



MOTIVATION AND CONTRIBUTIONS

RESEARCH QUESTIONS

- Q1** When is partial identifiability *problematic*?
Q2 What is the “best” approach to cope with partial identifiability in Reward Learning (ReL) problems?
Q3 Can we apply it in a provably efficient manner?

CONTRIBUTIONS

- Novel **quantitative** framework for ReL
- **Robust** approach to tackle partial identifiability
- Provably efficient robust algorithm for ReL

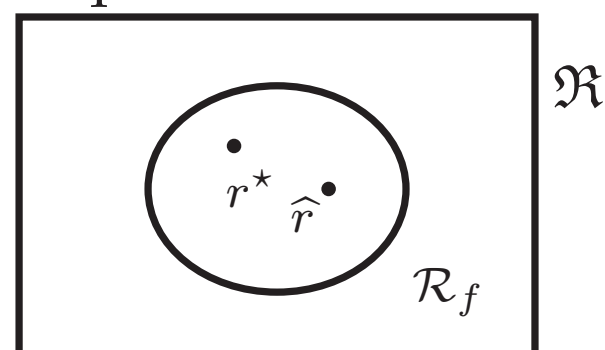
PRELIMINARIES

REL AND PARTIAL IDENTIFIABILITY

- Target reward r^* → unknown, to learn
- Feedback f → gives information on r^*
- Application g → for which we need r^*
- Partial identifiability → f doesn't *uniquely* identify r^*

EXAMPLE

- f = demonstrations $\mathcal{D}^E = \{(s_h, a_h)\} \sim \pi^E$ (IRL)
- f' = preferences $\mathcal{D} = \{(\omega_1, \omega_2, \sigma)\}$ (RLHF)
- g = find an optimal policy under r^*
- $\mathcal{R}_f$ = rewards equivalent to r^* under f



PREVIOUS FRAMEWORKS

Q1 When is partial identifiability problematic?
JEON ET AL. (2020)

- Don't consider applications
- No criterion

SKALSE ET AL. (2023)

- Qualitative model for applications
- *Criterion*: f suffices if all the rewards in $\mathcal{R}_f$ have the same meaning as r^* for g (e.g., same optimal policies)
- **Too restrictive**: what if the meaning is close but different? (e.g., if $\hat{r}$ induces an ϵ -optimal policy for r^*)

A NEW QUANTITATIVE FRAMEWORK

INGREDIENTS

- Model g as a pair $(\mathcal{X}_g, \mathcal{L}_g)$
- $\mathcal{X}_g \rightarrow$ **set** of objects to deploy (e.g., set of policies $\mathcal{X}_g := \Pi$)
- $\mathcal{L}_g : \mathfrak{R} \times \mathcal{X}_g \rightarrow \mathbb{R}_+ \rightarrow$ **loss** of an object for a reward (e.g., suboptimality $\mathcal{L}_g(r, \pi) := J^*(r) - J^\pi(r)$)

OBJECTIVE

$$x^* \in \arg \min_{x \in \mathcal{X}_g} \mathcal{L}_g(r^*, x)$$

A ROBUST APPROACH

But r^* is still unknown → robust approach!

$$\mathcal{L}_g(r^*, x) \leq \max_{r \in \mathcal{R}_f} \mathcal{L}_g(r, x)$$

Q1 When is partial identifiability problematic?

→ *Quantitative* criterion: if the worst-case error $\mathcal{I}_{f,g} := \min_{x \in \mathcal{X}_g} \max_{r \in \mathcal{R}_f} \mathcal{L}_g(r, x)$ is above some given threshold

Q2 What is the “best” approach to cope with partial identifiability in Reward Learning (ReL) problems?

→ *Robust* approach: deploy the best object in the worst case $x_{f,g} := \arg \min_{x \in \mathcal{X}_g} \max_{r \in \mathcal{R}_f} \mathcal{L}_g(r, x)$

ROB-REL - A ROBUST ALGORITHM

Q3 Can we apply it in a provably efficient manner?

→ Yes in at least the following class of ReL problems

SETTING

- Application g : assess which policy between π^1 and π^2 is preferred under r^*
 → $\mathcal{X}_g = \mathbb{R}, \mathcal{L}_g(r, x) = |x - (J^{\pi^1}(r) - J^{\pi^2}(r))|$
- Feedback f_1 : demonstrations from (sub-)optimal expert $\pi^E \rightarrow \mathcal{R}_{f_1} = \{r \in \mathfrak{R} : J^{\pi^E}(r) \geq J^*(r) - \epsilon\}$
- Feedback f_2 : hard trajectory comparisons between ω_1 and $\omega_2 \rightarrow \mathcal{R}_{f_2} = \{r \in \mathfrak{R} : G(\omega_1; r) \geq G(\omega_2; r)\}$
- Feedback f_3 : hard policy comparisons between π'_1 and $\pi'_2 \rightarrow \mathcal{R}_{f_3} = \{r \in \mathfrak{R} : J^{\pi'_1}(r) \geq J^{\pi'_2}(r)\}$
- Finite data: transition models and policies have to be estimated from online interaction and offline data

