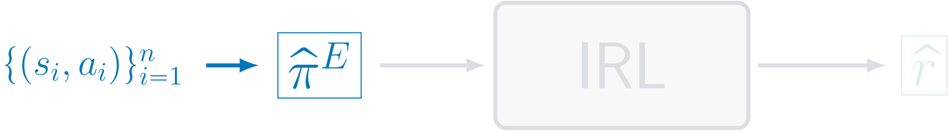
Provably Efficient Learning of Transferable Rewards

Alberto Maria Metelli* Giorgia Ramponi* Alessandro Concetti Marcello Restelli

July 2021
Thirty-eighth International Conference on Machine Learning

Inverse Reinforcement Learning

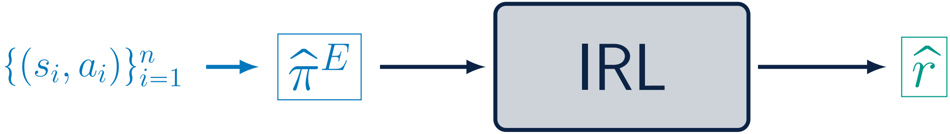
- **Goal**: learn *one* **reward** function from **expert**'s demonstrations (Ng and Russell, 2000)



- IRL problem is **ambiguous** (Abbeel and Ng, 2004)!
- **Feasible Reward Set**: set of all rewards making π^E optimal (Ng and Russell, 2000)

$$\mathcal{R} = \left\{ r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}} : \pi^E \in \text{Greedy}\left(Q^*_{\mathcal{M} \cup r}\right) \right\}$$

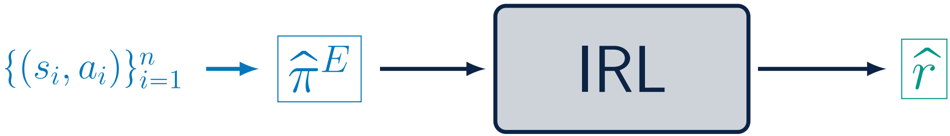
- **Goal**: learn *one* **reward** function from **expert**'s demonstrations (Ng and Russell, 2000)



- IRL problem is **ambiguous** (Abbeel and Ng, 2004)!
- **Feasible Reward Set**: set of all rewards making π^E optimal (Ng and Russell, 2000)

$$\mathcal{R} = \left\{ r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}} : \pi^E \in \text{Greedy}\left(Q^*_{\mathcal{M} \cup r}\right) \right\}$$

- **Goal**: learn *one* **reward** function from **expert**'s demonstrations (Ng and Russell, 2000)

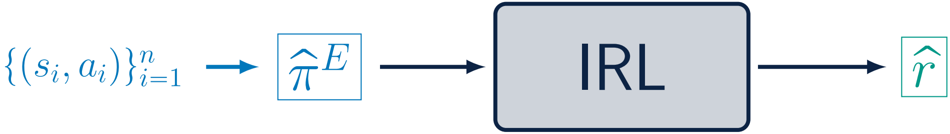


- IRL problem is **ambiguous** (Abbeel and Ng, 2004)!
- **Feasible Reward Set**: set of all rewards making π^E optimal (Ng and Russell, 2000)

$$\mathcal{R} = \left\{ r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}} : \pi^E \in \text{Greedy}\left(Q^*_{\mathcal{M} \cup r}\right) \right\}$$

Inverse Reinforcement Learning

- **Goal**: learn *one* **reward** function from **expert**'s demonstrations (Ng and Russell, 2000)



- IRL problem is **ambiguous** (Abbeel and Ng, 2004)!
- **Feasible Reward Set**: set of all rewards making π^E optimal (Ng and Russell, 2000)

$$\mathcal{R} = \left\{ r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}} : \pi^E \in \text{Greedy}\left(Q^*_{\mathcal{M} \cup r}\right) \right\}$$

