How does Inverse RL Scale to Large State Spaces? A Provably Efficient Approach

F. Lazzati, M. Mutti, A. M. Metelli

38th Conference on Neural Information Processing Systems (NeurIPS 2024)

Inverse Reinforcement Learning (IRL)

Introduction

- **IRL**: Given $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ and π^E, find *"the"* reward r that makes π^E optimal

Inverse Reinforcement Learning (IRL)

Introduction

- **IRL**: Given $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ and π^E, find *"the"* reward r that makes π^E optimal
- *Ill-posedness*: many reward functions make π^E optimal

Inverse Reinforcement Learning (IRL)

Introduction

- **IRL**: Given $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ and π^E, find *"the"* reward r that makes π^E optimal

- *Ill-posedness*: many reward functions make π^E optimal

- Feasible Set: $\mathcal{R}_{p,\pi^E} := \{r : J^*(r; p) = J^{\pi^E}(r; p)\}$

Learning Setting

Introduction

- p, π^E *unknown*

Learning Setting

Introduction

- p, π^E *unknown*
- τ trajectories with forward model for $p \to \widehat{p}$

Learning Setting

Introduction

- p, π^E *unknown*
- τ trajectories with forward model for $p \rightarrow \widehat{p}$
- τ^E trajectories in batch dataset for $\pi^E \rightarrow \widehat{\pi}^E$

Learning Setting

Introduction

- p, π^E *unknown*
- τ trajectories with forward model for $p \rightarrow \widehat{p}$
- τ^E trajectories in batch dataset for $\pi^E \rightarrow \widehat{\pi}^E$
- $\widehat{p}, \widehat{\pi}^E \rightarrow \widehat{\mathcal{R}}$

Learning Setting

Introduction

- p, π^E *unknown*
- τ trajectories with forward model for $p \to \widehat{p}$
- τ^E trajectories in batch dataset for $\pi^E \to \widehat{\pi}^E$
- $\widehat{p}, \widehat{\pi}^E \to \widehat{\mathcal{R}}$
- Previous works analyse how many τ, τ^E are needed to obtain $\widehat{\mathcal{R}} \approx \mathcal{R}_{p,\pi^E}$ in the <u>tabular setting</u>

What about Linear MDPs?

π^E known

Limitations of the Feasible Set

Theorem

Let $\underline{\pi^E \text{ known}}$. Then, we can design an algorithm such that

$$\mathcal{H}\big(\widehat{\mathcal{R}}, \mathcal{R}_{p,\pi^E}\big) \leq \epsilon \quad \text{w.p. } 1-\delta,$$

with a number of exploration episodes:

$$\tau \leq \widetilde{\mathcal{O}}\left(\frac{H^5 d}{\epsilon^2}\left(d + \log\frac{1}{\delta}\right)\right).$$

π^E unknown

Limitations of the Feasible Set

Theorem

Let π^E unknown. Assume to have access to a *generative model* for π^E. Then, any algorithm must collect at least

$$\tau^E \geq \Omega(S)$$

samples to obtain

$$\mathcal{H}\left(\widehat{\mathcal{R}}, \mathcal{R}_{p,\pi^E}\right) \leq \epsilon \quad \text{w.p. } 1-\delta.$$

The feasible set cannot be learned efficiently in Linear MDPs!

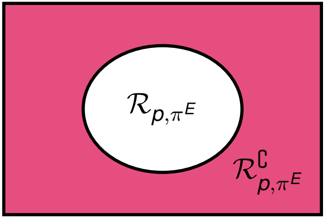
Rewards Compatibility

A New Framework

- The *feasible set*

  $\mathcal{R}_{p,\pi^E} := \{r : J^*(r;p) = J^{\pi^E}(r;p)\}$

  binary classifies rewards



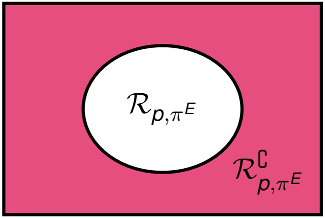
Rewards Compatibility

A New Framework

- The *feasible set*

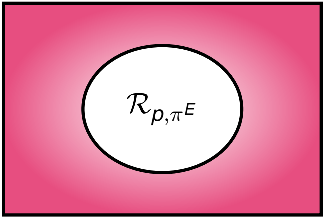
  $\mathcal{R}_{p,\pi^E} \coloneqq \{r : J^*(r;p) = J^{\pi^E}(r;p)\}$

  binary classifies rewards



- Some rewards are *more "compatible"* than others:

