



MOTIVATION AND CONTRIBUTIONS

MOTIVATION

- Humans are *risk-sensitive* and *non-Markov*
- Existing Imitation Learning (IL) algorithms are *risk-neutral* and/or work with *Markov* policies

CONTRIBUTIONS

- New **risk-sensitive** IL formulation
- Provably efficient* algorithms with **non-Markov** policies when the reward is *known*
- Statistical insights when the reward is *unknown*

NOTATION

- Wasserstein distance $\mathcal{W}(p, q) := \int_{\mathbb{R}} |F_p(x) - F_q(x)| dx$
- Tabular MDP\(\mathcal{R}\) $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, s_0, p)$
- Return $G(\omega; r) := \sum_h r_h(s_h, a_h)$
- Return distribution $\eta_r^\pi(g) := \mathbb{P}^\pi(G(\omega; r) = g)$
- Expected return $J_r^\pi = \mathbb{E}[\eta_r^\pi]$
- Occupancy measure $d_h^\pi(s, a) := \mathbb{P}^\pi(s_h = s, a_h = a)$
- Non-Markov policies $\Pi^{\text{NM}} := \{\pi_h : (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S} \rightarrow \Delta(\mathcal{A})\}$
- Markov policies $\Pi^{\text{M}} := \{\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})\} \subseteq \Pi^{\text{NM}}$

IMITATION LEARNING

- r^E expert's reward
- $\pi^E \in \Pi^{\text{NM}}$ (non-Markov) expert's policy
- $\mathcal{D}^E = \{(s_h, a_h)\} \sim \pi^E$ demonstrations
- Objective:

$$\hat{\pi} \in \arg \min_{\pi \in \Pi^{\text{NM}}} |J_{r^E}^{\pi^E} - J_{r^E}^\pi|$$

- When r^E is unknown:

$$\hat{\pi} \in \arg \min_{\pi \in \Pi^{\text{NM}}} \max_r |J_{r^E}^{\pi^E} - J_r^\pi|$$

- Equivalent to **occupancy measure matching** (Abbeel and Ng, 2004)

$$\hat{\pi} \in \arg \min_{\pi \in \Pi^{\text{NM}}} \sum_h \|d_h^{\pi^E} - d_h^\pi\|$$

- A minimum always exists in Π^{M} (Puterman, 1994)
- **But** this does *not* control the return distribution

$$d^{\pi^E} \approx d^\pi \not\Rightarrow \eta_{r^E}^{\pi^E} \approx \eta_{r^E}^\pi$$

RETURN DISTRIBUTION MATCHING

FORMULATION

When r^E is *known*:

$$\hat{\pi} \in \arg \min_{\pi \in \Pi^{\text{NM}}} \mathcal{W}(\eta_{r^E}^{\pi^E}, \eta_{r^E}^\pi)$$

When r^E is *unknown*:

$$\hat{\pi} \in \arg \min_{\pi \in \Pi^{\text{NM}}} \max_r \mathcal{W}(\eta_{r^E}^{\pi^E}, \eta_r^\pi)$$

OBSERVATIONS

- More expressive than vanilla IL $\rightarrow |J_{r^E}^{\pi^E} - J_{r^E}^{\hat{\pi}}| \leq \mathcal{W}(\eta_{r^E}^{\pi^E}, \eta_{r^E}^{\hat{\pi}})$
- Markov policies do not suffice:

Proposition 1 There exist an MDP $\mathcal{M}_{r^E}$ and $\pi^E \in \Pi^{\text{NM}}$ such that any Markov $\pi \in \Pi^{\text{M}}$ satisfies $\mathcal{W}(\eta_{r^E}^{\pi^E}, \eta_{r^E}^\pi) \geq 0.5$.

- Stronger metrics (e.g., total variation) are sample inefficient:

Theorem 2 There exist an MDP $\mathcal{M}_{r^E}$ and $\pi^E \in \Pi^{\text{NM}}$ such that, even with $|\mathcal{D}^E| = (S-1)^{H-1} - 1$ trajectories, any algorithm satisfies $\mathbb{E}_{\mathcal{D}^E \sim \mathbb{P}^{\pi^E}} \text{TV}(\eta_{r^E}^{\pi^E}, \eta_{r^E}^{\hat{\pi}}) > 0.1$.

KNOWN REWARD SETTING

AN EFFICIENT CLASS OF NON-MARKOV POLICIES

Adopt a **discretization approach** (Bastani et al., 2022)

- Given $\theta \in (0, 1]$, build a discretized reward: $r_{\theta, h}^E(s, a) := \arg \min_{x \in \{0, \theta, 2\theta, \dots, \lfloor 1/\theta \rfloor \theta\}} |x - r_h^E(s, a)|$
- Markov policies in the return-augmented space: $\Pi(r_\theta^E) := \left\{ \pi \in \Pi^{\text{NM}} \mid \pi(a|s, \omega) = f(s, G(\omega; r_\theta^E), a, h) \right\}$

Lemma 3 For any MDP $\mathcal{M}_{r^E}$ and $\pi^E \in \Pi^{\text{NM}}$, the set $\Pi(r_\theta^E)$ contains at least a π such that: $\mathcal{W}(\eta_{r^E}^{\pi^E}, \eta_{r^E}^\pi) \leq H\theta$.

