

MOTIVATION AND CONTRIBUTIONS

MOTIVATION

- **Transferring** behavior to new environments
- IRL is an **ill-posed** problem (Ng et al., 2000)

CONTRIBUTIONS

- Novel **unbiased** criterion for reward selection
- **Closed-form** expression
- Provably efficient algorithms

PROBLEM SETTING

FEASIBLE SETS

$$\mathcal{R}_{\mathcal{M},\pi_E}^{\text{OPT},\bar{\mathcal{S}}} := \{r : \pi_E(s) \in \arg \max_a Q^*(s, a; r) \forall s \in \bar{\mathcal{S}}\}$$

$$\mathcal{R}_{\mathcal{M},\pi_E}^{\text{MCE},\bar{\mathcal{S}}} := \{r : \pi_E(a|s) \propto e^{\frac{1}{\lambda} Q_\lambda^*(s, a; r)} \forall s \in \bar{\mathcal{S}}, \forall a\}$$

$$\mathcal{R}_{\mathcal{M},\pi_E}^{\text{BIRL},\bar{\mathcal{S}}} := \{r : \pi_E(a|s) \propto e^{\frac{1}{\beta} Q^*(s, a; r)} \forall s \in \bar{\mathcal{S}}, \forall a\}$$

PROBLEM FORMULATION

- Given $\mathcal{D}_E \sim \pi_E$
- Given $m \in \{\text{OPT}, \text{MCE}, \text{BIRL}\}$
- Find:

$$\hat{\pi}^* \in \arg \min_{\pi' \in \Pi_{c,k}} \left(\max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; r_E) - V^{\pi'}(s'_0; r_E) \right),$$

where $\Pi_{c,k} = \{\pi \in \Pi : V^\pi(s'_0; c) \leq k\}$.

EXISTING METHODS

“BEST-CASE” METHODS

(Ziebart et al., 2008)

Min suboptimality for any r from the feasible set:

$$\hat{\pi}_B \in \arg \min_{\pi' \in \Pi_{c,k}} \min_{r \in \mathcal{R}_{\mathcal{M},\pi_E}^m} \left(\max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; r) - V^{\pi'}(s'_0; r) \right)$$

“WORST-CASE” METHODS

(Lazzati and Metelli, 2026)

Min suboptimality for worst $r \in \mathfrak{R} = [-1, +1]^{SA}$

$$\hat{\pi}_W \in \arg \min_{\pi' \in \Pi_{c,k}} \max_{r \in \mathcal{R}_{\mathcal{M},\pi_E}^m \cap \mathfrak{R}} \left(\max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; r) - V^{\pi'}(s'_0; r) \right)$$

“ON-AVERAGE” METHODS

(Ramachandran and Amir, 2007)

Min average suboptimality w.r.t. $w : \mathbb{R}^{SA} \rightarrow [0, 1]$

$$\hat{\pi}_w \in \arg \min_{\pi' \in \Pi_{c,k}} \int \mathcal{R}_{\mathcal{M},\pi_E}^m(w(r)) \left(\max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; r) - V^{\pi'}(s'_0; r) \right) dr$$

LIMITATIONS OF EXISTING METHODS

- “Best-Case” Methods → Overly **optimistic** in practice (Finn et al., 2016)
- “Worst-Case” Methods → Too **conservative** and computationally inefficient
- “On-Average” Methods → **Biased** in the space of policies for common $\bar{w}(r) := 1$ if $r \in \mathfrak{R} = [-1, +1]^{SA}$ else 0

Proposition 1 There exists an MDP $\mathcal{R}\mathcal{M}$ and two deterministic policies $\pi_1, \pi_2 \in \Pi$ such that: $\text{vol}(\mathfrak{R} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{\text{OPT},\mathcal{S}}) \neq \text{vol}(\mathfrak{R} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\text{OPT},\mathcal{S}})$.

Proposition 2 There exists an MDP $\mathcal{R}\mathcal{M}$ and two stochastic policies $\pi_1, \pi_2 \in \Pi$ such that, for all $m \in \{\text{MCE}, \text{BIRL}\}$: $\text{vol}(\mathfrak{R} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{m,\mathcal{S}}) \neq \text{vol}(\mathfrak{R} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{m,\mathcal{S}})$.



