pi_{r}^{T_{\text{test}}}$

Figure 3: Comparison between true and recovered reward (r and r_{true}) from 4 experts in `WindyGridworld` with $|\mathcal{S}| = 400$. We notice that the reward function is not identified (see (a), (b), (c)). However, when we use the recovered reward in subplot (a) to train an optimal policy under unseen dynamics we recover the optimal policy under the true reward in subplot (b). The subplot (d) shows the policy $\pi_{r_{\text{true}}}^{T_{\text{test}}}$ recovered from the true reward in a new environment T_{test} and (e) shows the difference between the policy recovered from r_{true} and from the recovered reward denoted as $\pi_{r}^{T_{\text{test}}}$.

directions at step t, w_t are sampled i.i.d. from the distribution P_{wind}, and is hence an exogenous variable. While the reward is not identifiable whatever the number of experts, we can generalize to a new environment with an arbitrary wind distribution by observing enough experts in environments with different wind distributions.

In Figure 2b, we see that we can obtain better identifiability (although never full identifiability) when increasing the number of experts. Once we have observed 4 experts, we do not get further identifiability by observing more experts, hence leading to generalizability as shown in Figure 2a and Figure 3. We conjecture that this number of experts is linked to the number of values that the exogenous variable, i.e. the wind direction, can take.



(a) Generalizability (b) Identifiability

Figure 2: Figure 2a shows the difference between right and left term of Theorem 11. Figure 2b shows the difference between columns and rank of the matrix in Theorem 3.We have identifiability or generalizability respectively when those values are 0.

Furthermore, although the actions in Gridworld do not exactly commute (because of the boundary), observing two experts in the same environment with different discount factors enables generalizing to a different discount factor (see Figure 6 in Appendix C). The condition of Corollary 12 is hence sufficient but not necessary.

Experiments on `Strebulaev-Whited` The `Strebulaev-Whited` environment is the neoclassical investment model in which a firm has a Cobb-Douglas production function with decreasing returns to scale, as in [75]. The goal of the agent is to maximize profits discounted at rate $0 < \gamma < 1$. The state of the agent is defined by the capital level $k \geq 0$ and an exogenously given persistent stochastic productivity shock z. We can summarize the state by $s = (k, z)$. The next state $s' = (k', z')$ is determined separately for k' and z'. We have that $k' = (1 - \delta)k + ak$, where δ is the depreciation rate of physical capital and a is today's rate of investment which is the action in the model. The variable z' evolves according to $\ln z' = \rho \ln z + \epsilon$ where $\epsilon \sim N(0, \sigma_\epsilon)$.

The continuous variables k and z are discretized according to the scheme proposed in [76]. Hence, we obtain a discrete process with K^2 possible values for the state variable $s = (k, z)$ (so $|\mathcal{S}| = K^2$) and K values for the action a. In the experiments in Figure 7 in Appendix C , we choose $K = 20$ and consider two environments with different values of σ_ϵ set to 0.02 and 0.04, respectively. We observe that the rank of the identifiability matrix is 552. Since $552 < 2|\mathcal{S}| - 1 = 799$, the reward function is not identifiable up to a constant as expected in MDPs with exogenous states. Nonetheless, when we consider a third environment with $\sigma_\epsilon = 0.6$, the generalizability condition in Theorem 11 is satisfied. Hence, the expert behavior can be predicted in the third environment (see Figure 7e in Appendix C).

5.3 Identifiability experiments with a restricted reward class

The final result presents a numerical validation of Theorem 7 in the environment `Strebulaev-Whited` with exogenous state variable. In this model, the true reward function can be expressed as a linear combination of the three features given by $f_{s,a} = [z((1-\delta)k+ak)^\theta, (1-\delta)k, ak]^T$, where $s = (k, z)$ and the parameter $\rho \in (0, 1)$ captures the curvature of the production function. We set $\rho = 0.55$. The first feature corresponds to the firm's output or sales which is available from the firm's income statement, the second feature is the firm's current capital stock net of depreciation which is available from the balance sheet, and the third feature is the level of investment that determines the future level of capital stock. The true reward function can be written as $r(s, a) = w^T f_{s,a}$ with $w = [1, 1, -1]^T$. It can be interpreted as follows: the agent's reward of investment is an increase in output/sales, $w_1 = 1 > 0$, while the cost of capital is 1 and, hence, investment is costly, $w_3 = -1 < 0$. At the same time, the capital stock is valuable and can be liquidated at a price of $w_2 = 1 > 0$.

Knowing these features, we can verify that the rank of the matrix in Equation (8), is 803 which is equal to $2|S| + d$ in this environment ($|S| = 400$ and $d = 3$). Invoking Theorem 7, we can conclude that the reward function is identifiable exactly, which is verified numerically in Figure 8 in Appendix C. Expressing the reward in terms of features hence helps identifiability and interpretability.

6 Conclusion

In this paper, we analyze conditions that guarantee identifiability of the reward function (up to an additive constant) from multiple observed experts maximizing the same reward and facing different transition dynamics. This allows us to train optimal policies in any other environment sharing the same reward with the environments of the observed experts. On the other hand, in order to generalize to unknown environments, such strong reward identification is not required, and we provide a milder necessary and sufficient condition for generalizability. We also provide identifiability results in a variety of settings, i.e., linearly parameterized reward, approximated transition matrices, observation of any number of experts, as well as a non-identifiability result in the presence of exogenous variables. In the following, we list the main limitations of our work that will be the subject of future studies.

Observing experts in different environments. We saw that observing a single expert in one environment cannot lead to reward identification in our setting. We hence need to observe at least two experts acting in different enough environments. To motivate this assumption, note that varying environments are ubiquitous in RL, in particular in Robust RL which deals with the training of experts that perform well in different environments, where the transition dynamics can vary to some extent. It is hence rather common to consider that the transition dynamics of a given environment can change. This was studied, e.g., in [77–80], where the authors considered different Mujoco environments with varying friction coefficients, or object masses, which influence the dynamics. Also, instead of observing different experts in different environments, we could imagine that we observe a single expert in a single environment that varies over time (but with fixed reward), and that the expert adapts to these changes. Such observations would provide us optimal actions in environments with different transition dynamics, and thus our results would apply. This is of particular interest in economics/finance where the environment is in constant evolution.

Assuming entropy regularized experts. When observing real world data, we have to face the fact that humans do not follow this idealized mathematical model. However, it turns out that our results still hold for the more general class of regularized MDPs [81] where we replace the entropy with any strongly convex function (see Appendix D). Whether the flexibility in the choice of the strongly convex regularizer allows to better capture real-world behaviors is an open question.

Acknowledgements

This work has received financial support from the Enterprise for Society Center (E4S) and SNF project 100018_192584. This work was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_205011. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725594 - time-data).

References

[1] Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] W. Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis)design for autonomous driving, 2021.

[3] T Osa, J Pajarinen, G Neumann, JA Bagnell, P Abbeel, and J Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018.

[4] Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance, 2020.

[5] J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1947.

[6] John W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136, 1964.

[7] Kenneth Joseph Arrow. *Aspects of the theory of risk-bearing*. Helsinki: Yrjo Jahnsonian Sa tio, 1965.

[8] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[9] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[10] John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica*, 55(5):999–1033, 1987.

[11] V. Joseph Hotz and Robert A. Miller. Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3):497–529, 1993.

[12] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[13] Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

[14] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1st edition, 1998.

[15] Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Annual Conference on Computational Learning Theory (COLT)*, 1998.

[16] A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2000.

[17] Bilal Piot, Matthieu Geist, and Olivier Pietquin. Learning from demonstrations: Is it worth estimating a reward function? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 17–32. Springer, 2013.

[18] Bilal Piot, Matthieu Geist, and Olivier Pietquin. Bridging the gap between imitation learning and inverse reinforcement learning. *IEEE transactions on neural networks and learning systems*, 28(8):1814–1826, 2016.

[19] J. Ho, J. K. Gupta, and S. Ermon. Model-free imitation learning with policy optimization. In *International Conference on Machine Learning (ICML)*, 2016.

[20] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2018.

[21] Kareem Amin and Satinder Singh. Towards resolving unidentifiability in inverse reinforcement learning, 2016.

[22] Kareem Amin, Nan Jiang, and Satinder Singh. Repeated inverse reinforcement learning, 2017.

[23] Amarildo Likmeta, Alberto Maria Metelli, Giorgia Ramponi, Andrea Tirinzoni, Matteo Giuliani, and Marcello Restelli. Dealing with multiple experts and non-stationarity in inverse reinforcement learning: an application to real-life problems. *Machine Learning*, 110(9):2541–2576, 2021.

[24] Rati Devidze, Goran Radanovic, Parameswaran Kamalaruban, and Adish Singla. Explicable reward design for reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34, 2021.

[25] Daniel Jarrett, Ioana Bica, and Mihaela van der Schaar. Strictly batch imitation learning by energy-based distribution matching, 2021.

[26] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping.

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

[28] Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions, 2020.

[29] Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[30] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2004.

[31] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *International Conference on Machine Learning (ICML)*, 2006.

[32] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *National Conference on Artificial Intelligence (AAAI)*, 2008.

[33] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, Carnegie Mellon University, 2010.