  $\overline{\mathcal{C}}_{p,\pi^E}(r) \coloneqq J^*(r;p) - J^{\pi^E}(r;p)$



IRL Classification Formulation

A New Framework

- IRL Classification Problem: $(\mathcal{M}, \pi^E, \mathcal{R}, \Delta)$

  $\forall r \in \mathcal{R}$: **if** $\overline{\mathcal{C}}_{p,\pi^E}(r) \leq \Delta$ **then** **return** True, **else return** False.

- IRL Classification Algorithm: **Input:** $r \in \mathcal{R}$, **output:** boolean.

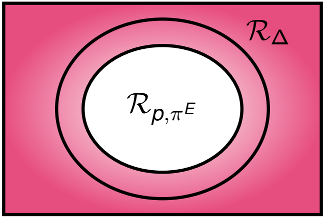
IRL Classification Formulation

A New Framework

- IRL Classification Problem: $(\mathcal{M}, \pi^E, \mathcal{R}, \Delta)$

  $\forall r \in \mathcal{R}$: **if** $\overline{\mathcal{C}}_{p,\pi^E}(r) \leq \Delta$ **then** **return** True, **else return** False.

- IRL Classification Algorithm: **Input:** $r \in \mathcal{R}$, **output:** boolean.



Learning Setting

A New Framework

- p, π^E *unknown*
- τ trajectories with *forward model* for $p \to \widehat{p}$
- τ^E trajectories in *batch dataset* for $\pi^E \to \widehat{\pi}^E$

Learning Setting

A New Framework

- p, π^E *unknown*
- τ trajectories with *forward model* for $p \to \widehat{p}$
- τ^E trajectories in *batch dataset* for $\pi^E \to \widehat{\pi}^E$

PAC Algorithm: Let $\epsilon, \delta \in (0, 1)$. An algorithm $\mathfrak{A}$ is (ϵ, δ)-**PAC** for the *IRL classification problem* if:

$$\sup_{r \in \mathcal{R}} \left| \overline{\mathcal{C}}_{p,\pi^E}(r) - \widehat{\mathcal{C}}(r) \right| \leq \epsilon \quad \text{w.p. } 1 - \delta.$$

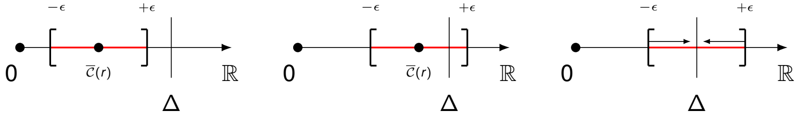
Learning Setting

A New Framework

- p, π^E *unknown*
- τ trajectories with *forward model* for $p \to \widehat{p}$
- τ^E trajectories in *batch dataset* for $\pi^E \to \widehat{\pi}^E$

PAC Algorithm: Let $\epsilon, \delta \in (0,1)$. An algorithm $\mathfrak{A}$ is (ϵ, δ)-**PAC** for the *IRL classification problem* if:

$$\sup_{r \in \mathcal{R}} \left| \overline{\mathcal{C}}_{p,\pi^E}(r) - \widehat{\mathcal{C}}(r) \right| \leq \epsilon \quad \text{w.p. } 1-\delta.$$



The Algorithm

CATY-IRL

`CATY-IRL` (`C`omp`AT`ibilit`Y` for `IRL`) is made of two phases:

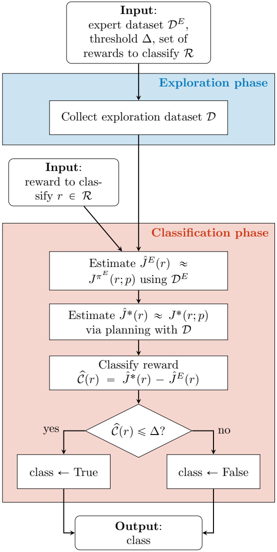
- Exploration phase
- Classification phase

The Algorithm

CATY-IRL

`CATY-IRL` (`C`omp`AT`ibilit`Y` for `IRL`) is made of two phases:

