Regularized Inverse Reinforcement Learning

Wonseok Jeon[1,2], Chen-Yang Su[1,2], Paul Barde[1,2], Thang Doan[1,2], Derek Nowrouzezahrai[1,2], Joelle Pineau[1,2,3]

[1]Mila - Quebec AI Institute
[2]McGill University
[3]Facebook AI Research

ICLR 2021

Agent-Environment Interaction

- Markov Decision Process
  - A set of states S
  - A set of actions A
  - A transition distribution $T(\cdot|s,a) \in \Delta^S$ (Δ^X: A set of probs on X)
  - A reward function $r(s,a)$
  - A discount factor γ
  - An initial state distribution $P_0 \in \Delta^S$
- Policy $\pi(\cdot|s) \in \Delta^A$
  - The agent's probability of choosing an action
- Joint distribution
  - $s_0 \sim P_0, a_i \sim \pi(\cdot|s_i), s_{i+1} \sim T(\cdot|s_i,a_i), i \geq 0.$

Reinforcement Learning

- Return $R = \sum_{i=0}^{\infty} \gamma^i r(s_i, a_i)$.
- Learning objective $\pi_* \in \text{argmax}_\pi \mathbb{E}_\pi[R]$.
  - Values

$$V_\pi(s) = \mathbb{E}_\pi \left[R | s_0 = s \right].$$
$$Q_\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim T(\cdot|s,a)} V_\pi(s').$$

  - (Unique) optimal Q value

$$Q_*(s, a) = \max_\pi Q_\pi(s, a), \forall s, a.$$

  - Optimal policy via greediness ($\langle f, g \rangle = \sum_{a \in A} f(a) g(a)$)

$$\max_{\pi(\cdot|s)} \langle \pi(\cdot|s), Q_*(s, \cdot) \rangle$$

Reinforcement Learning

- π_* may not be unique.

$$\begin{tikzcd}
r(s,a) = 1 \arrow[r] & Q_\pi(s,a) = \frac{1}{1-\gamma} \arrow[r] \arrow[dr] \arrow[ddr] \arrow[dddr] \arrow[ddddr] & \pi_*^{(0)} \\
 & & \pi_*^{(1)} \\
 & & \pi_*^{(2)} \\
 & & \pi_*^{(3)} \\
 & & \pi_*^{(4)} \\
 & & \vdots
\end{tikzcd}$$

Reinforcement Learning

- π_* may not be unique.

$$r(s,a)=1 \longrightarrow Q_\pi(s,a)=\frac{1}{1-\gamma} \qquad \pi_*^{(0)}$$

$$r(s,a)=2 \longrightarrow Q_\pi(s,a)=\frac{2}{1-\gamma} \qquad \pi_*^{(1)}$$

$$r(s,a)=3 \longrightarrow Q_\pi(s,a)=\frac{3}{1-\gamma} \qquad \pi_*^{(2)}$$

$$r(s,a)=4 \longrightarrow Q_\pi(s,a)=\frac{4}{1-\gamma} \qquad \pi_*^{(3)}$$

$$r(s,a)=5 \longrightarrow Q_\pi(s,a)=\frac{5}{1-\gamma} \qquad \pi_*^{(4)}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

Regularized Reinforcement Learning [Geist et al., ICML 2019]

- Regularized return $R^{\Omega} = \sum_{i=0}^{\infty} \gamma^i (r(s_i, a_i) - \Omega(\pi(\cdot|s_i)))$.
  - A strongly convex function $\Omega : \Delta^A \to \mathbb{R}$
- Learning objective $\pi_* \in \operatorname{argmax}_{\pi} \mathbb{E}_{\pi}[R^{\Omega}]$.
  - Values

$$V_{\pi}^{\Omega}(s) = \mathbb{E}_{\pi}\left[R^{\Omega}|s_0 = s\right].$$
$$Q_{\pi}^{\Omega}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim T(\cdot|s,a)} V_{\pi}^{\Omega}(s').$$

