Identifiability in Inverse Reinforcement Learning

Team: BRICS

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

1 Introduction to IRL

Inverse reinforcement learning (IRL) is the process of estimating a reward function from demonstrations of optimal behavior.

Optimal behavior → IRL → Estimated reward function

Problem: In IRL, under what conditions are the recovered reward function unique (up to constants)?
Application example: Recovering surgeon's reward for knot-tying from video, then fine-tuning a robot to replicate suturing motions.
Regularized reward: To resolve ambiguity, we consider Maximum Entropy RL:

$$J_{\text{MaxEnt}}(\pi; d, r, T) = \mathbb{E}_\pi\Big[\sum_{t=0}^{T} \gamma^t \big(r(x_t, a_t) - \lambda \log \pi(a_t \mid x_t)\big)\Big]$$

Definition 1. MDP is called ***weakly identifiable*** if it holds that: the two reward functions coincide up to constant along all trajectories $\iff$ policies induced by the rewards coincide.

Definition 2. MDP is called ***strongly identifiable*** if it holds that: the two reward functions coincides up to constant at every point (state, action) $\iff$ policies induced by the rewards coincide.

Definition 3. We say that $(\mathcal{T}_1, \gamma_1)$ and $(\mathcal{T}_2, \gamma_2)$ ***generalize*** to $(\mathcal{T}_3, \gamma_3)$ if any reward consistent with the experts in Env 1 and Env 2 yields an optimal expert in Env 3.

2 Main results

Theorem 1 (Adapted from [1]). *For all domains $d := (\mathcal{X}, \mathcal{A}, \mathcal{T}, \mathcal{T}_0, \gamma)$ with deterministic transition $\mathcal{T}(s' \mid s, a) \in \{0, 1\}$ and initial state $|supp(\mathcal{T}_0)| = 1$, the $\mathcal{P}_{MDP}[R; d, T, J_{MaxEnt}]$ is weakly identifiable.*

Theorem 2 (Strong Identification Criteria [1]). *For all (d, r, T, J) such that G_d is* strongly connected,

- *(Sufficiency) $\mathcal{P}_{MDP}[R; d; T; J]$ is* weakly ID, *G_d is T_0-coverable, and $T \geq 2T_0 \Rightarrow \mathcal{P}_{MDP}[R; d; T; J]$ is strongly ID.*
- *(Necessity) $\mathcal{P}_{MDP}[R; d; T; J]$ is strongly ID $\Rightarrow \mathcal{P}_{MDP}[R; d, T, J]$ is weakly ID, G_d is coverable.*

Corollary 1 (Adapted from [2]). *Suppose the MDP is stochastic, and satisfies one of the assumptions of Corollary in [2] (easy to check). If*

$$\operatorname{rank}\{\mathcal{T}(\cdot \mid s, a) : a \in \mathcal{A}\} = \#\{s' : \mathcal{T}(s' \mid s, a) > 0 \text{ for some } a \in \mathcal{A}\}$$

then for any initial s_0, there exists horizon T such that the IRL problem is strongly identifiable.

Theorem 3 (Action Independent rewards [2]). *The IRL problem admits a solution with action-independent reward $f : \mathcal{S} \to \mathbb{R}$ iff*

$$\lambda(\log \pi(a) - \log \pi(a_0)) = \gamma(\mathcal{T}(s_j \mid s_i, a) - \mathcal{T}(s_j \mid s_i, a_0))v, \quad \forall a \in \mathcal{A}$$

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

Two Environments with $\gamma_1 \neq \gamma_2$, identical $\mathcal{S}$ and $\mathcal{A}$

$\mathcal{T}_1 \neq \mathcal{T}_2$ → **Theorem 3** [3]:

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

$\mathcal{T}_1 = \mathcal{T}_2$ → **Corollary 5** [3]:

$$\operatorname{rank}\begin{pmatrix} \mathcal{T}_{a_1} - \mathcal{T}_{a_2} \\ \vdots \\ \mathcal{T}_{a_1} - \mathcal{T}_{a_{|\mathcal{A}|}} \end{pmatrix} = |\mathcal{S}| - 1$$

Theorem 4 (Generalizability [3]). *$(\mathcal{T}^1, \gamma_1)$, $(\mathcal{T}^2, \gamma_2)$ generalize to $(\mathcal{T}^3, \gamma_3)$ if and only if*