- Exploration phase
- Classification phase



Sample Complexity Analysis

CATY-IRL

Theorem

In tabular MDPs, `CATY-IRL` executed with RF-Express (Menard et al., 2021) is (ϵ, δ)-PAC with a sample complexity:

$$\tau^E \leq \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\epsilon^2}\log\frac{1}{\delta}\Big), \qquad \tau \leq \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\epsilon^2}\Big(S+\log\frac{1}{\delta}\Big)\Big).$$

Sample Complexity Analysis

CATY-IRL

Theorem

In tabular MDPs, `CATY-IRL` executed with RF-Express (Menard et al., 2021) is (ϵ,δ)-PAC with a sample complexity:

$$\tau^E \leq \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\epsilon^2}\log\frac{1}{\delta}\Big), \qquad \tau \leq \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\epsilon^2}\Big(S+\log\frac{1}{\delta}\Big)\Big).$$

Theorem

In linear MDPs, `CATY-IRL` executed with RFLin (Wagenmaker et al., 2022) is (ϵ,δ)-PAC with a sample complexity:

$$\tau^E \leq \widetilde{\mathcal{O}}\Big(\frac{H^3d}{\epsilon^2}\log\frac{1}{\delta}\Big), \qquad \tau \leq \widetilde{\mathcal{O}}\Big(\frac{H^5d}{\epsilon^2}\Big(d+\log\frac{1}{\delta}\Big)\Big).$$

Theoretical Limits of IRL and RFE

Statistical Barriers

Theorem

IRL Classification and RFE enjoy the same lower bound to the sample complexity in the *tabular* setting, which is matched, respectively, by `CATY-IRL` and RF-Express (Menard et al., 2021):

$$\tau \geq \Omega\left(\frac{H^3 SA}{\epsilon^2}\left(S + \log\frac{1}{\delta}\right)\right).$$

Theoretical Limits of IRL and RFE

Statistical Barriers

Theorem

IRL Classification and RFE enjoy the same lower bound to the sample complexity in the *tabular* setting, which is matched, respectively, by `CATY-IRL` and RF-Express (Menard et al., 2021):

$$\tau \geq \Omega\left(\frac{H^3 SA}{\epsilon^2}\left(S + \log\frac{1}{\delta}\right)\right).$$

This improves over the state-of-the-art lower bound of RFE by one H factor (Jin et al., 2020).

Objective-Free Exploration (OFE)

A Unifying Exploration Framework

*What is the **most efficient** exploration strategy that can be performed in an unknown environment?*

Objective-Free Exploration (OFE)

A Unifying Exploration Framework

*What is the **most efficient** exploration strategy that can be performed in an unknown environment?*

It *depends* on the subsequent task that shall be solved!

Objective-Free Exploration (OFE)

A Unifying Exploration Framework

*What is the **most efficient** exploration strategy that can be performed in an unknown environment?*

It *depends* on the subsequent task that shall be solved!

Definition

Given a tuple $(\mathcal{M}, \mathscr{F}, (\epsilon, \delta))$, where $\mathcal{M}$ is an *unknown* environment and $\mathscr{F}$ is a certain class of tasks, the Objective-Free Exploration (OFE) problem aims to find an exploration strategy of the environment $\mathcal{M}$ that permits to solve *any* task $f \in \mathscr{F}$ in an (ϵ, δ)-correct manner.

Summary of Contributions

Conclusion

- Non-learnability of the feasible set in Linear MDPs

Summary of Contributions

Conclusion

- Non-learnability of the feasible set in Linear MDPs
- *Rewards compatibility*

Summary of Contributions

Conclusion

- Non-learnability of the feasible set in Linear MDPs
- *Rewards compatibility*
- `CATY-IRL`, an efficient algorithm for IRL classification

Summary of Contributions

Conclusion

- Non-learnability of the feasible set in Linear MDPs
- *Rewards compatibility*
- `CATY-IRL`, an efficient algorithm for IRL classification
- Matching *lower bound* for the tabular setting

Summary of Contributions

Conclusion

- Non-learnability of the feasible set in Linear MDPs
- *Rewards compatibility*
- `CATY-IRL`, an efficient algorithm for IRL classification
- Matching *lower bound* for the tabular setting
- *Objective-free exploration* (OFE)