  - (Unique) optimal Q value

$$Q_*^{\Omega}(s, a) = \max_{\pi} Q_{\pi}^{\Omega}(s, a), \forall s, a.$$

  - Optimal policy via greediness ($\langle f, g \rangle = \sum_{a \in A} f(a) g(a)$).

$$\max_{\pi(\cdot|s)} \langle \pi(\cdot|s), Q_*^{\Omega}(s, \cdot) \rangle - \Omega(\pi(\cdot|s))$$

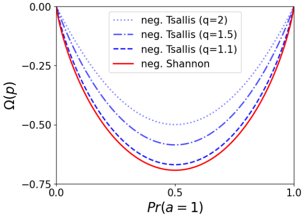
Regularized Reinforcement Learning [Geist et al., ICML 2019]

- e.g.,
  - Shannon-entropy-regularized reinforcement learning

$$\Omega(\pi(\cdot|s)) = -H(\pi(\cdot|s))$$

  - Tsallis-entropy-regularized reinforcement learning ($k > 0, q > 1$)

$$\Omega(\pi(\cdot|s)) = -T_q^k(\pi(\cdot|s))$$



Regularized Reinforcement Learning[Geist et al., ICML 2019]

- The **convex conjugate** $\Omega^* : \mathbb{R}^A \to \mathbb{R}$ of $\Omega : \Delta^A \to \mathbb{R}$

$$\Omega^*(Q_*^\Omega(s,\cdot)) = \max_{\pi(\cdot|s)\in\Delta^A} \langle \pi(\cdot|s), Q_*^\Omega(s,\cdot)\rangle - \Omega(\pi(\cdot|s)).$$

  - For a strongly convex Ω, the maximizer is **unique** and is equal to

$$\pi_*(\cdot|s) = \nabla\Omega^*(Q_*^\Omega(s,\cdot)).$$

Regularized Reinforcement Learning[Geist et al., ICML 2019]

- π_* is unique! Let $\Omega(\pi(\cdot|s)) = -H(\pi(\cdot|s))$.

$$r(s,a) = 1 \longrightarrow Q_\pi^\Omega(s,a) = \frac{1}{1-\gamma} + \sum_{i=1}^\infty \gamma^i H(\pi(\cdot|s_i)) \longrightarrow \pi_*(a|s) = \frac{1}{|A|}$$

(Maxium Entropy)

Regularized Reinforcement Learning [Geist et al., ICML 2019]

- π_* is unique! Let $\Omega(\pi(\cdot|s)) = -H(\pi(\cdot|s))$.

$$r(s,a) = 1 \longrightarrow Q_\pi^\Omega(s,a) = \frac{1}{1-\gamma} + \sum_{i=1}^\infty \gamma^i H(\pi(\cdot|s_i)) \longrightarrow \pi_*(a|s) = \frac{1}{|A|}$$

(Maxium Entropy)

$$r(s,a) = 2 \longrightarrow Q_\pi^\Omega(s,a) = \frac{2}{1-\gamma} + \sum_{i=1}^\infty \gamma^i H(\pi(\cdot|s_i))$$

$$r(s,a) = 3 \longrightarrow Q_\pi^\Omega(s,a) = \frac{3}{1-\gamma} + \sum_{i=1}^\infty \gamma^i H(\pi(\cdot|s_i))$$

$$r(s,a) = 4 \longrightarrow Q_\pi^\Omega(s,a) = \frac{4}{1-\gamma} + \sum_{i=1}^\infty \gamma^i H(\pi(\cdot|s_i))$$

$$r(s,a) = 5 \longrightarrow Q_\pi^\Omega(s,a) = \frac{5}{1-\gamma} + \sum_{i=1}^\infty \gamma^i H(\pi(\cdot|s_i))$$

$$\vdots \qquad\qquad \vdots$$

Inverse Reinforcement Learning [Ng et al., ICML 2000]