[34] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *Proc. Intl Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[35] Sreejith Balakrishnan, Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Harold Soh. Efficient exploration of reward functions in inverse reinforcement learning via bayesian optimization, 2020.

[36] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *International Joint Conference on Artifical Intelligence (IJCAI)*, 2007.

[37] Daniel S. Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast bayesian reward inference from preferences, 2020.

[38] Matteo Pirotta and Marcello Restelli. Inverse reinforcement learning through policy gradient minimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[39] Giorgia Ramponi, Gianluca Drappo, and Marcello Restelli. Inverse reinforcement learning from a gradient-based learner. 2020.

[40] Rakhoon Hwang, Hanjin Lee, and Hyung Ju Hwang. Option compatible reward inverse reinforcement learning. *Pattern Recognition Letters*, 154:83–89, 2022.

[41] Alberto Maria Metelli, Matteo Pirotta, and Marcello Restelli. Compatible reward inverse reinforcement learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[42] Tanmay Gangwani and Jian Peng. State-only imitation with transition dynamics mismatch. In *Proc. Intl Conf. on Learning Representations (ICLR)*, 2020.

[43] Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018.

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

[45] Daniel S. Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 330–359. PMLR, 30 Oct–01 Nov 2020.

[46] Daniel S Brown, Jordan Schneider, Anca Dragan, and Scott Niekum. Value alignment verification. 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 1105–1115. PMLR, 18–24 Jul 2021.

[47] 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.

[48] K. Dvijotham and E. Todorov. Inverse optimal control with linearly-solvable MDPs. In *International Conference on Machine Learning (ICML)*, 2010.

[49] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2004.

[50] S. Levine, Z. Popović, and V. Koltun. Nonlinear inverse reinforcement learning with Gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

[51] Alberto Maria Metelli, Giorgia Ramponi, Alessandro Concetti, and Marcello Restelli. Provably efficient learning of transferable rewards. In *International Conference on Machine Learning*, pages 7665–7676. PMLR, 2021.

[52] Tom Everitt and Marcus Hutter. Avoiding wireheading with value reinforcement learning. In *International Conference on Artificial General Intelligence*, pages 12–22. Springer, 2016.

[53] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.

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

[55] Srivatsan Srinivasan and Finale Doshi-Velez. Interpretable batch irl to extract clinician goals in icu hypotension management. *AMIA Summits on Translational Science Proceedings*, 2020:636, 2020.

[56] Daniel Jarrett, Alihan Hüyük, and Mihaela Van Der Schaar. Inverse decision modeling: Learning interpretable representations of behavior. In *International Conference on Machine Learning*, pages 4755–4771. PMLR, 2021.

[57] David Abel, Will Dabney, Anna Harutyunyan, Mark K Ho, Michael Littman, Doina Precup, and Satinder Singh. On the expressivity of markov reward. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7799–7812. Curran Associates, Inc., 2021.

[58] Samuelson Paul. A note on the pure theory of consumer behavior. *Economica*, 5(17):61–71, 1938.

[59] Paul A Samuelson. Consumption theory in terms of revealed preference. *Economica*, 15(60):243–253, 1948.

[60] Thomas Demuynck and Per Hjertstrand. Samuelson's approach to revealed preference theory: Some recent advances. *Paul Samuelson*, pages 193–227, 2019.

[61] Federico Echenique. New developments in revealed preference theory: decisions under risk, uncertainty, and intertemporal choice. *Annual Review of Economics*, 12:299–316, 2020.

[62] James Tobin. A proposal for international monetary reform. *Eastern economic journal*, 4(3/4):153–159, 1978.

[63] S. Joshi, R. Khardon, P. Tadepalli, A. Raghavan, and A. Fern. Solving relational mdps with exogenous events and additive rewards, 2013.

[64] Thomas G. Dietterich, George Trimponias, and Zhitang Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning, 2018.

[65] Sean McGregor, Rachel Houtman, Claire Montgomery, Ronald Metoyer, and Thomas G. Dietterich. Factoring exogenous state for model-free monte carlo, 2017.

[66] Vincent Liu, James Wright, and Martha White. Exploiting action impact regularity and exogenous state variables for offline reinforcement learning, 2021.

[67] Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2000.

[68] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2006.

[69] U. Syed, M. Bowling, and R.E. Schapire. Apprenticeship learning using linear programming. In *International Conference on Machine Learning (ICML)*, 2008.

[70] Eric Heim. A practitioner's guide to maximum causal entropy inverse reinforcement learning, starting from markov decision processes. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA PITTSBURGH United States, 2019.

[71] Michael Kearns. Near-optimal reinforcement learning in polynomial time. In *Machine Learning*, pages 260–268. Morgan Kaufmann, 1998.

[72] Mohammad Gheshlaghi Azar, Remi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model, 2012.

[73] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 9.1–9.24, Edinburgh, Scotland, 25–27 Jun 2012. PMLR.

[74] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Dimension-free tail inequalities for sums of random matrices, 2011.

[75] Ilya A Strebulaev and Toni M Whited. Dynamic models and structural estimation in corporate finance. *Final pre-publication version, published in Foundations and Trends in Finance*, 6:1–163, 2012.

[76] George Tauchen. Finite state markov-chain approximations to univariate and vector autoregressions. *Economics Letters*, 20(2):177–181, 1986.

[77] Parameswaran Kamalaruban, Yu-Ting Huang, Ya-Ping Hsieh, Paul Rolland, Cheng Shi, and Volkan Cevher. Robust reinforcement learning via adversarial training with langevin dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[78] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2017.

[79] Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2019.

[80] Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Craig Innes, Subramanian Ramamoorthy, and Adrian Weller. Robust learning from observation with model misspecification. *arXiv preprint arXiv:2202.06003*, 2022.

[81] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A Theory of Regularized Markov Decision Processes. In *International Conference on Machine Learning (ICML)*, 2019.

[82] Terence Tao. 254a, notes 3a: Eigenvalues and sums of hermitian matrices. *Terence Tao's blog*, 2010.

[83] Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, and Volkan Cevher. Robust inverse reinforcement learning under transition dynamics mismatch. *Advances in Neural Information Processing Systems*, 34:25917–25931, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the supplementary.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

We provide hereafter the proofs of the statements made in the main body.

A.1 Proof of Theorem 3

Let $r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ be the reward function common to the two experts, and let $v^1, v^2 \in \mathbb{R}^{|\mathcal{S}|}$ be the entropy regularized values functions associated experts 1 and 2 respectively and the reward function r. Then, according to Theorem 1, we have that $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned} r(s,a) &= \lambda \log \pi^1(a|s) - \gamma_1 \sum_{s'\in\mathcal{S}} T^1(s'|s,a) v^1(s') + v^1(s) \\ &= \lambda \log \pi^2(a|s) - \gamma_2 \sum_{s'\in\mathcal{S}} T^2(s'|s,a) v^2(s') + v^2(s) \end{aligned}$$

where π^1, π^2 denote the policies of experts 1 and 2 respectively. We hence deduce that $\forall a \in \mathcal{A}$,

$$\begin{pmatrix} I - \gamma_1 T^1_a & -(I - \gamma_2 T^2_a) \end{pmatrix} \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \lambda \log \pi^2(a|\cdot) - \lambda \log \pi^1(a|\cdot). \tag{13}$$

By including all available actions to the experts, we can write

$$\begin{pmatrix} I - \gamma_1 T^1_{a_1} & -(I - \gamma_2 T^2_{a_1}) \\ \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & -(I - \gamma_2 T^2_{a_{|\mathcal{A}|}}) \end{pmatrix} \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \begin{pmatrix} \lambda \log \pi^2(\cdot|a_1) - \lambda \log \pi^1(\cdot|a_1) \\ \vdots \\ \lambda \log \pi^2(\cdot|a_{|\mathcal{A}|}) - \lambda \log \pi^1(\cdot|a_{|\mathcal{A}|}) \end{pmatrix}. \tag{14}$$

Reward identifiability is directly related to the size of the solution space of the linear system (14). Since we assume that both experts are optimal with respect to a *true* reward function r, we know that the associated value function solves equation (14), and hence that this system is feasible. The solution space then depends on the rank of the matrix in the left hand side of (14), which we denote by A.

We first show that there always exists an eigenvector of A associated with eigenvalue 0. Indeed, since the matrices T_a are transition matrices, their rows must sum to 1, which can be written as $T_a \mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ is a $\mathcal{S}$ dimensional column vector of 1's. Hence,

$$A \begin{pmatrix} \frac{1}{1-\gamma_1}\mathbf{1} \\ \frac{1}{1-\gamma_2}\mathbf{1} \end{pmatrix} = \begin{pmatrix} \frac{1}{1-\gamma_1}(\mathbf{1} - \gamma_1 T^1_{a_1}\mathbf{1}) - \frac{1}{1-\gamma_2}(\mathbf{1} - \gamma_2 T^2_{a_1}\mathbf{1}) \\ \vdots \\ \frac{1}{1-\gamma_1}(\mathbf{1} - \gamma_1 T^1_{a_{|\mathcal{A}|}}\mathbf{1}) - \frac{1}{1-\gamma_2}(\mathbf{1} - \gamma_2 T^2_{a_{|\mathcal{A}|}}\mathbf{1}) \end{pmatrix} = \mathbf{0}$$

Hence, the vector $\begin{pmatrix} \frac{1}{1-\gamma_1}\mathbf{1} \\ \frac{1}{1-\gamma_2}\mathbf{1} \end{pmatrix}$ is an eigenvector of A with eigenvalue 0, and corresponds to the invariance of the optimal policy under addition of a constant to the reward function.

