Task-Guided Inverse Reinforcement Learning under Partial Information

TEXAS — The University of Texas at Austin

Franck Djeumou, Murat Cubuktepe, Craig Lennon, and Ufuk Topcu

autonomous SYSTEMS GROUP

IRL under partial information

Inverse reinforcement learning (IRL) with a learner acting under partial information

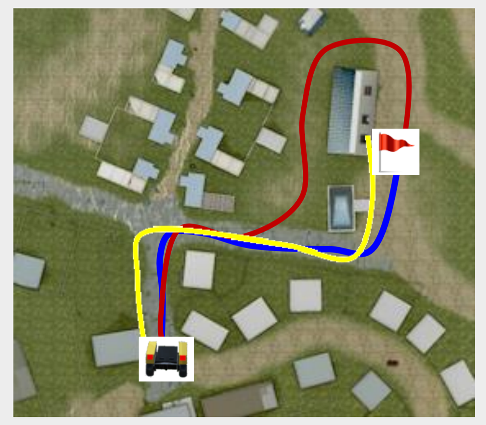
Reward-free POMDP

$\mathcal{M} = (S, A, P, Z, O)$

Inverse reinforcement learning

Reward function R

Policy that induces a similar behavior to the expert's



Expert demonstrations (observation-action pairs)

$\mathcal{D} = (z_1, \alpha_1, z_2, \alpha_2, \dots)$

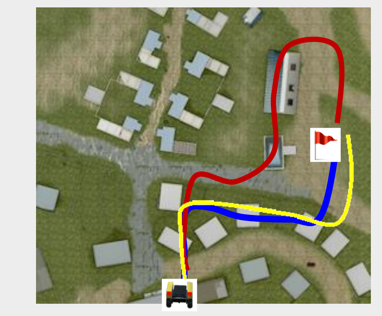
Challenges in IRL under partial information

1. IRL is an ill-defined problem, many reward functions can induce the same behavior

Solution: Use **(causal)** entropy to randomize while acting similar to demonstrations (causal entropy only depends on past observations and not the future)

2. Information asymmetry between the expert and the learning agent

- The agent may not obtain behavior similar to the expert's, even with known reward function



Expert's view

Learner's view

3. Each step of IRL requires to solve a policy synthesis problem on the POMDP

- Computationally intractable: Nonconvex optimization problem
- Optimal policy may require infinite memories

Solution approach

Key idea: Task knowledge as temporal logic specification alleviates information asymmetry

Given: $\mathcal{M}$ (POMDP), φ (specification), $\mathcal{D}$ (expert demonstrations) and ψ (feature functions),

Learn: σ (policy), θ (reward parameter), and $R = \psi(\theta)$ (unknown reward function) such that

maximize $H_\sigma \triangleq \mathbb{E}_\sigma[-\log \sigma_{z,\alpha}]$ — maximize the causal entropy of the policy

subject to $\mathcal{M}_\sigma \models \varphi$ — policy satisfies specification (side information)

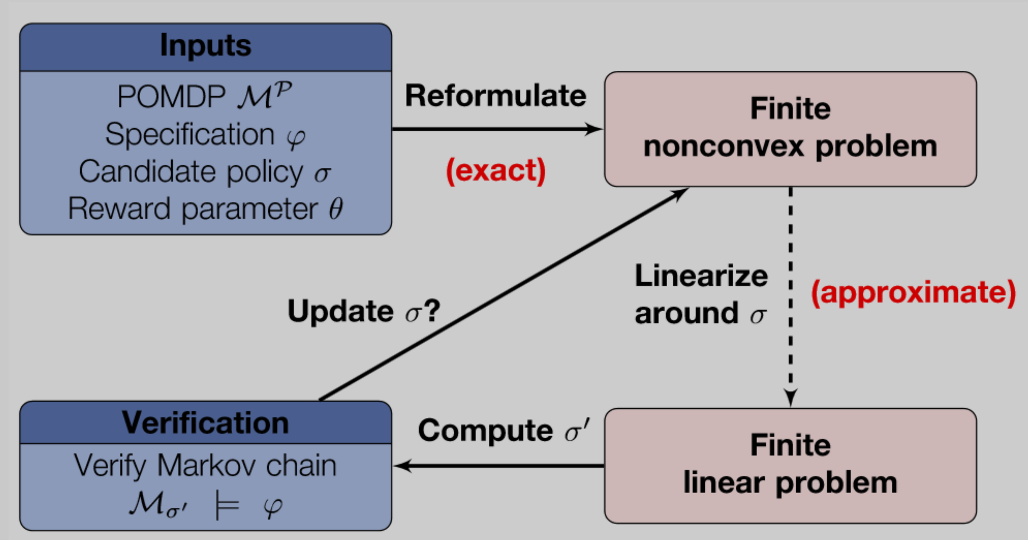
$\mathbb{E}_\sigma[\psi] = \mathbb{E}_\mathcal{D}[\psi]$ — obtain a similar behavior to the expert's by matching features