- Expert Policy $\pi_E(\cdot|s) \in \Delta^A$
  - The expert's probability of choosing an action
- Return $R(r) = \sum_{i=0}^{\infty} \gamma^i r(s_i, a_i)$.
- Learning objective

$$IRL(\pi_E) := \text{argmax}_{r \in \mathbb{R}^{S \times A}} \left\{ \underbrace{\mathbb{E}_{\pi_E}[R(r)]}_{\substack{\text{expert's} \\ \text{return}}} - \underbrace{\max_{\pi} \mathbb{E}_{\pi}[R(r)]}_{\substack{\text{optimal} \\ \text{return}}} \right\}.$$

Inverse Reinforcement Learning [Ng et al., ICML 2000]

- IRL has degenerate solutions.
    - *Constant rewards are IRL solutions for any policies.*

$$r(s,a)=1 \longleftarrow Q_\pi(s,a)=\frac{1}{1-\gamma} \qquad \pi_E^{(0)}$$

$$r(s,a)=2 \longleftarrow Q_\pi(s,a)=\frac{2}{1-\gamma} \qquad \pi_E^{(1)}$$

$$r(s,a)=3 \longleftarrow Q_\pi(s,a)=\frac{3}{1-\gamma} \qquad \pi_E^{(2)}$$

$$r(s,a)=4 \longleftarrow Q_\pi(s,a)=\frac{4}{1-\gamma} \qquad \pi_E^{(3)}$$

$$r(s,a)=5 \longleftarrow Q_\pi(s,a)=\frac{5}{1-\gamma} \qquad \pi_E^{(4)}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

Regularized Inverse Reinforcement Learning [Geist et al., ICML 2019]

- Expert Policy $\pi_E(\cdot|s) \in \Delta^A$
  - The expert's probability of choosing an action
- Regularized return $R^\Omega = \sum_{i=0}^{\infty} \gamma^i (r(s_i, a_i) - \Omega(\pi(\cdot|s_i)))$.
- Learning objective

$$IRL(\pi_E) := \text{argmax}_{r \in \mathbb{R}^{S \times A}} \left\{ \underbrace{\mathbb{E}_{\pi_E}[R^\Omega(r)]}_{\substack{\text{expert's} \\ \text{regularized} \\ \text{return}}} - \underbrace{\max_\pi \mathbb{E}_\pi[R^\Omega(r)]}_{\substack{\text{optimal} \\ \text{regularized} \\ \text{return}}} \right\}.$$

Regularized Inverse Reinforcement Learning [Geist et al., ICML 2019]

- Regularized IRL does not suffer from degeneracy.
  - *Constant rewards correspond to uniform policy, e.g.,* $\Omega = -H$

$$\pi_E(a|s) = \frac{1}{|A|}$$

(Maxium Entropy)

$$r(s,a) = 1 \longleftarrow Q_\pi^\Omega(s,a) = \frac{1}{1-\gamma} + \sum_{i=1}^{\infty} \gamma^i H(\pi(\cdot|s_i)) \longleftarrow \pi_E$$

$$r(s,a) = 2 \longleftarrow Q_\pi^\Omega(s,a) = \frac{2}{1-\gamma} + \sum_{i=1}^{\infty} \gamma^i H(\pi(\cdot|s_i)) \longleftarrow \pi_E$$

$$r(s,a) = 3 \longleftarrow Q_\pi^\Omega(s,a) = \frac{3}{1-\gamma} + \sum_{i=1}^{\infty} \gamma^i H(\pi(\cdot|s_i)) \longleftarrow \pi_E$$

$$r(s,a) = 4 \longleftarrow Q_\pi^\Omega(s,a) = \frac{4}{1-\gamma} + \sum_{i=1}^{\infty} \gamma^i H(\pi(\cdot|s_i)) \longleftarrow \pi_E$$

$$r(s,a) = 5 \longleftarrow Q_\pi^\Omega(s,a) = \frac{5}{1-\gamma} + \sum_{i=1}^{\infty} \gamma^i H(\pi(\cdot|s_i)) \longleftarrow \pi_E$$

$$\vdots \qquad\qquad \vdots$$

Motivation

- *Tractable solutions for regularized inverse RL?*
- *An algorithm to derive learn a solution?*