Suppose now that $\text{rank}(A) = 2|\mathcal{S}| - 1$. Since A has $2\mathcal{S}$ columns, this implies that the only eigenvector with eigenvalue 0 is $\begin{pmatrix} \frac{1}{1-\gamma_1}\mathbf{1} \\ \frac{1}{1-\gamma_2}\mathbf{1} \end{pmatrix}$, and thus that we can recover the value function v^1 (or v^2 equivalently) up to an additive constant. Using Theorem 1 again, it implies that we can also recover the reward function up to a constant.

On the other hand, suppose that $\text{rank}(A) < 2|\mathcal{S}| - 1$. Then, there exists another vector in $\text{Ker}(A)$ which is linearly independent of $\begin{pmatrix} \frac{1}{1-\gamma_1}\mathbf{1} \\ \frac{1}{1-\gamma_2}\mathbf{1} \end{pmatrix}$, and whose addition to the value function would not change the optimal policy. However, it is easy to check that the only eigenvector of A with eigenvalue 0 of the form $\begin{pmatrix} c_1\mathbf{1} \\ c_2\mathbf{1} \end{pmatrix}$ with $c_1, c_2 \in \mathbb{R}$ is proportional to $\begin{pmatrix} \frac{1}{1-\gamma_1}\mathbf{1} \\ \frac{1}{1-\gamma_2}\mathbf{1} \end{pmatrix}$. Hence, any other vector in $\text{Ker}(A)$ would induce a modification of the value and reward functions more complex than just adding a constant. The provided condition is hence also necessary.

Equivalence with Definition 2. It turns out that our rank condition (5) is equivalent to the value-distinguishing assumption of Definition 2. To show this, we first notice that, if v^1, v^2 satisfy equation (2), and if v^1 is a constant vector, then v^2 must also be a constant vector, and vice versa. Indeed, equation (2) can be written as

$$(I - \gamma_1 T_a^1)v^1 = (I - \gamma_2 T_a^2)v^2 \ \forall a \in \mathcal{A}.$$

Since, $\forall a \in \mathcal{A}, i = 1, 2$, $\mathbf{1}$ is an eigenvector of T_a^i with eigenvalue 1, then $\mathbf{1}$ is also an eigenvector of $(I - \gamma_2 T_a^2)^{-1}$ with eigenvalue $\frac{1}{1-\gamma_2}$. Hence, if $v^1 = c\mathbf{1}$ is a constant vector, then $v^2 = (I - \gamma_2 T_a^2)^{-1}(I - \gamma_1 T_a^1)v^1 = c\frac{1-\gamma_1}{1-\gamma_2}\mathbf{1}$ is also a constant vector, and the associated constant is determined by the constant of v^1. Thus, the condition of Definition 2 can be rewritten as

$$(I - \gamma_1 T_a^1)v^1 = (I - \gamma_2 T_a^2)v^2 \ \forall a \in \mathcal{A} \Rightarrow (v^1, v^2) = (c\mathbf{1}, c\frac{1-\gamma_1}{1-\gamma_2}\mathbf{1}) \text{ for some } c \in \mathbb{R}.$$

This is hence equivalent to

$$\dim\left(\text{Ker}\begin{pmatrix} I - \gamma_1 T_{a_1}^1 & I - \gamma_2 T_{a_1}^2 \\ \vdots & \vdots \\ I - \gamma_1 T_{a_{|\mathcal{A}|}}^1 & I - \gamma_2 T_{a_{|\mathcal{A}|}}^2 \end{pmatrix}\right) = 1.$$

which is equivalent to equation (5).

A.2 Proof of Corollary 4

Let $v^1, v^2, v^3 \in \mathbb{R}^{|\mathcal{S}|}$ be the entropy regularized value functions associated with experts $1, 2$ and 3 respectively. Following the proof of Theorem 3, these vectors must satisfy

$$\begin{pmatrix} I - \gamma_1 T_{a_1}^1 & -(I - \gamma_2 T_{a_1}^2) & \mathbf{0} \\ \vdots & \vdots & \vdots \\ I - \gamma_1 T_{a_{|\mathcal{A}|}}^1 & -(I - \gamma_2 T_{a_{|\mathcal{A}|}}^2) & \mathbf{0} \\ I - \gamma_1 T_{a_1}^1 & \mathbf{0} & -(I - \gamma_3 T_{a_1}^3) \\ \vdots & \vdots & \vdots \\ I - \gamma_1 T_{a_{|\mathcal{A}|}}^1 & \mathbf{0} & -(I - \gamma_3 T_{a_{|\mathcal{A}|}}^3) \end{pmatrix} \begin{pmatrix} v^1 \\ v^2 \\ v^3 \end{pmatrix} = \begin{pmatrix} \lambda \log \pi^2(\cdot|a_1) - \lambda \log \pi^1(\cdot|a_1) \\ \vdots \\ \lambda \log \pi^2(\cdot|a_{|\mathcal{A}|}) - \lambda \log \pi^1(\cdot|a_{|\mathcal{A}|}) \\ \lambda \log \pi^3(\cdot|a_1) - \lambda \log \pi^1(\cdot|a_1) \\ \vdots \\ \lambda \log \pi^3(\cdot|a_{|\mathcal{A}|}) - \lambda \log \pi^1(\cdot|a_{|\mathcal{A}|}) \end{pmatrix}. \tag{15}$$

Similarly as previously, we can easily show that the vector $\begin{pmatrix} \frac{1}{1-\gamma_1}\mathbf{1} \\ \frac{1}{1-\gamma_2}\mathbf{1} \\ \frac{1}{1-\gamma_3}\mathbf{1} \end{pmatrix} \in \text{Ker}(A')$, where A' denotes the matrix on the left of equation (15) In order for the reward to be recovered up to a constant, we hence need that there is no other linearly independent vector in $\text{Ker}(A')$, i.e., that $\text{rank}(A') = 3|\mathcal{S}| - 1$.

A.3 Proof of Corollary 5

We want to show that

$$\dim\left(\text{Ker}\begin{pmatrix} I - \gamma_1 T_{a_1} & -(I - \gamma_2 T_{a_1}) \\ \vdots & \vdots \\ I - \gamma_1 T_{a_{|\mathcal{A}|}} & -(I - \gamma_2 T_{a_{|\mathcal{A}|}}) \end{pmatrix}\right) = 1. \tag{16}$$

Suppose that $\begin{pmatrix} v^1 \\ v^2 \end{pmatrix} \in \text{Ker}\begin{pmatrix} I - \gamma_1 T_{a_1} & -(I - \gamma_2 T_{a_1}) \\ \vdots & \vdots \\ I - \gamma_1 T_{a_{|\mathcal{A}|}} & -(I - \gamma_2 T_{a_{|\mathcal{A}|}}) \end{pmatrix}$, i.e.,

$$(I - \gamma_1 T_a)v^1 = (I - \gamma_2 T_a)v^2 \ \forall a \in \mathcal{A}, \tag{17}$$

or equivalently

$$v^1 - v^2 = T_a(\gamma_1 v^1 - \gamma_2 v^2)\ \forall a \in \mathcal{A}. \tag{18}$$

Subtracting equation (18) for $a = a_1$ and $a = a_i$, we get

$$(T_{a_1} - T_{a_i})(\gamma_1 v^1 - \gamma_2 v^2) = 0\ \forall i. \tag{19}$$

Using equation (7) and the fact that the vector $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}|}$ always belongs to Ker $\begin{pmatrix} T_{a_1} - T_{a_2} \\ \vdots \\ T_{a_1} - T_{a_{|\mathcal{A}|}} \end{pmatrix}$, we have that Ker $\begin{pmatrix} T_{a_1} - T_{a_2} \\ \vdots \\ T_{a_1} - T_{a_{|\mathcal{A}|}} \end{pmatrix} = \text{Span}(\mathbf{1})$. Thus, we deduce from equation (19) that

$$\gamma_1 v^1 - \gamma_2 v^2 = c\mathbf{1} \tag{20}$$

for some $c \in \mathbb{R}$. Using the fact that for any $a \in \mathcal{A}$, $\mathbf{1}$ is an eigenvector of T_a with eigenvalue 1, we deduce from (18) and (20) that

$$v^1 - v^2 = T_a c\mathbf{1} = c\mathbf{1}. \tag{21}$$

Solving equations (20) and (21) for v^1 and v^2, we find $v^1 = \frac{c(1-\gamma_2)}{\gamma_1 - \gamma_2}\mathbf{1}$ and $v^2 = \frac{c(1-\gamma_1)}{\gamma_2 - \gamma_1}\mathbf{1}$. Therefore,

$$\text{Ker} \begin{pmatrix} I - \gamma_1 T^1_{a_1} & -(I - \gamma_2 T^2_{a_1}) \\ \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & -(I - \gamma_2 T^2_{a_{|\mathcal{A}|}}) \end{pmatrix} = \left\{ \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} : v^1 = \frac{c(1-\gamma_2)}{\gamma_1 - \gamma_2}\mathbf{1}, v^2 = \frac{c(1-\gamma_1)}{\gamma_2 - \gamma_1}\mathbf{1} \text{ for } c \in \mathbb{R} \right\}$$

which shows that condition (16) holds. On the other hand, if condition (7) does not hold, then Ker $\begin{pmatrix} T_{a_1} - T_{a_2} \\ \vdots \\ T_{a_1} - T_{a_{|\mathcal{A}|}} \end{pmatrix}$ contains another vector v_0 which is not a constant vector, so the reward cannot be recovered up to a constant.