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

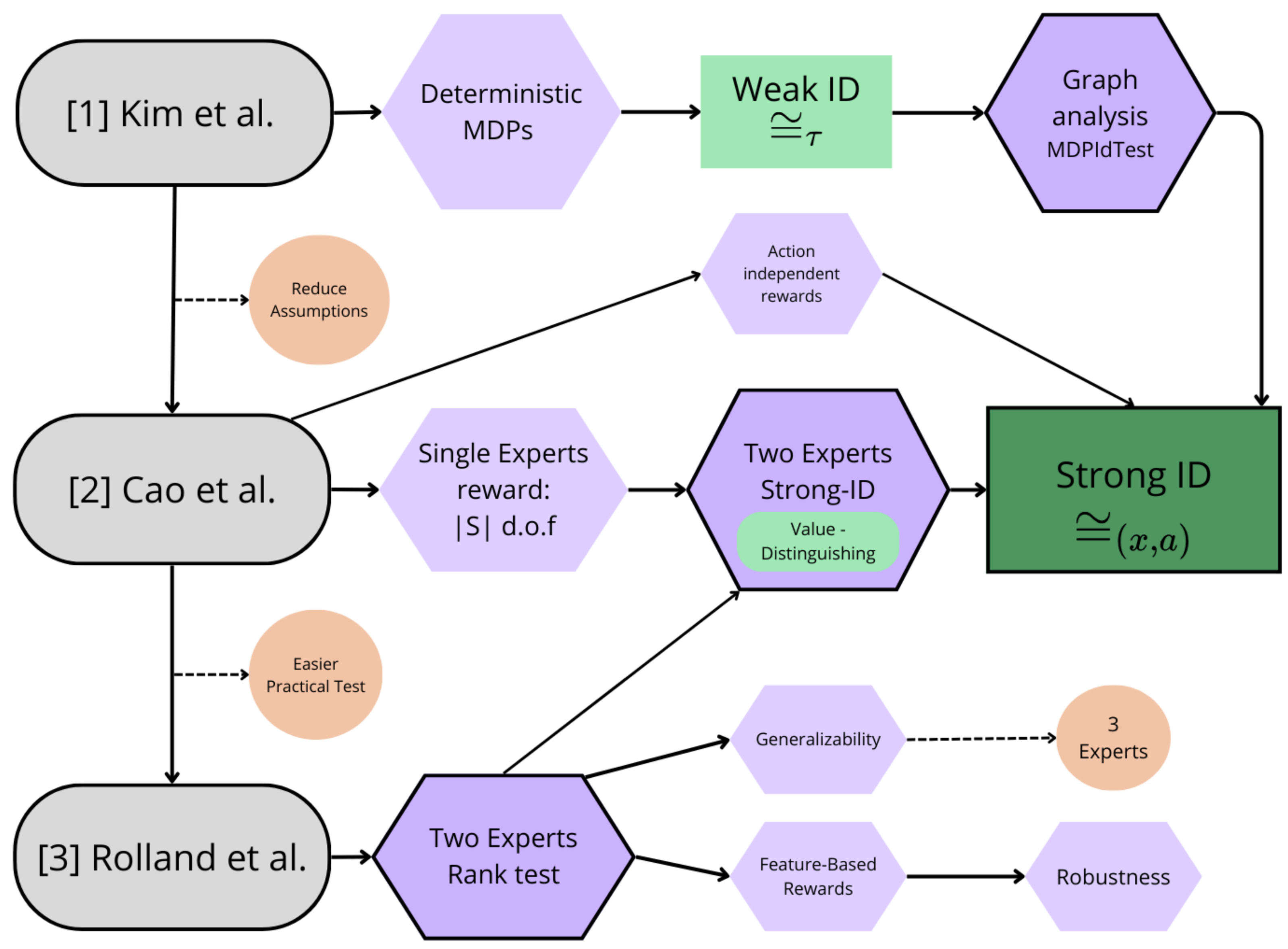
Summary

1. Solved MDP-s: multiple experts; deterministic; action independent reward;
2. General stochastic MDP: some sufficiency conditions.

3 Comparative analysis

All three papers [1], [2], and [3] try to classify all MDP-s for which the problem with inverse reward is well-posed.

- [1] is a more fundamental paper. It introduces the theory on weak and strong identifiabilities;
- [2] starts the theory on multiple agents, finds the criteria for action independent reward MDP-s, introduces new sufficient conditions for stochastic MDP-s;
- [3] focuses on multiple agents case, finishing [2]'s theory.



4 Open Questions

Open problem 1. *Find more families of weakly identifiable MDP-s.*

Open problem 2. *Verify if the results from [1], [2], and [3] can be generalized for continuous state and actions spaces.*

Open problem 3. *What identifiability and generalizability guarantees can be obtained under partial observability (POMDP) assumption?*

References

[1] K. Kim, S. Garg, K. Shiragur, and S. Ermon, *Reward identification in inverse reinforcement learning.* International Conference on Machine Learning, 2021.

[2] H. Cao, S. Cohen, and L. Szpruch, *Identifiability in inverse reinforcement learning (IRL).* Advances in Neural Information Processing Systems, 2021.

[3] P. Rolland, L. Viano, N. Schürhoff, B. Nikolov, and V. Cevher. *Identifiability and generalizability from multiple experts in inverse reinforcement learning.* Advances in Neural Information Processing Systems, 2022.