A Solution of Regularized IRL

Theorem (A Solution of Regularized IRL)

$$t(s, a; \pi_E) = [\nabla\Omega(\pi_E(\cdot|s))]_a - \langle \pi_E(\cdot|s), \nabla\Omega(\pi_E(\cdot|s))\rangle + \Omega(\pi(\cdot|s)).$$

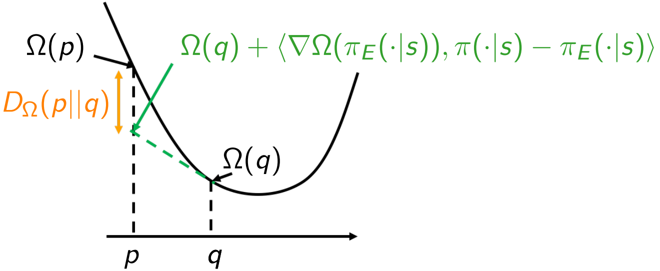
A Solution of Regularized IRL

Theorem (A Solution of Regularized IRL)

$$t(s,a;\pi_E) = [\nabla\Omega(\pi_E(\cdot|s))]_a - \langle \pi_E(\cdot|s), \nabla\Omega(\pi_E(\cdot|s))\rangle + \Omega(\pi(\cdot|s)).$$

- Proof. For Bregman divergence $D_\Omega(p||q)$,

$$\text{argmax}_\pi\, \mathbb{E}_\pi[R^\Omega(t(\cdot,\cdot;\pi_E))] = \text{argmin}_\pi\, \mathbb{E}_\pi\left[\sum_{i=0}^\infty \gamma^i D_\Omega(\pi(\cdot|s_i)||\pi_E(\cdot|s_i))\right] = \pi_E$$



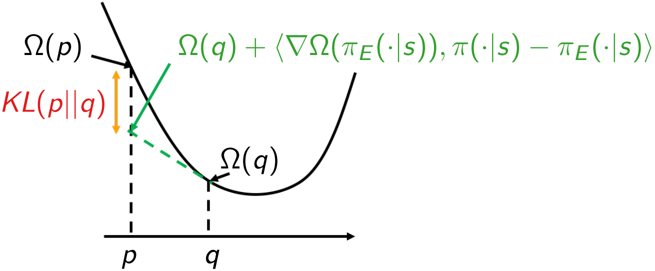
A Solution of MaxEnt IRL [Ziebart et al., 2008, Ho et al., 2016]

Theorem (A Solution of Regularized IRL)

$$t(s,a;\pi_E) = [\nabla\Omega(\pi_E(\cdot|s))]_a - \langle \pi_E(\cdot|s), \nabla\Omega(\pi_E(\cdot|s))\rangle + \Omega(\pi(\cdot|s)).$$

- Proof. For KL divergence $KL(p||q)$, $t(s,a;\pi_E) = \log \pi_E(a|s)$.

$$\operatorname{argmax}_\pi \mathbb{E}_\pi[R^\Omega(t(\cdot,\cdot;\pi_E))] = \operatorname{argmin}_\pi \mathbb{E}_\pi\left[\sum_{i=0}^\infty \gamma^i KL(\pi(\cdot|s_i)||\pi_E(\cdot|s_i))\right] = \pi_E$$



Optimal Advantage Function

Theorem (A Solution of Regularized IRL)

$$\underbrace{t(s,a;\pi_E)}_{A^{\Omega}_{\pi_E}(s,a)} = \underbrace{[\nabla\Omega(\pi_E(\cdot|s))]_a}_{Q^{\Omega}_{\pi_E}(s,a)} - \underbrace{\{\langle\pi_E(\cdot|s), \nabla\Omega(\pi_E(\cdot|s))\rangle - \Omega(\pi(\cdot|s))\}}_{V^{\Omega}_{\pi_E}(s)}.$$

Additional Solutions via Reward Shaping [Ng et al., 1999]

Theorem (Potential-based reward shaping)

Let π^* be the optimal policy of regularized RL with a reward $r \in \mathbb{R}^{S \times A}$. Then for $\Phi \in \mathbb{R}^S$, using

$$r_\Phi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim T(\cdot|s,a)} \Phi(s') - \Phi(s)$$

as a reward also leads to π^*.