A.4 Proof of Theorem 7

Suppose that $\mathbf{1} \in \text{Im} \begin{pmatrix} f_{a_1} \\ \vdots \\ f_{a_{|\mathcal{A}|}} \end{pmatrix}$, i.e., $\exists w \in \mathbb{R}^d$ such that $\begin{pmatrix} f_{a_1} \\ \vdots \\ f_{a_{|\mathcal{A}|}} \end{pmatrix} w = \mathbf{1}$. This implies that

$$\begin{pmatrix} I - \gamma_1 T^1_{a_1} & I - \gamma_2 T^2_{a_1} & \mathbf{0} \\ \vdots & \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & I - \gamma_2 T^2_{a_{|\mathcal{A}|}} & \mathbf{0} \\ I - \gamma_1 T^1_{a_1} & \mathbf{0} & f_{a_1} \\ \vdots & \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & \mathbf{0} & f_{a_{|\mathcal{A}|}} \end{pmatrix} \begin{pmatrix} \frac{1}{1-\gamma_1}\mathbf{1} \\ -\frac{1}{1-\gamma_2}\mathbf{1} \\ -w \end{pmatrix} = \mathbf{0}. \tag{22}$$

Suppose that condition (8) holds, i.e., that

$$\dim \left(\text{Ker} \begin{pmatrix} I - \gamma_1 T^1_{a_1} & I - \gamma_2 T^2_{a_1} & \mathbf{0} \\ \vdots & \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & I - \gamma_2 T^2_{a_{|\mathcal{A}|}} & \mathbf{0} \\ I - \gamma_1 T^1_{a_1} & \mathbf{0} & f_{a_1} \\ \vdots & \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & \mathbf{0} & f_{a_{|\mathcal{A}|}} \end{pmatrix} \right) = 1. \tag{23}$$

Equations (22) and (23) thus imply that

$$\text{Ker}\begin{pmatrix} I-\gamma_1 T^1_{a_1} & I-\gamma_2 T^2_{a_1} & \mathbf{0} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & I-\gamma_2 T^2_{a_{|\mathcal{A}|}} & \mathbf{0} \\ I-\gamma_1 T^1_{a_1} & \mathbf{0} & f_{a_1} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & \mathbf{0} & f_{a_{|\mathcal{A}|}} \end{pmatrix} = \text{Span}\begin{pmatrix} \frac{1}{1-\gamma_1}\mathbf{1} \\ -\frac{1}{1-\gamma_2}\mathbf{1} \\ -w \end{pmatrix}. \tag{24}$$

This means that for any v^1, v^2 satisfying $(I-\gamma_1 T^1_a)v^1 = (I-\gamma_1 T^2_a)v^2 \ \forall a \in \mathcal{A}$ and such that $\exists w \in \mathbb{R}^d$, $(I-\gamma_1 T^1_a)v^1 = f_a w \ \forall a \in \mathcal{A}$, then $v^1 \propto \mathbf{1}$.

Now suppose that we recover a reward function $r(s,a) = w^T f_{s,a}$ compatible with the two experts, i.e.,

$$r(\cdot, a) = r^*(\cdot, a) + (I-\gamma_1 T^1_a)v^1 \tag{25}$$

where $r^*(s,a) = w^{*T} f_{s,a}$ denotes the true reward and v^1 satisfies $(I-\gamma_1 T^1_a)v^1 = (I-\gamma_2 T^2_a)v^2 \ \forall a \in \mathcal{A}$. Then, $(I-\gamma_1 T^1_a)v^1 = r(\cdot,a) - r^*(\cdot,a) = f_a(w-w^*)$, and hence $\exists \tilde{w} \in \mathbb{R}^d$ such that $(I-\gamma_1 T^1_a)v^1 = f_a \tilde{w} \ \forall a \in \mathcal{A}$. Thus, $v^1 \propto \mathbf{1}$ and the reward is recovered up to a constant.

Suppose now that $\mathbf{1} \notin \text{Im}\begin{pmatrix} f_{a_1} \\ \vdots \\ f_{a_{|\mathcal{A}|}} \end{pmatrix}$. Then, the condition

$$\text{rank}\begin{pmatrix} I-\gamma_1 T^1_{a_1} & I-\gamma_2 T^2_{a_1} & \mathbf{0} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & I-\gamma_2 T^2_{a_{|\mathcal{A}|}} & \mathbf{0} \\ I-\gamma_1 T^1_{a_1} & \mathbf{0} & f_{a_1} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T^1_{a_{|\mathcal{A}|}} & \mathbf{0} & f_{a_{|\mathcal{A}|}} \end{pmatrix} = 2|\mathcal{S}| + d. \tag{26}$$

means that this matrix is full rank and hence that its kernel is $\{\mathbf{0}\}$. Thus, if we recover a reward of the form (25), following the same argument as previously, this means that $v^1 = 0$, and thus that the reward function is recovered exactly.

A.5 Proof of Corollary 6

Without loss of generality, let us assume that the exogenous variable can only take two possible values, i.e., the state space is defined as $\mathcal{S} = \{(s,e) : s \in \mathcal{S}_0, e \in \{e_1, e_2\}\}$, where e denotes the exogenous variable and $\mathcal{S}_0$ contains all other variables. Exogenity of variable e implies that $\forall e \in \{e_1, e_2\}$, $p(e^{t+1} = e_1 | s^t = s, e^t = e, a^t = a) = p(e^{t+1} = e_1 | e^t = e)$ does not depend on s nor a.

Suppose that we order the states as $\{\{(e_1, s)\}_{s \in \mathcal{S}_0}, \{(e_2, s)\}_{s \in \mathcal{S}_0}\}$. Then, the transition matrix for each expert i associated with action a has the following form:

$$T^i_a = \begin{pmatrix} p^i_1 T^i_{a,1} & (1-p^i_1)T^i_{a,1} \\ (1-p^i_2)T^i_{a,2} & p^i_2 T^i_{a,2} \end{pmatrix} \tag{27}$$

where for each expert i and exogenous variable $e_j, j = 1, 2$, $p^i_j = p^i(e^{t+1} = e_j | e^t = e_j)$ and $T^i_{a,j} \in \mathbb{R}^{|\mathcal{S}_0| \times |\mathcal{S}_0|}$ denotes the transition matrix of expert i for state variables in $\mathcal{S}_0$ knowing that the current value of state variable e is e_j, i.e. $T^i_{a,j}(s,s') = p^i(s^{t+1} = s' | s^t = s, e^t = e_j, a^t = a) \forall s, s' \in \mathcal{S}_0$ where p^i denotes the state transition probability in environment i.

We first show the result in the case of two experts. The matrix $A = \begin{pmatrix} I - \gamma_1 T^1_{a_1} & I - \gamma_2 T^2_{a_1} \\ \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & I - \gamma_2 T^2_{a_{|\mathcal{A}|}} \end{pmatrix}$ has the following form:

$$A = \begin{pmatrix} I - \gamma_1 p^1_1 T^1_{a_1,1} & -\gamma_1(1-p^1_1)T^1_{a_1,1} & I - \gamma_2 p^2_1 T^2_{a_1,1} & -\gamma_2(1-p^2_1)T^2_{a_1,1} \\ -\gamma_1(1-p^1_2)T^1_{a_1,2} & I - \gamma_1 p^1_2 T^1_{a_1,2} & -\gamma_2(1-p^2_2)T^2_{a_1,2} & I - \gamma_2 p^2_2 T^2_{a_1,2} \\ \vdots & \vdots & \vdots & \vdots \\ I - \gamma_1 p^1_1 T^1_{a_{|\mathcal{A}|},1} & -\gamma_1(1-p^1_1)T^1_{a_{|\mathcal{A}|},1} & I - \gamma_2 p^2_1 T^2_{a_{|\mathcal{A}|},1} & -\gamma_2(1-p^2_1)T^2_{a_{|\mathcal{A}|},1} \\ -\gamma_1(1-p^1_2)T^1_{a_{|\mathcal{A}|},2} & I - \gamma_1 p^1_2 T^1_{a_{|\mathcal{A}|},2} & -\gamma_2(1-p^2_2)T^2_{a_{|\mathcal{A}|},2} & I - \gamma_2 p^2_2 T^2_{a_{|\mathcal{A}|},2} \end{pmatrix}$$