ALGORITHMS

Algorithm 1: RS-BC

Input: $\mathcal{D}^E = \{(s_h^i, a_h^i)\}_i, r^E, \theta$
 $N_h(s, g, a) \leftarrow \sum_i \mathbb{1}\{s_h^i = s, a_h^i = a, G(\omega_h^i; r_\theta^E) = g\}$
 $\hat{\pi}(a|s, \omega) \leftarrow \begin{cases} \frac{N_h(s, G(\omega; r_\theta^E), a)}{\sum_{a'} N_h(s, G(\omega; r_\theta^E), a')} & \text{if den} > 0 \\ \frac{1}{A} & \text{otherwise} \end{cases}$
Return $\hat{\pi}$

Theorem 4 Set $\theta = \epsilon/(4H)$. Then, with probability $1 - \delta$, the policy $\hat{\pi}$ output by Algorithm 1 satisfies $\mathcal{W}(\eta_{r^E}^{\pi^E}, \eta_{r^E}^{\hat{\pi}}) \leq \epsilon$, with a number of samples:

$$N \leq \tilde{\mathcal{O}}\left(\frac{SH^6 \ln \frac{1}{\delta}}{\epsilon^3} \left(A + \ln \frac{1}{\delta}\right)\right)$$

Algorithm 2: RS-KT

Input: $\mathcal{D}^E = \{(s_h^i, a_h^i)\}_i, r^E, \theta, p$
 $\hat{\eta}(g) \leftarrow \frac{1}{N} \sum_{i \in [N]} \mathbb{1}\{G(\omega_h^i; r_\theta^E) = g\}$
 $\hat{\pi} \in \arg \min_{\pi \in \Pi(r_\theta^E)} \mathcal{W}(\eta_{r^E}^{\pi^E}, \hat{\eta}) \quad // \text{LP}$
Return $\hat{\pi}$

Theorem 5 Set $\theta = \epsilon/(7H)$ and assume that the optimization problem is solved exactly. Then, with probability $1 - \delta$, the policy $\hat{\pi}$ output by Algorithm 2 satisfies $\mathcal{W}(\eta_{r^E}^{\pi^E}, \eta_{r^E}^{\hat{\pi}}) \leq \epsilon$, with a number of samples:

$$N \leq \mathcal{O}\left(\frac{H^2}{\epsilon^2} \ln \frac{1}{\delta}\right)$$

UNKNOWN REWARD SETTING

Discretize any reward r :

$$r_{\theta, h}(s, a) := \arg \min_{x \in \{0, \theta, 2\theta, \dots, \lfloor 1/\theta \rfloor \theta\}} |x - r_h(s, a)|$$

Algorithm 3: RS-KT- variant

Input: $\mathcal{D}^E = \{(s_h^i, a_h^i)\}_i, p$
Define $\hat{\eta}_r(g) := \frac{1}{N} \sum_i \mathbb{1}\{G(\omega^i; r_\theta) = g\} \quad \forall r$
 $\hat{\pi} \in \arg \min_{\pi \in \Pi^{\text{NM}}} \max_r \mathcal{W}(\eta_r^\pi, \hat{\eta}_r) \quad // \text{oracle}$
Return $\hat{\pi}$

Theorem 6 Set $\theta = \epsilon/(2H)$. Then, with probability $1 - \delta$, the policy $\hat{\pi}$ output by the oracle-based Algorithm 3 satisfies $\max_r \mathcal{W}(\eta_r^{\pi^E}, \eta_r^{\hat{\pi}}) \leq 2\epsilon$, with a number of samples:

$$N \leq \tilde{\mathcal{O}}\left(\frac{SAH^3}{\epsilon^2} \ln \frac{1}{\delta}\right)$$

We conjecture that no computationally and sample efficient algorithm exists even for this known-transition setting

OPEN QUESTIONS

- How to scale in a provably efficient manner to large/continuous state-action spaces?
- Empirical study in financial domains
- Simulations to compare practical (e.g., deep) implementations of **RS-BC** with vanilla BC (Pomerleau, 1988)
- ...

REFERENCES

- Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning 21 (ICML)*, 2004.
- Osbert Bastani, Jason Yecheng Ma, Estelle Shen, and Wanqiao Xu. Regret bounds for risk-sensitive reinforcement learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.
- Dean A. Pomerleau. Alvin: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems 1 (NeurIPS)*, 1988.
- Martin Lee Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.