Regularized IRL in Continuous Control

- Tractable when
  - $\pi_E(\cdot|s)$ follows independent normal distributions.
  - Negative Tsallis entropy regularizer $\Omega(\pi(\cdot|s)) = -T_q^k(\pi(\cdot|s))$

Regularized IRL in Continuous Control

- If $\pi(\cdot|s)$ follows independent normal distributions, the Bregman divergence is also tractable.
  - $\pi = \mathcal{N}(\mu, \sigma^2)$ and $\pi_E = \mathcal{N}(0, (e^{-3})^2)$
  - For larger q, means and variances are matched more tightly.

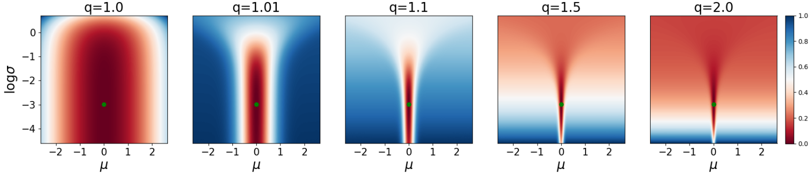


Figure: Bregman divergence $D_\Omega(\pi||\pi_E)$

Algorithmic Consideration

Algorithm 1 Regularized Adversarial IRL(RAIRL)

1: Expert demonstration $\mathcal{D}_E \sim \pi_E$.
2: **for** each iteration **do**
3: $\mathcal{D}_\pi := \{(s,a)\} \sim \pi$.
4: Reward learning (binary classification)

$$\max_{r \in \mathbb{R}^{S \times A}} \mathbb{E}_{(s,a)\sim\mathcal{D}_E} \log D_{r,\pi}(s,a) + \mathbb{E}_{(s,a)\sim d_\pi} \log(1 - D_{r,\pi}(s,a))$$

$$D_{r,\pi}(s,a) = \sigma(r(s,a) - t(s,a;\pi))$$

5: Policy optimization via Regularized Actor Critic [Yang et al., NeurIPS 2019]:

$$\max_\pi \mathbb{E}[R^\Omega(r)|\pi]$$

6: **end for**
7: **Output:** π_E, $t(s,a;\pi_E)$.

Experiments

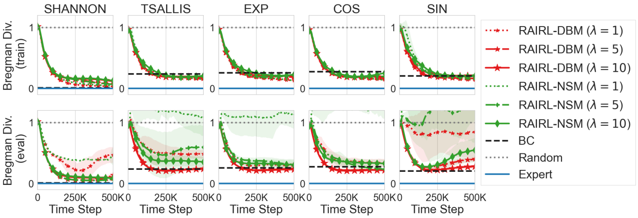


Figure: BermudaWorld (Continuous Observation, Discrete Action)

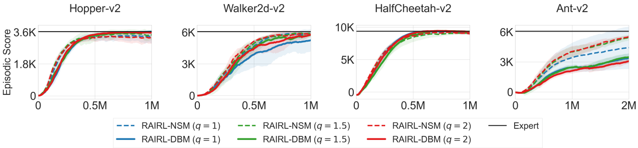
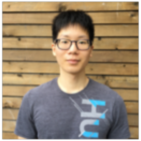


Figure: MuJoCo (Continuous Observation, Continuous Action)

For more information, please check our paper and poster!

Poster Session 3

May 3rd, 2021, 5 pm-7 pm (PDT)



Mila

McGill