We know that $v_0 = \begin{pmatrix} \frac{1}{1-\gamma_1}\mathbf{1} \\ \frac{1}{1-\gamma_1}\mathbf{1} \\ -\frac{1}{1-\gamma_2}\mathbf{1} \\ -\frac{1}{1-\gamma_2}\mathbf{1} \end{pmatrix}$ is an eigenvector of A with eigenvalue 0, corresponding to an addition of a constant to the reward. In order to show that the reward is not identifiable, we need to find another vector in Ker(A) linearly independent of v_0. We search for such a vector of the form $v_1 = \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \\ c_1\mathbf{1} \\ c_2\mathbf{1} \end{pmatrix}$. Using the fact that $\mathbf{1}$ is an eigenvector of any transition matrix with eigenvalue 1, the condition $v_1 \in$ Ker(A) is equivalent to

$$\begin{cases} -\gamma_1(1-p^1_1) + c_1(1-\gamma_2 p^2_1) - c_2\gamma_2(1-p^2_1) = 0 \\ 1 - \gamma_1 p^1_2 - c_1\gamma_2(1-p^2_2) + c_2(1-\gamma_2 p^2_2) = 0. \end{cases}$$

This system of equations turns out to have a unique solution for (c_1, c_2) since

$$\begin{aligned} \det \begin{pmatrix} 1-\gamma_2 p^2_1 & -\gamma_2(1-p^2_1) \\ -\gamma_2(1-p^2_2) & 1-\gamma_2 p^2_2 \end{pmatrix} &= (1-\gamma_2 p^2_1)(1-\gamma_2 p^2_2) - (\gamma_2 - \gamma_2 p^2_1)(\gamma_2 - \gamma_2 p^2_2) \\ &= (1-\gamma_2)(1+\gamma_2 - \gamma_2 p^2_1 - \gamma_2 p^2_2) > 0 \end{aligned}$$

since $0 \leq \gamma_2 < 1$. Hence, Ker(A) contains at least two linearly independent vector, and thus rank$(A) < 2|\mathcal{S}| - 1$. So, according to Theorem 3, the reward function is not identifiable up to a constant.

This means that, in addition to a global constant that we can add to the reward, we can also add a constant only to the rewards associated with a specific value of the exogenous variable. The proof naturally extends to the case of multiple experts, and when the exogenous variable can take more than two values. Actually, in the latter case, we can find even more linearly independent vectors in Ker(A), corresponding to adding a constant to the rewards associated with each possible value of the exogenous variable.

A.6 Proof of Theorem 8

Define $A = \begin{pmatrix} I - \gamma_1 T^1_{a_1} & I - \gamma_2 T^2_{a_1} \\ \vdots & \vdots \\ I - \gamma_1 T^1_{a_{|\mathcal{A}|}} & I - \gamma_2 T^2_{a_{|\mathcal{A}|}} \end{pmatrix}$ and $\hat{A} = \begin{pmatrix} I - \gamma_1 \hat{T}^1_{a_1} & I - \gamma_2 \hat{T}^2_{a_1} \\ \vdots & \vdots \\ I - \gamma_1 \hat{T}^1_{a_{|\mathcal{A}|}} & I - \gamma_2 \hat{T}^2_{a_{|\mathcal{A}|}} \end{pmatrix}$. For an arbitrary matrix M, let $\sigma_2(M)$ denote the second smallest singular value of M. Note that the condition (5) for A is equivalent to $\sigma_2(A) > 0$. From Weyl's inequality for singular values[82], we have that

$$|\sigma_2(A) - \sigma_2(\hat{A})| \leq \|A - \hat{A}\|_2.$$

Moreover,

$$
\begin{aligned}
\|A-\hat{A}\|_2 &= \left\| \begin{pmatrix} \gamma_1(T^1_{a_1}-\hat{T}^1_{a_1}) & \gamma_2(T^2_{a_1}-\hat{T}^2_{a_1}) \\ \vdots & \vdots \\ \gamma_1(T^1_{a_{|\mathcal{A}|}}-\hat{T}^1_{a_{|\mathcal{A}|}}) & \gamma_2(T^2_{a_{|\mathcal{A}|}}-\hat{T}^2_{a_{|\mathcal{A}|}}) \end{pmatrix} \right\|_2 \\
&\leq \sqrt{2}\max(\gamma_1,\gamma_2)\max\left(\left\| \begin{pmatrix} (T^1_{a_1}-\hat{T}^1_{a_1}) \\ \vdots \\ T^1_{a_{|\mathcal{A}|}}-\hat{T}^1_{a_{|\mathcal{A}|}} \end{pmatrix} \right\|_2, \left\| \begin{pmatrix} (T^2_{a_1}-\hat{T}^2_{a_1}) \\ \vdots \\ T^2_{a_{|\mathcal{A}|}}-\hat{T}^2_{a_{|\mathcal{A}|}} \end{pmatrix} \right\|_2 \right) \\
&\leq \sqrt{2|\mathcal{A}|}\max(\gamma_1,\gamma_2)\epsilon.
\end{aligned}
$$

Therefore, $\sigma_2(A) \geq \sigma_2(\hat{A}) - \sqrt{2|\mathcal{A}|}\max(\gamma_1,\gamma_2)\epsilon$, and hence $\sigma_2(A) > 0$ provided that $\sigma_2(\hat{A}) > \sqrt{2|\mathcal{A}|}\max(\gamma_1,\gamma_2)\epsilon$.

A.7 Proof of Theorem 9

Proof. $\hat{T}_a$ can be constructed as follows. Sample $\frac{N}{|\mathcal{S}|}$ states $\{s'_i\}_{i=1}^{\frac{N}{|\mathcal{S}|}}$ from the distribution $T(\cdot|s,a)$ for every state $s \in \mathcal{S}$. Let $N(s)$ denote the number of times state s has been sampled, i.e. $N(s) = \frac{N}{|\mathcal{S}|}$. Form the matrix $\tilde{T}_i = [\frac{\mathbf{1}(s_i=s,s'_i=s')N}{N(s)}]_{s,s'}$ It holds that $\forall i$, $\mathbb{E}[\tilde{T}_i] = T_a$, $\lambda_{\max}(\tilde{T}_i - T_a) \leq |\mathcal{S}|+1$, $\lambda_{\max}(\mathbb{E}[(\tilde{T}_i - T_a)^2]) \leq |\mathcal{S}|^2$ and $\mathrm{Trace}(\mathbb{E}[(\tilde{T}_i - T_a)^2]) \leq |\mathcal{S}|^2$. Then, the result follows applying Lemma 10 in [73] and assuming $\delta < 1/e$. Finally, we conclude with a covering argument over the set $\mathcal{A}$. □

A.8 Proof of Theorem 11

Lemma 13. *The condition of equation 12 holds if and only if* $\forall v^1, v^2 \in \mathbb{R}^{|\mathcal{S}|}$ *satisfying* $(I - \gamma_1 T^1_a)v^1 = (I-\gamma_2 T^2_a)v^2$, $\forall a \in \mathcal{A}$, *there exists* $v^3 \in \mathbb{R}^{|\mathcal{S}|}$ *such that* $(I-\gamma_3 T^3_a)v^3 = (I-\gamma_1 T^1_a)v^1$, $\forall a \in \mathcal{A}$.

Proof. Denote by A_1, A_2 the matrices shown and the left and right hand side of equation (12) respectively, so that the equation reads $\mathrm{rank}(A_1) = \mathrm{rank}(A_2) - |\mathcal{S}|$, or equivalently $2|\mathcal{S}| - \mathrm{rank}(A_1) = 3|\mathcal{S}| - \mathrm{rank}(A_2)$. Using the rank theorem, it follows that $\dim(\mathrm{Ker}(A_1)) = \dim(\mathrm{Ker}(A_2))$, i.e.,

$$
\begin{aligned}
&\dim(\{(v^1,v^2) \in \mathbb{R}^{2|\mathcal{S}|} : (I-\gamma_1 T^1_a)v^1 = (I-\gamma_2 T^2_a)v^2 \ \forall a \in \mathcal{A}\}) \\
&\quad = \dim(\{(v^1,v^2,v^3) \in \mathbb{R}^{3|\mathcal{S}|} : (I-\gamma_1 T^1_a)v^1 = (I-\gamma_2 T^2_a)v^2 = (I-\gamma_3 T^3_a)v^3 \ \forall a \in \mathcal{A}\}).
\end{aligned} \tag{28}
$$

Since all matrices $I - \gamma_3 T^3_a$ are invertible for any $a \in \mathcal{A}$, it follows that for any $(v^1,v^2) \in \mathbb{R}^{2|\mathcal{S}|}$, there can exist at most one vector $v^3 \in \mathbb{R}^{|\mathcal{S}|}$ such that $(I-\gamma_1 T^1_a)v^1 = (I-\gamma_2 T^2_a)v^2 = (I-\gamma_3 T^3_a)v^3 \ \forall a \in \mathcal{A}$. We hence deduce that

$$
\begin{aligned}
&\dim(\{(v^1,v^2,v^3) \in \mathbb{R}^{3|\mathcal{S}|} : (I-\gamma_1 T^1_a)v^1 = (I-\gamma_2 T^2_a)v^2 = (I-\gamma_3 T^3_a)v^3 \ \forall a \in \mathcal{A}\}) \\
&= \dim(\{(v^1,v^2) \in \mathbb{R}^{2|\mathcal{S}|} : \exists v^3 \in \mathbb{R}^{|\mathcal{S}|}, (I-\gamma_1 T^1_a)v^1 = (I-\gamma_2 T^2_a)v^2 = (I-\gamma_3 T^3_a)v^3 \ \forall a \in \mathcal{A}\}).
\end{aligned} \tag{29}
$$