Questions

(Q1) *How does the error on the* **transition model** $\widehat{P}$ *and on the* **expert's policy** $\widehat{\pi}^E$ *propagate to the* **recovered reward** $\widehat{r} \in \widehat{\mathcal{R}}$ *?*

(Q2) *How does the error on the* **recovered reward** $\widehat{r} \in \widehat{\mathcal{R}}$ *affect the* **value function** $Q^*_{\mathcal{M}' \cup \widehat{r}}$ *in a* **different environment** $\mathcal{M}'$*?*

Questions

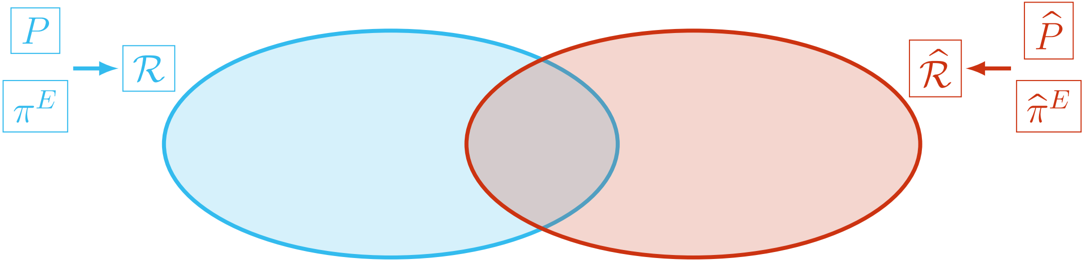
(Q1) *How does the error on the* **transition model** $\widehat{P}$ *and on the* **expert's policy** $\widehat{\pi}^E$ *propagate to the* **recovered reward** $\widehat{r} \in \widehat{\mathcal{R}}$ *?*

(Q2) *How does the error on the* **recovered reward** $\widehat{r} \in \widehat{\mathcal{R}}$ *affect the* **value function** $Q^*_{\mathcal{M}' \cup \widehat{r}}$ *in a* **different environment** $\mathcal{M}'$*?*

(Q1) Reward Error Propagation

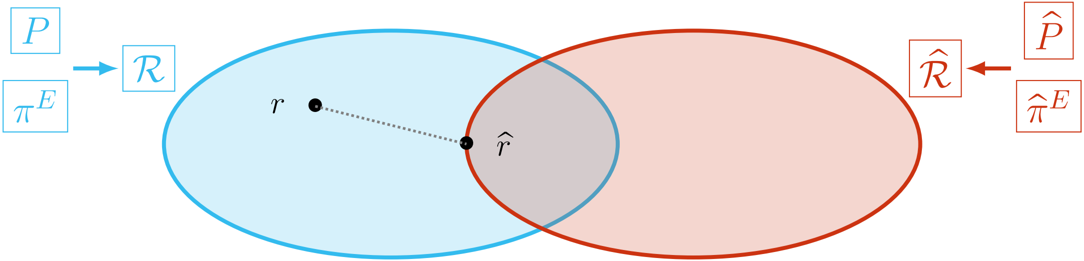
(Q1) *How does the error on the* **transition model** $\widehat{P}$ *and on the* **expert's policy** $\widehat{\pi}^E$ *propagate to the* **recovered reward** $\widehat{r} \in \widehat{\mathcal{R}}$ *?*

$$\inf_{\widehat{r} \in \widehat{\mathcal{R}}} |r - \widehat{r}| \leqslant \frac{R_{\max}}{1-\gamma} \text{Distance}(\pi^E, \widehat{\pi}^E) + \frac{\gamma R_{\max}}{1-\gamma} \text{Distance}\left(P, \widehat{P}\right)$$



(Q1) *How does the error on the* **transition model** $\widehat{P}$ *and on the* **expert's policy** $\widehat{\pi}^E$ *propagate to the* **recovered reward** $\widehat{r} \in \widehat{\mathcal{R}}$ *?*