OUR APPROACH

NOVEL PRIORS OVER REWARDS

$$\mathcal{R}_{\mathcal{M}}^{\text{OPT}} := \{r \in \mathbb{R}^{SA} \mid \forall \pi \in \Pi^*(r), \forall s, a : |V^\pi(s; r)| \leq C_1^{\text{OPT}} k_\pi \wedge |A^\pi(s, a; r)| \leq C_2^{\text{OPT}}\},$$

$$\mathcal{R}_{\mathcal{M}}^{\text{MCE}} := \{r \in \mathbb{R}^{SA} \mid \forall s, a : |V_\lambda^*(s; r)| \leq C_1^{\text{MCE}} \wedge |A_\lambda^*(s, a; r)| \leq C_2^{\text{MCE}}\},$$

$$\mathcal{R}_{\mathcal{M}}^{\text{BIRL}} := \{r \in \mathbb{R}^{SA} \mid \forall s, a : |V^*(s; r)| \leq C_1^{\text{BIRL}} \wedge |A^*(s, a; r)| \leq C_2^{\text{BIRL}}\},$$

Proposition 3 (Close to hypercube) Let $\mathcal{M}$ be any MDP $\mathcal{R}\mathcal{M}$. For some choices of the constants, then, for any $m \in \{\text{OPT}, \text{MCE}, \text{BIRL}\}$: $\mathfrak{R} \subseteq \mathcal{R}_{\mathcal{M}}^m \subseteq [-\rho C_1^m - C_2^m, +\rho C_1^m]^{SA}$.

Proposition 4 (No bias)

- For any MDP $\mathcal{R}\mathcal{M}$ and any pair of deterministic policies $\pi_1, \pi_2 \in \Pi$, it holds that: $\text{vol}(\mathcal{R}_{\mathcal{M}}^{\text{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{\text{OPT},\mathcal{S}}) = \text{vol}(\mathcal{R}_{\mathcal{M}}^{\text{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\text{OPT},\mathcal{S}})$.
- For any MDP $\mathcal{R}\mathcal{M}$ and any pair of stochastic policies $\pi_1, \pi_2 \in \Pi$ such that $\pi_1(a|s) \geq \pi_{\min}$ and $\pi_2(a|s) \geq \pi_{\min}$ in every s, a , if $C_2^{\text{MCE}} \geq \lambda \log(1/\pi_{\min})$, then: $\text{vol}(\mathcal{R}_{\mathcal{M}}^{\text{MCE}} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{\text{MCE},\mathcal{S}}) = \text{vol}(\mathcal{R}_{\mathcal{M}}^{\text{MCE}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\text{MCE},\mathcal{S}})$.
- For any MDP $\mathcal{R}\mathcal{M}$ and any pair of stochastic policies $\pi_1, \pi_2 \in \Pi$ such that $\pi_1(a|s) \geq \pi_{\min}$ and $\pi_2(a|s) \geq \pi_{\min}$ in every s, a , if $C_2^{\text{BIRL}} \geq \beta \log(1/(A\pi_{\min}))$, then: $\text{vol}(\mathcal{R}_{\mathcal{M}}^{\text{BIRL}} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{\text{BIRL},\mathcal{S}}) = \text{vol}(\mathcal{R}_{\mathcal{M}}^{\text{BIRL}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\text{BIRL},\mathcal{S}})$.

CLOSED-FORM EXPRESSIONS AND ALGORITHMS

Algorithm 1: OPT

Input: Data $\mathcal{D}_E$
 $\hat{\mathcal{S}} \leftarrow$ visited states
 $\hat{\mathcal{Z}} \leftarrow$
visited state-action pairs
 $\hat{r}^{\text{OPT}}(s, a) \leftarrow$

$$\begin{cases} 1 & \text{if } (s, a) \in \hat{\mathcal{Z}} \\ \frac{1}{A} & \text{if } (s, a) \notin \hat{\mathcal{Z}} \\ 0 & \text{otherwise} \end{cases}$$