Plugging equation (29) in (28), we have

$$
\begin{aligned}
&\dim(\{(v^1,v^2) \in \mathbb{R}^{2|\mathcal{S}|} : (I-\gamma_1 T^1_a)v^1 = (I-\gamma_2 T^2_a)v^2 \ \forall a \in \mathcal{A}\}) \\
&= \dim(\{(v^1,v^2) \in \mathbb{R}^{2|\mathcal{S}|} : \exists v^3 \in \mathbb{R}^{|\mathcal{S}|}, (I-\gamma_1 T^1_a)v^1 = (I-\gamma_2 T^2_a)v^2 = (I-\gamma_3 T^3_a)v^3 \ \forall a \in \mathcal{A}\}).
\end{aligned} \tag{30}
$$

Moreover, we can clearly see that

$$\{(v^1, v^2) \in \mathbb{R}^{2|\mathcal{S}|} : \exists v^3 \in \mathbb{R}^{|\mathcal{S}|}, (I - \gamma_1 T_a^1)v^1 = (I - \gamma_2 T_a^2)v^2 = (I - \gamma_3 T_a^3)v^3 \ \forall a \in \mathcal{A}\}$$
$$\subseteq \{(v^1, v^2) \in \mathbb{R}^{2|\mathcal{S}|} : (I - \gamma_1 T_a^1)v^1 = (I - \gamma_2 T_a^2)v^2 \ \forall a \in \mathcal{A}\}.$$

Thus, together with equation (30), we can conclude that

$$\{(v^1, v^2) \in \mathbb{R}^{2|\mathcal{S}|} : \exists v^3 \in \mathbb{R}^{|\mathcal{S}|}, (I - \gamma_1 T_a^1)v^1 = (I - \gamma_2 T_a^2)v^2 = (I - \gamma_3 T_a^3)v^3 \ \forall a \in \mathcal{A}\}$$
$$= \{(v^1, v^2) \in \mathbb{R}^{2|\mathcal{S}|} : (I - \gamma_1 T_a^1)v^1 = (I - \gamma_2 T_a^2)v^2 \ \forall a \in \mathcal{A}\}$$

which shows the result.

Suppose now that condition 12 does not hold, i.e.,

$$\dim(\{(v^1, v^2) \in \mathbb{R}^{2|\mathcal{S}|} : (I - \gamma_1 T_a^1)v^1 = (I - \gamma_2 T_a^2)v^2 \ \forall a \in \mathcal{A}\})$$
$$> \dim(\{(v^1, v^2) \in \mathbb{R}^{3|\mathcal{S}|} : \exists v^3 \in \mathbb{R}^{|\mathcal{S}|}, (I - \gamma_1 T_a^1)v^1 = (I - \gamma_2 T_a^2)v^2 = (I - \gamma_3 T_a^3)v^3 \ \forall a \in \mathcal{A}\}). \tag{31}$$

This directly implies that there must exist a pair (v^1, v^2), such that there exists no $v^3 \in \mathbb{R}^{|\mathcal{S}|}$ satisfying $(I - \gamma_3 T_a^3)v^3 = (I - \gamma_1 T_a^1)v^1 \ \forall a \in \mathcal{A}$ hence finalizing the proof.

□

We now turn to the proof of Theorem 11. Let r^* be the ground truth reward, and suppose that we recover some reward function r from policies π^1, π^2, i.e., π^1, π^2 are optimal with respect to both rewards r and r^* on $(T^1, \gamma_1), (T^2, \gamma_2)$ respectively. Suppose that we train a policy π^3 optimally with respect to r on (T^3, γ_3). We want to show that π^3 is also optimal with respect to the true reward r^*.

Let v^i, v_*^i be the value vectors associated to expert $i = 1, 2$ with respect to rewards r and r^* respectively, i.e., such that

$$r(\cdot, a) = \lambda \log \pi^1(a|\cdot) + (I - \gamma_1 T_a^1)v^1 = \lambda \log \pi^2(a|\cdot) + (I - \gamma_2 T_a^2)v^2 \tag{32}$$
$$r^*(\cdot, a) = \lambda \log \pi^1(a|\cdot) + (I - \gamma_1 T_a^1)v_*^1 = \lambda \log \pi^2(a|\cdot) + (I - \gamma_2 T_a^2)v_*^2. \tag{33}$$

Let v^3 be the value vector associated with expert 3 with respect to reward r, i.e., such that $\forall a \in \mathcal{A}$

$$r(\cdot, a) = \lambda \log \pi^3(a|\cdot) + (I - \gamma_3 T_a^3)v^3. \tag{34}$$

We need to show that there exists a vector $v_*^3 \in \mathbb{R}^{|\mathcal{S}|}$ such that $\forall a \in \mathcal{A}$

$$r^*(\cdot, a) = \lambda \log \pi^3(a|\cdot) + (I - \gamma_3 T_a^3)v_*^3. \tag{35}$$

Using equations (32), (33) and (34), we have $\forall a \in \mathcal{A}$

$$r^*(\cdot, a) = \lambda \log \pi^1(a|\cdot) + (I - \gamma_1 T_a^1)v_*^1 \tag{36}$$
$$= r(\cdot, a) - (I - \gamma_1 T_a^1)v^1 + (I - \gamma_1 T_a^1)v_*^1 \tag{37}$$
$$= \lambda \log \pi^3(a|\cdot) + (I - \gamma_3 T_a^3)v^3 + (I - \gamma_1 T_a^1)(v^1 - v_*^1). \tag{38}$$

Moreover, subtracting equations (32) and (33), we have

$$(I - \gamma_1 T_a^1)(v^1 - v_*^1) = (I - \gamma_2 T_a^2)(v^2 - v_*^2)$$

Therefore, using our assumption and Lemma 13, there exists a vector $\tilde{v}_3 \in \mathbb{R}^{|\mathcal{S}|}$ such that $(I - \gamma_1 T_a^1)(v^1 - v_*^1) = (I - \gamma_3 T_a^3)\tilde{v}_3$. Hence, combined with equation (38), we conclude that there exists $v_*^3 \in \mathbb{R}^{|\mathcal{S}|}$ such that $\forall a \in \mathcal{A}$

$$r^*(\cdot, a) = \lambda \log \pi^3(a|\cdot) + (I - \gamma_3 T_a^3)v_*^3. \tag{39}$$

Using Theorem 1, we conclude that r^* belongs to the set rewards compatible with π^3, and hence that π^3, which has been optimized for r, is also optimal for the ground truth reward r^*.

On the other hand, if condition 12 does not hold, according to Lemma 13, we can construct a reward function r compatible with experts 1 and 2 that cannot be written in the form $r(\cdot, a) = \lambda \log \pi^3(a|\cdot) + (I - \gamma_3 T_a^3)v^3$ for some $v^3 \in \mathbb{R}^{|\mathcal{S}|}$. Hence, thanks to Theorem 1, the policy π^3 cannot be optimal for such a reward function. Hence, there will necessarily exist some recovered reward functions that would lead to a sub-optimal policy in environment 3.

A.9 Proof of Corollary 12

For the setup describe in this corollary, we need to verify condition (12). By the rank theorem, this condition is equivalent to

$$\dim\left(\text{Ker}\begin{pmatrix} I-\gamma_1 T_{a_1} & I-\gamma_2 T_{a_1} \\ \vdots & \vdots \\ I-\gamma_1 T_{a_{|\mathcal{A}|}} & I-\gamma_2 T_{a_{|\mathcal{A}|}} \end{pmatrix}\right) = \dim\left(\text{Ker}\begin{pmatrix} I-\gamma_1 T_{a_1} & I-\gamma_2 T_{a_1} & \mathbf{0} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T_{a_{|\mathcal{A}|}} & I-\gamma_2 T_{a_{|\mathcal{A}|}} & \mathbf{0} \\ I-\gamma_1 T_{a_1} & \mathbf{0} & I-\gamma_3 T_{a_1} \\ \vdots & \vdots & \vdots \\ I-\gamma_1 T_{a_{|\mathcal{A}|}} & \mathbf{0} & I-\gamma_3 T_{a_{|\mathcal{A}|}} \end{pmatrix}\right). \tag{40}$$

To this end, we will show that any element $(v^1, v^2) \in \mathbb{R}^{2|\mathcal{S}|}$ of the kernel space of the left hand side is associated a single element $(v^1, v^2, v^3) \in \mathbb{R}^{3|\mathcal{S}|}$ of the kernel space of the right hand side. More precisely, we need to show that for any v^1, v^2 satisfying

$$(I-\gamma_1 T_a)v^1 = (I-\gamma_2 T_a)v^2 \ \forall a \in \mathcal{A},$$

there exists a unique $v^3 \in \mathbb{R}^{|\mathcal{S}|}$ such that

$$(I-\gamma_1 T_a)v^1 = (I-\gamma_3 T_a)v^3 \ \forall a \in \mathcal{A}.$$

Consider the action a_0 satisfying by assumption that T_{a_0} commutes with all other matrices T_a, $a \in \mathcal{A}$. Define $v^3 = (I-\gamma_3 T_{a_0})^{-1}(I-\gamma_1 T_{a_0})v^1$. Notice that for any $a \in \mathcal{A}$, we can write $I-\gamma_3 T_a = \alpha(I-\gamma_1 T_a) + (1-\alpha)(I-\gamma_2 T_a)$ where $\alpha = \frac{\gamma_3-\gamma_2}{\gamma_1-\gamma_2}$. Moreover, recall that, if any two invertible matrices A and B commute, then A and B^{-1} also commute.