SCPForward: Scalable policy synthesis for POMDPs

Inverse reinforcement learning as a two-player game

$$L(\theta,\sigma) \triangleq \min_\theta \boxed{\max_\sigma \; H_\sigma + (\mathbb{E}_\sigma[\psi(\theta)]} - \mathbb{E}_\mathcal{D}[\psi(\theta)])$$

causal-entropy-regularized POMDP synthesis problem for fixed θ



If σ' improves over σ and $\mathcal{M}_{\sigma'} \models \varphi$: update σ', enlarge trust region

Else: keep σ, shrink trust region

Theorem: Our algorithm provides sound and locally optimal solutions for the policy synthesis problem

Gradient descent for learning the reward parameter

The learning problem can be formulated as finding a saddle point to

$$L(\theta,\sigma) \triangleq \boxed{\min_\theta \max_\sigma \; H_\sigma + (\mathbb{E}_\sigma[\psi(\theta)] - \mathbb{E}_\mathcal{D}[\psi(\theta)])}$$

learning problem for fixed σ

$$\nabla_\theta L(\theta,\sigma) = \sum_{(z,\alpha)\in Z\times A} \mathbb{P}(z,\alpha|\sigma)\; \nabla_\theta \psi_\theta(z,\alpha) - \frac{1}{|\mathcal{D}|} \sum_{(z,\alpha)\in\mathcal{D}} \nabla_\theta \psi_\theta(z,\alpha)$$

(gradient with respect to θ; probability of (z,α) under policy σ; gradient of ψ with respect to θ)

Approach: iterate between σ and θ until $|\mathbb{E}_\sigma[\psi(\theta)] - \mathbb{E}_\mathcal{D}[\psi(\theta)]| \leq \epsilon$

Experiments

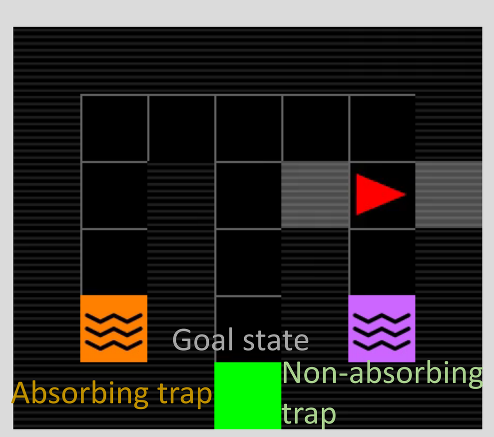
An example: Robot navigation in a maze

A robot navigates in a maze to reach the exit

POMDP: Partial observability over the location in the maze

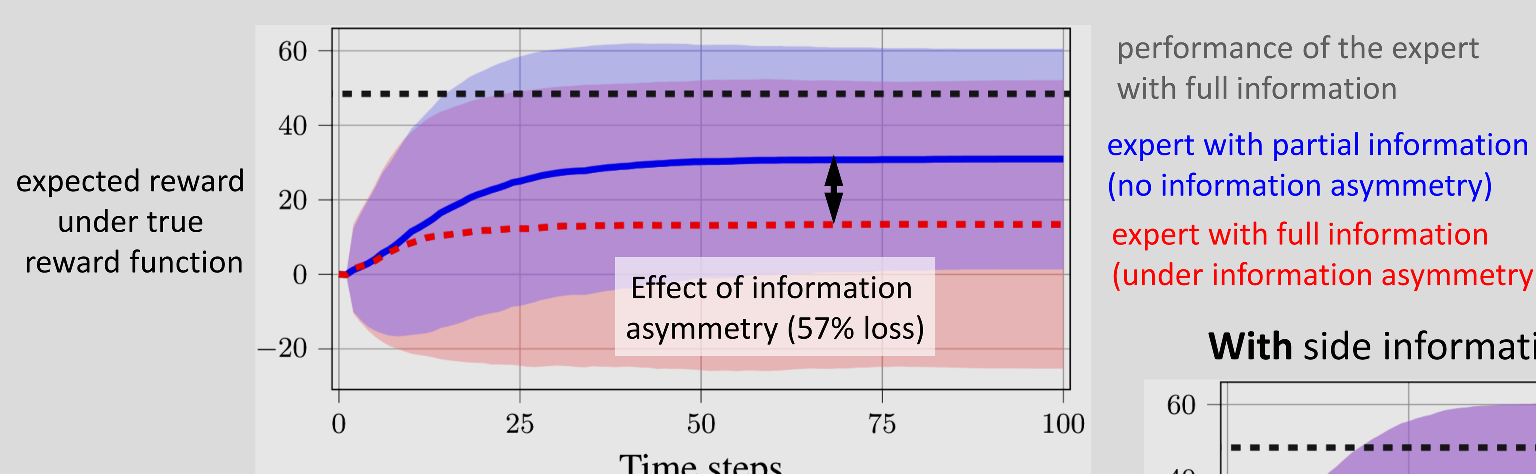
Specification φ (side information): Avoid trap states