Return $\hat{r}^{\text{OPT}}$

Algorithm 2: MCE

Input: Data $\mathcal{D}_E$, parameter $\pi'_{\min}$
 $N(s, a) \leftarrow$ first-time state-action counts
 $N(s) \leftarrow \sum_{a \in \mathcal{A}} N(s, a)$
 $\hat{\pi}(a|s) \leftarrow \max\{\pi'_{\min}, \frac{N(s, a)}{\max\{1, N(s)\}}\}$
 $\hat{r}^{\text{MCE}}(s, a) \leftarrow \log \hat{\pi}(a|s)$
Return $\hat{r}^{\text{MCE}}$

Algorithm 3: BIRL

Input: Data $\mathcal{D}_E$, parameter $\pi'_{\min}$
 $N(s, a) \leftarrow$ first-time state-action counts
 $N(s) \leftarrow \sum_{a \in \mathcal{A}} N(s, a)$
 $\hat{\pi}(a|s) \leftarrow \max\{\pi'_{\min}, \frac{N(s, a)}{\max\{1, N(s)\}}\}$
 $\hat{r}^{\text{BIRL}}(s, a) \leftarrow$

$$\begin{cases} \log \frac{\hat{\pi}(a|s)}{\max_{a'} \hat{\pi}(a'|s)} & \text{if } N(s) > 0, \\ \log \pi'_{\min} & \text{if } N(s) = 0. \end{cases}$$

Return $\hat{r}^{\text{BIRL}}$

SAMPLE COMPLEXITY ANALYSIS

Theorem 5 Let $\delta \in (0, 1)$ and $H \geq S$. If π_E is deterministic, then, with probability at least $1 - \delta$, it holds that $\hat{r}^{\text{OPT}} = r_{\mathcal{M},\pi_E}^{\text{OPT}}$ provided that:

$$NH \geq \frac{H \log \frac{|\mathcal{S}_{\mathcal{M},\pi_E}|}{\delta}}{p_{\mathcal{M},\pi_E}^{\min,H}}.$$

where $p_{\mathcal{M},\pi_E}^{\min,H}$ is the minimum probability with which π_E visits any state in its support in the first H steps.

Theorem 6 Let $\epsilon, \delta \in (0, 1)$ and $H \geq S$. If $C_2^{\text{MCE}} \geq \lambda \log(1/\pi_{\min})$ and $\pi_{\min} \geq \pi'_{\min}$, then, with probability $1 - \delta$, we have $\|\hat{r}^{\text{MCE}} - r_{\mathcal{M},\pi_E}^{\text{MCE}}\|_\infty \leq \epsilon$ provided that:

$$NH \geq \frac{H \log^2 \frac{4SA}{\delta}}{p_{\mathcal{M},\pi_E}^{\min,H}} \cdot \frac{16}{\epsilon^2 \pi'_{\min}}.$$

Theorem 7 Let $\epsilon, \delta \in (0, 1)$ and $H \geq S$. If $C_2^{\text{BIRL}} \geq \beta \log(1/(A\pi_{\min}))$ and $\pi_{\min} \geq \pi'_{\min}$, then, with probability $1 - \delta$, we have $\|\hat{r}^{\text{BIRL}} - r_{\mathcal{M},\pi_E}^{\text{BIRL}}\|_\infty \leq \epsilon$ provided that:

$$NH \geq \frac{H \log^2 \frac{8SA}{\delta}}{p_{\mathcal{M},\pi_E}^{\min,H}} \cdot \frac{33}{\epsilon^2 \pi'_{\min}}.$$

OPEN QUESTIONS

- How to scale in a provably efficient manner to large/continuous state-action spaces?
- Empirical study in real-world environments, e.g., financial domains
- Simulations to compare practical (e.g., deep) implementations of **RS-BC** with vanilla BC

REFERENCES

- Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning 33 (ICML)*, 2016.
- Filippo Lazzati and Alberto Maria Metelli. Robustness in the face of partial identifiability in reward learning. In *International Conference on Learning Representations (ICLR)*, 2026.
- Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 1, page 2, 2000.
- Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *International Joint Conference on Artificial Intelligence 20 (IJCAI)*, 2007.
- Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence 23 (AAAI)*, 2008.