ROB-REL

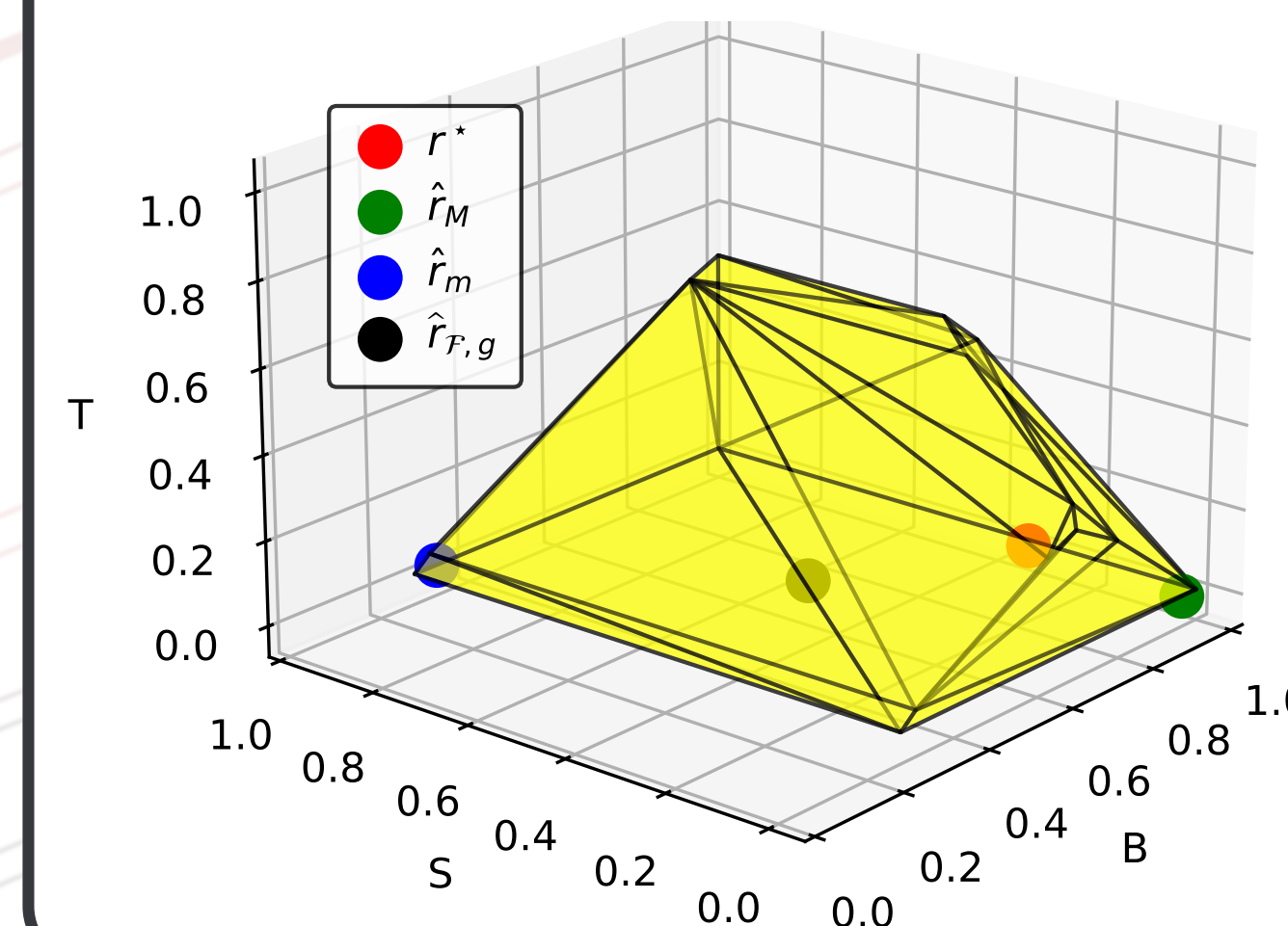
- Estimate the transition model with **RF-Express** (Menard et al., 2021)
- Estimate the occupancy measure of the policies with the empirical estimates
- Note that $x_{f,g} = (M + m)/2$ and $\mathcal{I}_{f,g} = (M - m)/2$, where M, m are max/min over $r \in \mathcal{R}_f$ of $(J^{\pi^1}(r) - J^{\pi^2}(r))$
- Solve the optimization problem with the **primal-dual subgradient method** (Nedic and Ozdaglar, 2009)

Theorem 1 In tabular MDPs, under Slater's condition, for suitable choice of hyperparameters, **Rob-Rel** outputs $\hat{\mathcal{I}}, \hat{x}$ such that, with probability $1 - \delta$: $\mathcal{L}_g(r^*, \hat{x}) \leq \mathcal{I}_{f,g} + \epsilon$ and $|\mathcal{I}_{f,g} - \hat{\mathcal{I}}| \leq \epsilon$, with a number of samples and iterations:

$$N_\pi \leq \tilde{\mathcal{O}}\left(\frac{SAH^5}{\epsilon^2 \xi^2} \log \frac{1}{\delta}\right), \quad N_p \leq \tilde{\mathcal{O}}\left(\frac{SAH^5}{\epsilon^2 \xi^2} \left(S + \log \frac{1}{\delta}\right)\right), \quad N_{it} \leq \tilde{\mathcal{O}}\left(\frac{(SA)^{3/2} H^5 m}{\epsilon^2 \xi^4}\right).$$

ILLUSTRATIVE SIMULATION

- Small MDP with $S = 12, A = 3, H = 5$
- Noisy transition model
- 7 feedback
- State-only 3-dimensional reward
- Error reduction: $1.64 \rightarrow 0.82$



OPEN QUESTIONS

- What are other ReL problems in which this approach is doable?
- How to apply it to non-convex feedback?
- What are meaningful approximations?
- What about problems with $|\mathcal{S}| = |\mathcal{A}| = +\infty$?
- Deep nets function approximation
- Empirical study
- ...

REFERENCES

- Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.
- Pierre Menard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning 38 (ICML)*, 2021.
- A. Nedic and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 2009.
- Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning 40 (ICML)*, 2023.