Using these properties, we then have for any $a \in \mathcal{A}$,

$$\begin{aligned}
(I-\gamma_3 T_a)v^3 &= \alpha(I-\gamma_1 T_a)v^3 + (1-\alpha)(I-\gamma_2 T_a)v^3 \\
&= \alpha(I-\gamma_1 T_a)(I-\gamma_3 T_{a_0})^{-1}(I-\gamma_1 T_{a_0})v^1 + (1-\alpha)(I-\gamma_2 T_a)(I-\gamma_3 T_{a_0})^{-1}(I-\gamma_1 T_{a_0})v^1 \\
&= \alpha(I-\gamma_1 T_a)(I-\gamma_3 T_{a_0})^{-1}(I-\gamma_1 T_{a_0})v^1 + (1-\alpha)(I-\gamma_2 T_a)(I-\gamma_3 T_{a_0})^{-1}(I-\gamma_2 T_{a_0})v^2 \\
&= \alpha(I-\gamma_1 T_a)(I-\gamma_3 T_{a_0})^{-1}(I-\gamma_1 T_{a_0})v^1 + (1-\alpha)(I-\gamma_3 T_{a_0})^{-1}(I-\gamma_2 T_{a_0})(I-\gamma_2 T_a)v^2 \\
&= \alpha(I-\gamma_1 T_a)(I-\gamma_3 T_{a_0})^{-1}(I-\gamma_1 T_{a_0})v^1 + (1-\alpha)(I-\gamma_3 T_{a_0})^{-1}(I-\gamma_2 T_{a_0})(I-\gamma_1 T_a)v^1 \\
&= (I-\gamma_1 T_a)(I-\gamma_3 T_{a_0})^{-1}(\alpha(I-\gamma_1 T_{a_0}) + (1-\alpha)(I-\gamma_2 T_{a_0}))v^1 \\
&= (I-\gamma_1 T_a)v^1.
\end{aligned}$$

Uniqueness of v^3 is trivial since the matrices $(I-\gamma_3 T_a)$ are invertible, which shows that condition (40) holds.

Counter-example when the commutativity constraint does not hold. We now provide a simple example showing that the required generalizability condition (12) does not always hold in the case where the commutativity condition breaks. Suppose $|\mathcal{S}| = 3$, $|\mathcal{A}| = 2$ and

$$T_{a_1} = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0.5 & 0.5 \end{pmatrix}, \quad T_{a_2} = \begin{pmatrix} 0.3 & 0.4 & 0.3 \\ 0.7 & 0.1 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{pmatrix}. \tag{41}$$

These matrices do not commute and we have for any discount factors $\gamma_1, \gamma_2, \gamma_3$ all different,

$$4 = \text{rank}\begin{pmatrix} I-\gamma_1 T_{a_1} & I-\gamma_2 T_{a_1} \\ I-\gamma_1 T_{a_2} & I-\gamma_2 T_{a_2} \end{pmatrix} \neq \text{rank}\begin{pmatrix} I-\gamma_1 T_{a_1} & I-\gamma_2 T_{a_1} & \mathbf{0} \\ I-\gamma_1 T_{a_2} & I-\gamma_2 T_{a_2} & \mathbf{0} \\ I-\gamma_1 T_{a_1} & \mathbf{0} & I-\gamma_3 T_{a_1} \\ I-\gamma_1 T_{a_2} & \mathbf{0} & I-\gamma_3 T_{a_2} \end{pmatrix} - |\mathcal{S}| = 5. \tag{42}$$

B Algorithms details

This section provides the detailed pseudocode of the procedures we introduced for reward identification (Algorithm 1), for generalizability (Algorithm 3) and identification when the reward function can be expressed as linear combination of known features (Algorithm 2).

Algorithm 1 `Identifiability Test`

Input: Expert transition matrices T_1, T_2, entropy-regularized optimal policies π_1, π_2.
Compute matrix

$$A := \begin{pmatrix} -(I - \gamma_1 T^1_{a_1}) & I - \gamma_2 T^2_{a_1} \\ \vdots & \vdots \\ -(I - \gamma_1 T^1_{a_{|\mathcal{A}|}}) & I - \gamma_2 T^2_{a_{|\mathcal{A}|}} \end{pmatrix} \tag{43}$$

if $\text{rank}(A) = 2|\mathcal{S}| - 1$ **then**
  Identifiable = True
  Form vector $b \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that $b(s,a) = \lambda \log \frac{\pi^1(a|s)}{\pi^2(a|s)}$ (ordered by states first)
  Recover value vectors $\begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = (A^T A)^{-1} A^T b$
  Recover the reward function as $r(s,a) = \lambda \log \pi^1(a|s) + \gamma \sum_{s'} T_1(s'|s,a) v^1(s') - v^1(s)$ or equivalently $r(s,a) = \lambda \log \pi^2(a|s) + \gamma \sum_{s'} T_2(s'|s,a) v^2(s') - v^2(s)$
else
  Identifiable = False
end if
Output: Identifiable and recovered reward r.

Algorithm 2 `Identifiability Test with linear reward function`

Input: Expert transition matrices T^1, T^2, entropy-regularized optimal policies π_1, π_2, features set $\{f_a\}_a$.
Compute matrix

$$A := \begin{pmatrix} -(I - \gamma_1 T^1_{a_1}) & I - \gamma_2 T^2_{a_1} & \mathbf{0} \\ \vdots & \vdots & \vdots \\ -(I - \gamma_1 T^1_{a_{|\mathcal{A}|}}) & I - \gamma_2 T^2_{a_{|\mathcal{A}|}} & \mathbf{0} \\ -(I - \gamma_1 T^1_{a_1}) & \mathbf{0} & f_{a_1} \\ \vdots & \vdots & \vdots \\ -(I - \gamma_1 T^1_{a_{|\mathcal{A}|}}) & \mathbf{0} & f_{a_{|\mathcal{A}|}} \end{pmatrix} \tag{44}$$

if $\text{rank}(A) = 2|\mathcal{S}| + d$ **then**
  Identifiable = True
  Form vectors $b_1, b_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ defined as $b_1(s,a) = \lambda \log \frac{\pi^1(a|s)}{\pi^2(a|s)}$, $b_2(s,a) = \lambda \log \pi^1(a|s)$ and $b \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}$ as $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$
  Recover value vectors and reward weights $\begin{pmatrix} v^1 \\ v^2 \\ w \end{pmatrix} = (A^T A)^{-1} A^T b$
  Recover the reward function as $r(s,a) = w^T f_{s,a}$
else
  Identifiable = False
end if
Output: Identifiable and recovered reward r.

Algorithm 3 `Generalization Test`

Input: Expert transition matrices T^1, T^2, transfer transition matrix T^3, entropy-regularized optimal policies π_1, π_2.
Compute matrix

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

if the condition in Equation (12) holds **then**
  Generalizable = True
  Form vector $b \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that $b(s,a) = \lambda \log \frac{\pi^1(a|s)}{\pi^2(a|s)}$
  Recover the value vectors $\begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = (A^T A)^{-1} A^T b$
  Recover the reward function as $r(s,a) = \lambda \log \pi^1(a|s) + \gamma \sum_{s'} T_1(s'|s,a) v^1(s') - v^1(s)$
  Recover the optimal entropy regularized policy π^3 in T^3 using the recovered reward r with any RL algorithm.
else
  Generalizable = False
end if
Output: Generalizable and recovered policy π^3.

Algorithms 1 and 3 can be generalized to an arbitrary number of experts. Indeed, denoting the matrix in Equation (43) as A_2, we can construct the matrix A_n for n experts recursively as follows:

$$A_n := \begin{pmatrix} A_{n-1} & & \mathbf{0} \\ -(I - \gamma_1 T^1_{a_1}) & \mathbf{0} & I - \gamma_n T^n_{a_1} \\ \vdots & \vdots & \vdots \\ -(I - \gamma_1 T^1_{a_{|\mathcal{A}|}}) & \mathbf{0} & I - \gamma_n T^n_{a_{|\mathcal{A}|}} \end{pmatrix} \tag{45}$$

Similarly, we can construct the vector b_n as

$$b_n := \begin{pmatrix} b_{n-1} \\ \lambda \log \frac{\pi^1(a|s)}{\pi^n(a|s)} \end{pmatrix} \tag{46}$$

where b_1 denotes the vector defined in the algorithms for 2 experts. The rest of the procedures remain unchanged.

C Additional experiments

This section provides the experimental results and environment details omitted from the main text.