Feature functions: Positive reward for reach the exit, Negative reward for each action and being on trap states



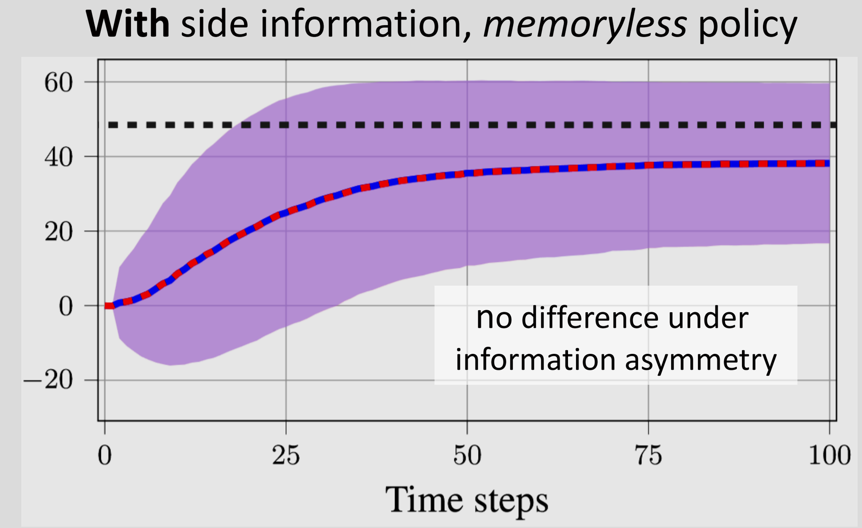
Side information alleviates the information asymmetry

Without side information, *memoryless* policy

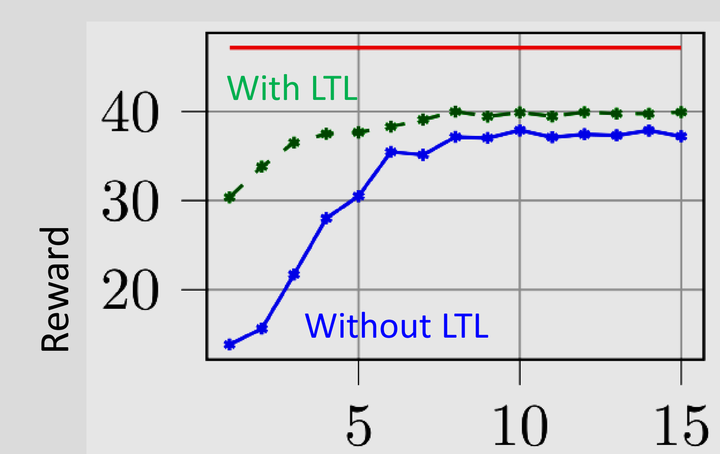


Performance decreases under information asymmetry

With side information, *memoryless* policy



No performance decrease under information asymmetry



Side information improves data efficiency

SCPForward is at least two orders of magnitude faster than existing POMDP solvers

Problem	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{S}\times\mathcal{O}\rvert$	$\lvert\mathcal{O}\rvert$	SCPForward R^θ_σ	SCPForward Time (s)	SARSOP R^θ_σ	SARSOP Time (s)
Maze	17	162	11	39.24	**0.1**	**47.83**	0.24
Maze (3-FSC)	49	777	31	44.98	**0.6**	NA	NA
Maze (10-FSC)	161	2891	101	46.32	2.04	NA	NA
Obstacle[10]	102	1126	5	19.71	8.79	**19.8**	**0.02**
Obstacle[10](5-FSC)	679	7545	31	19.77	38	NA	NA
Obstacle[25]	627	7306	5	19.59	14.22	**19.8**	**0.1**
Rock	550	4643	67	19.68	12.2	**19.83**	**0.05**
Rock (3-FSC)	1648	23203	199	19.8	15.25	NA	NA
Rock (5-FSC)	2746	41759	331	19.82	97.84	NA	NA
Intercept[5, 2, 0]	1321	5021	1025	**19.83**	**10.28**	**19.83**	13.71
Intercept[5, 2, 0.1]	1321	7041	1025	**19.81**	**13.18**	**19.81**	81.19
Evade[5, 2, 0]	2081	13561	1089	**97.3**	**26.25**	**97.3**	3600
Evade[5, 2, 0.1]	2081	16761	1089	**96.79**	**26.25**	95.28	3600
Evade[10, 2, 0]	36361	341121	18383	**94.97**	**3600**	–	–
Avoid[4, 2, 0]	2241	5697	1956	**9.86**	34.74	**9.86**	**9.19**
Avoid[4, 2, 0.1]	2241	8833	1956	**9.86**	**14.63**	**9.86**	210.47
Avoid[7, 2, 0]	19797	62133	3164	**9.72**	**3503**	–	–