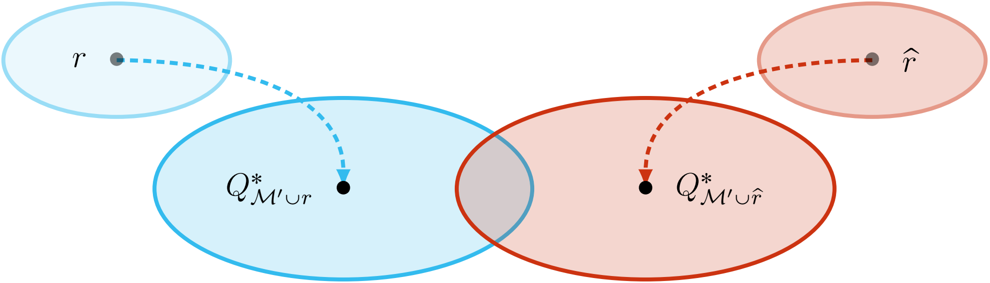
$$\inf_{\widehat{r} \in \widehat{\mathcal{R}}} |r - \widehat{r}| \leqslant \frac{R_{\max}}{1-\gamma} \text{Distance}(\pi^E, \widehat{\pi}^E) + \frac{\gamma R_{\max}}{1-\gamma} \text{Distance}\left(P, \widehat{P}\right)$$



(Q2) Transferred Reward Error Propagation

(Q2) *How does the error on the* **recovered reward** $\widehat{r} \in \widehat{\mathcal{R}}$ *affect the* **value function** $Q^*_{\mathcal{M}' \cup \widehat{r}}$ *in a* **different environment** $\mathcal{M}'$*?*

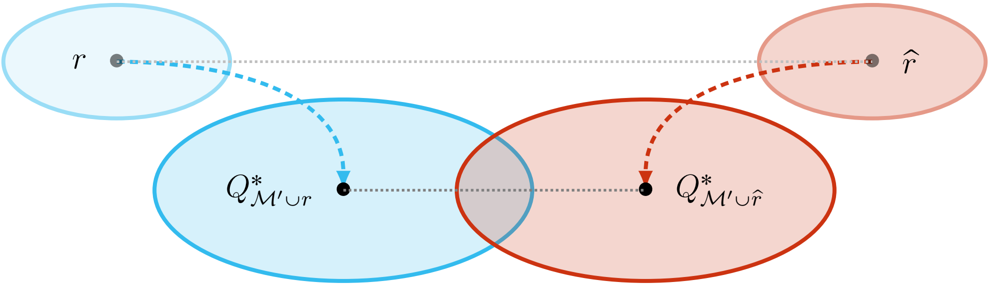
$$\left\| Q^*_{\mathcal{M}' \cup r} - Q^*_{\mathcal{M}' \cup \widehat{r}} \right\|_\infty \leqslant \max_{\pi \in \{\pi^*, \widehat{\pi}^*\}} \Big\| \underbrace{\left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P' \pi\right)^{-1}}_{\text{occupancy in } \mathcal{M}'} (r - \widehat{r}) \Big\|_\infty$$



(Q2) Transferred Reward Error Propagation

(Q2) *How does the error on the* **recovered reward** $\widehat{r} \in \widehat{\mathcal{R}}$ *affect the* **value function** $Q^*_{\mathcal{M}' \cup \widehat{r}}$ *in a* **different environment** $\mathcal{M}'$*?*

$$\left\| Q^*_{\mathcal{M}' \cup r} - Q^*_{\mathcal{M}' \cup \widehat{r}} \right\|_\infty \leqslant \max_{\pi \in \{\pi^*, \widehat{\pi}^*\}} \left\| \underbrace{\left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P' \pi\right)^{-1}}_{\text{occupancy in } \mathcal{M}'} (r - \widehat{r}) \right\|_\infty$$