Additional details for `Gridworld` In the main text, we omitted the description of the reward function. We provide it hereafter for completeness. The reward function is obtained assigning a value at every state according to the grid shown in Figure 5. This reward function would depend only on states. To obtain a state-action dependent reward function, we add a penalty of -30 for moving right, -20 for moving down, -10 for moving left and 0 for a step upwards.

Additional details for `WindyGridworld` In `WindyGridworld`, the agent moves of one step according to the next state sampled from $T_\alpha(s'|s,a) = (1-\alpha)T_{\det}(s'|s,a) + \alpha U(s'|s,a)$ where $T_{\det}(s'|s,a)$ as in `Gridworld`. In addition to that the agent takes an additional step according to the wind direction. The wind direction w is sampled from the wind distribution generated by sampling each entry of the non normalized P_{wind} from a normal distribution and normalizing the obtained vector. After sampling the wind direction we sample the corresponding next state from $T_{\det}(s'|s,w)$.

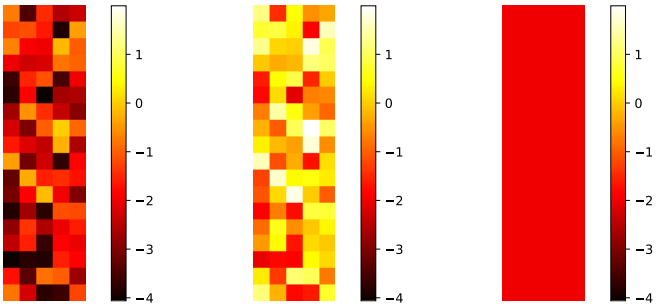
The reward function is the same used for the environment `Gridworld`.

Results on `Random-Matrices` We report in Figure 4, the results omitted from the main text. In Figure 4, we show the reward recovered with Algorithm 1 and the difference with respect to the true reward. It clearly emerges that the recovered reward is within a constant shift from the true reward function.

Results on `Gridworld` **with state only reward** We provide an additional result on `Gridworld` where we do not consider the penalty assigned to the different actions. In this case, the reward depends only on states but the learner is not informed about this feature. In Figure 5, we show the recovered reward. Given that the reward depends only on the states we show the 2D representation of the state space suppressing the action dimension. We used the `Gridworld` implementation released with [83].

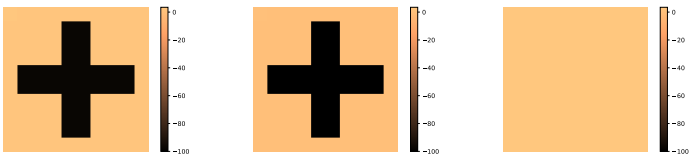
Results on `WindyGridworld` **with different discount factors** In the main text we showed that we need to observe 4 experts to generalize to a new wind distribution. Hereafter, we provide experiments on the generalization to a new environment with a different discount factor. We verified that in this case observing two experts is enough to generalize. The comparison between recovered rewards and policies can be found in Figure 6.

Computational resources The experiment can be reproduced with a standard laptop.



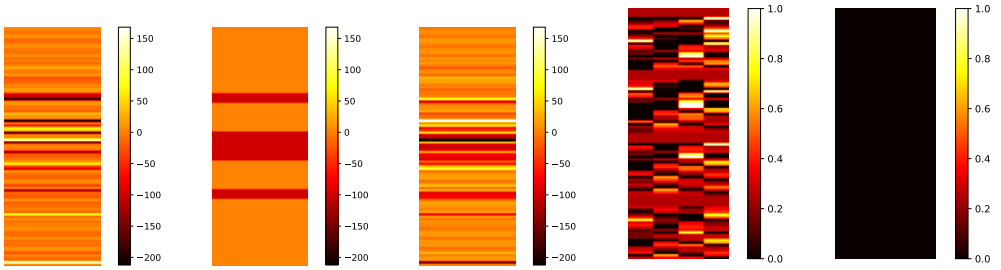
(a) r (b) r_{true} (c) $r - r_{\text{true}}$

Figure 4: Comparison between true and recovered reward in `Random-Matrices` with $|\mathcal{S}| = 18$ and $|\mathcal{A}| = 5$. On the vertical axis corresponds to the canonical ordering of the 18 states while the horizontal axis corresponds to the 5 actions.



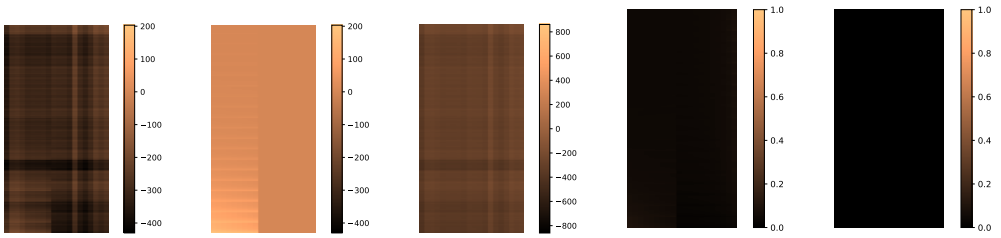
(a) r (b) r_{true} (c) $r - r_{\text{true}}$

Figure 5: Comparison between true and recovered reward in `Gridworld` with $|\mathcal{S}| = 100$ and the 4 actions up, down, left and right. It can be noticed that the reward function is recovered up to a constant shift.



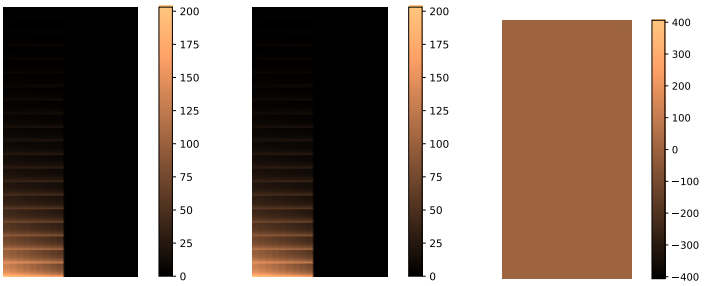
(a) r (b) r_{true} (c) $r - r_{\text{true}}$ (d) $\pi_{r_{\text{true}}}^{T_{\text{test}}}$ (e) $\pi_{r_{\text{true}}}^{T_{\text{test}}} - \pi_{r}^{T_{\text{test}}}$

Figure 6: Generalization in `WindyGridworld` with different discount factors. We observe two experts with discounts factor γ_1 and γ_2 with $\gamma_1 \neq \gamma_2$ and with common transition dynamics. Subplot (e) shows that the policy recovered from r_{true} in a new environment with a different γ_3 matches the policy obtained from the recovered reward.



(a) r (b) r_{true} (c) $r - r_{\text{true}}$ (d) $\pi_{r_{\text{true}}}^{T_{\text{test}}}$ (e) $\pi_{r_{\text{true}}}^{T_{\text{test}}} - \pi_{r}^{T_{\text{test}}}$

Figure 7: Comparison between true and recovered reward in `Strebulaev-Whited` with $|\mathcal{S}| = 400$ and the 20 actions. It can be clearly noticed that the reward function is not identified (see subplots (a), (b), (c)). However, when we use the recovered reward in subplot (a) to train an optimal policy under unseen dynamics we recover the optimal policy under the true reward in subplot (b). The subplots (d) show the policies recovered from the true reward and (e) shows the difference between the policy recovered from r_{true} and from the recovered reward.



(a) r (b) r_{true} (c) $r - r_{\text{true}}$

Figure 8: Comparison between true and recovered reward in `Strebulaev-Whited` assuming additional knowledge of the features $\{f_a\}_a$. It emerges that thanks to such additional information the reward function is identifiable.

D Extension to Regularized MDPs

It turns out that our identifiability result is valid more generally for regularized MDPs [81] where the entropy term in equation (1) is replaced by any other strongly convex differentiable function of the policy $\Omega(\pi)$.

Indeed, we can use Proposition 1 and Definition 1 in [81] to establish that for any value vector any v and reward r, there exists a unique policy that satisfies

$$\pi(a|s) = \nabla\Omega^*(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s'|s,a)v(s'))$$

where Ω^* denotes the Fenchel conjugate of Ω. By the distributivity property (iii) in Proposition 1 of [5], we can subtract a function dependent only on state in the argument without affecting the equality. This gives that for any v and r, there exists a unique π such that

$$\pi(a|s) = \nabla\Omega^*(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s'|s,a)v(s') - v(s))$$

Using the convexity of Ω, we have that $\nabla\Omega$ is the inverse map of $\nabla\Omega^*$. Hence we obtain

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

which is the equivalent of our Theorem 1 for general strongly convex regularizers. The only part changing is the left hand side. However, we saw in the analysis that reward identifiability was not depending on this part of the equation. When using a different regularizer, the recovered reward given observed expert policies will be different, but the identifiability condition remains the same.

This extension relax the assumption of entropy regularized experts but, unfortunately, epsilon-greedy or deterministic greedy policies would not fit this setting. Identifiability is more challenging with these kinds of experts because the knowledge of such policies only informs us with the action yielding the highest expected value, but no information about the relative difference with respect to other actions, in contrast with regularized stochastic policies.