- Problem Setting
  - **Generative model** of $\mathcal{M} = (P, \gamma)$ and possibility of sampling from π^E
  - Target MDP $\mathcal{M}' = (P', \gamma')$ **known**
- (ϵ, δ, n)-correct **Sampling Strategy**
  - Fix target rewards $(\overline{r}, \check{r}) \in \mathcal{R} \times \widehat{\mathcal{R}}$
  - After having collected n samples

$$\inf_{\widehat{r} \in \widehat{\mathcal{R}}} \| Q^*_{\mathcal{M}' \cup \overline{r}} - Q^*_{\mathcal{M}' \cup \widehat{r}} \|_\infty \leqslant \epsilon \quad \text{and} \quad \inf_{r \in \mathcal{R}} \| Q^*_{\mathcal{M}' \cup r} - Q^*_{\mathcal{M}' \cup \check{r}} \|_\infty \leqslant \epsilon \qquad \text{w.p. } 1 - \delta$$

- Problem Setting
  - **Generative model** of $\mathcal{M} = (P, \gamma)$ and possibility of sampling from π^E
  - Target MDP $\mathcal{M}' = (P', \gamma')$ **known**
- (ϵ, δ, n)-correct **Sampling Strategy**
  - Fix target rewards $(\overline{r}, \check{r}) \in \mathcal{R} \times \widehat{\mathcal{R}}$
  - After having collected n samples

$$\inf_{\widehat{r} \in \widehat{\mathcal{R}}} \| Q^*_{\mathcal{M}' \cup \overline{r}} - Q^*_{\mathcal{M}' \cup \widehat{r}} \|_\infty \leqslant \epsilon \quad \text{and} \quad \inf_{r \in \mathcal{R}} \| Q^*_{\mathcal{M}' \cup r} - Q^*_{\mathcal{M}' \cup \check{r}} \|_\infty \leqslant \epsilon \qquad \text{w.p. } 1 - \delta$$

Uniform Sampling

- Allocate samples **uniformly** over states and actions
- (ϵ, δ, n)-correct with $n = \sum_{s,a} n(s,a)$:

$$n(s,a) \leqslant \widetilde{\mathcal{O}} \left(\frac{\gamma^2 R_{\max}^2}{(1-\gamma')^2 (1-\gamma)^2 \epsilon^2} \right)$$

Uniform Sampling

- Allocate samples **uniformly** over states and actions
- (ϵ, δ, n)-correct with $n = \sum_{s,a} n(s,a)$:

$$n(s,a) \leqslant \widetilde{\mathcal{O}}\left(\frac{\gamma^2 R_{\max}^2}{(1-\gamma')^2(1-\gamma)^2\epsilon^2}\right)$$

- Allocate samples based on the **error propagation bound**
- (ϵ, δ, n)-correct with $n = \sum_{s,a} n(s,a)$:

- Allocate samples based on the **error propagation bound**
- (ϵ, δ, n)-correct with $n = \sum_{s,a} n(s,a)$:

$$n(s,a) \leqslant \widetilde{\mathcal{O}}\left(\min\left\{\frac{\gamma^2 R_{\max}^2}{(1-\gamma')^2(1-\gamma)^2\epsilon^2}, \frac{\gamma^2 R_{\max}^2 (\epsilon')^2}{(1-\gamma)^2 \boldsymbol{\Delta}(\boldsymbol{s},\boldsymbol{a})^{\mathbf{2}}\epsilon^2}\right\}\right)$$

$$\boldsymbol{\Delta}(\boldsymbol{s},\boldsymbol{a}) = V_{\widetilde{r}}^*(s) - Q_{\widetilde{r}}^*(s,a)$$

Thank You for Your Attention!

Code: `github.com/albertometelli/travel`

Contact: `albertomaria.metelli@polimi.it`

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In C. E. Brodley, editor, *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004. doi: 10.1145/1015330.1015430.

A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In P. Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 663–670. Morgan Kaufmann, 2000.