Robot Learning

Inverse Reinforcement Learning

Prof. Songhwai Oh
ECE, SNU

A. Y. Ng and S. Russell, "**Algorithms for inverse reinforcement learning**," in Proc. of the International Conference on Machine Learning (ICML), Jun. 2000.

INVERSE REINFORCEMENT LEARNING

Setting

- MDP: $M = (\mathcal{S}, \mathcal{A}, \{P_{sa}\}, \gamma, R)$
- Finite states and actions: $|\mathcal{S}| = N$, $|\mathcal{A}| = k$
- $|R| \leq R_{\max}$
- $\pi : \mathcal{S} \rightarrow \mathcal{A}$
- $V^{\pi}(s_1) = \mathbb{E}\left(R(s_1) + \gamma R(s_2) + \gamma^2 R(s_3) + \cdots \middle| \pi\right)$
- $Q^{\pi}(s, a) = R(s) + \gamma \mathbb{E}_{s' \sim P_{sa}(\cdot)}\left(V^{\pi}(s')\right)$
- $\mathbf{P}_a$: $N \times N$ matrix, for each action a
- $(\mathbf{P}_a)_{ij}$: the probability of making a transition to state j from state i if action a is taken.

Bellman Optimality

- Bellman equations:

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P_{s\pi(s)}(s') V^{\pi}(s')$$

$$Q^{\pi}(s,a) = R(s) + \gamma \sum_{s'} P_{sa}(s') V^{\pi}(s')$$

- Bellman optimality: π is an optimal policy for M if and only if, for all $s \in \mathcal{S}$,

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} Q^{\pi}(s,a)$$

Inverse Reinforcement Learning (IRL)

IRL Problem:

- Given: $\mathcal{S}, \mathcal{A} = \{a_1, \ldots, a_k\}, \{P_{sa}\}, \gamma, \pi$
- Goal: Find R such that π is an optimal policy
- Assumption (to make notation simpler): $\pi(s) \equiv a_1$

Theorem 3. $\pi(s) \equiv a_1$ is optimal if and only if, for all $a \in \{a_2, \ldots, a_k\}$,

$$(\mathbf{P}_{a_1} - \mathbf{P}_a)\left(\mathbb{I} - \gamma \mathbf{P}_{a_1}\right)^{-1} R \succeq 0,$$

where $x \succeq y$ if and only if $x_i \geq y_i$ for all i.

Proof

From the Bellman optimality

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} Q^{\pi}(s,a)$$

$$= \arg\max_{a \in \mathcal{A}} \left(R(s) + \gamma \sum_{s'} P_{sa}(s') V^{\pi}(s') \right)$$

$$= \arg\max_{a \in \mathcal{A}} \left(\sum_{s'} P_{sa}(s') V^{\pi}(s') \right)$$

$$\iff \quad \sum_{s'} P_{sa_1}(s') V^{\pi}(s') \geq \sum_{s'} P_{sa}(s') V^{\pi}(s') \qquad \forall s, a$$

$$\iff \quad \mathbf{P}_{a_1} V^{\pi} \succeq \mathbf{P}_{a} V^{\pi} \qquad \forall a \in \mathcal{A} \setminus a_1$$

$$
\begin{aligned}
V^\pi &= R + \gamma \mathbf{P}_{a_1} V^\pi \\
V^\pi - \gamma \mathbf{P}_{a_1} V^\pi &= R \\
\left(\mathbb{I} - \gamma \mathbf{P}_{a_1}\right) V^\pi &= R \\
V^\pi &= \left(\mathbb{I} - \gamma \mathbf{P}_{a_1}\right)^{-1} R
\end{aligned}
$$

$$
\begin{aligned}
& \pi(s) \in \arg\max_{a \in \mathcal{A}} Q^\pi(s,a) \\
\iff \quad & \mathbf{P}_{a_1} V^\pi \succeq \mathbf{P}_a V^\pi && \forall a \in \mathcal{A} \setminus a_1 \\
\iff \quad & \mathbf{P}_{a_1} \left(\mathbb{I} - \gamma \mathbf{P}_{a_1}\right)^{-1} R \succeq \mathbf{P}_a \left(\mathbb{I} - \gamma \mathbf{P}_{a_1}\right)^{-1} R && \forall a \in \mathcal{A} \setminus a_1 \\
\iff \quad & \left(\mathbf{P}_{a_1} - \mathbf{P}_a\right) \left(\mathbb{I} - \gamma \mathbf{P}_{a_1}\right)^{-1} R \succeq 0 && \forall a \in \mathcal{A} \setminus a_1
\end{aligned}
$$

□

- Trivial solution: $R \equiv 0$.
- There can be many R's satisfying the condition
- Need some criteria to find R

Optimization Problem for IRL

- Criterion: Find R which

$$\max \sum_{s \in \mathcal{S}} \left(Q^{\pi}(s, a_1) - \max_{a \in \mathcal{A} \setminus a_1} Q^{\pi}(s, a) \right)$$

to make other actions as bad as possible

- Optimization problem for IRL

$$\max \sum_{i=1}^{N} \min_{a \in \mathcal{A} \setminus a_1} \left(\mathbf{P}_{a_1}(i) - \mathbf{P}_a(i) \right) \left(\mathcal{I} - \gamma \mathrm{P}_{a_1} \right)^{-1} R - \lambda \|R\|_1$$

subject to

$$\begin{aligned} \left(\mathbf{P}_{a_1} - \mathbf{P}_a \right) \left(\mathbb{I} - \gamma \mathbf{P}_{a_1} \right)^{-1} R \succeq 0 \qquad & \forall a \in \mathcal{A} \setminus a_1 \\ |R_i| \leq R_{\max} \qquad & \forall i \end{aligned}$$

Linear Function Approximation for Infinite-State MDPs

- Infinite-state MDP: $S \in \mathbb{R}^n$
- ϕ_i: fixed basis function
- α_i: unknown parameters
- V_i^π: value function if $R = \phi_i$

$$R(s) = \sum_{i=1}^{d} \alpha_i \phi_i(s)$$

$$V^\pi = \sum_{i=1}^{d} \alpha_i V_i^\pi$$

- Given m sampled trajectories

$\hat{V}_i^\pi(s_0)$: average empirial return

$$\hat{V}^\pi(s_0) = \sum_{i=1}^{d} \alpha_i \hat{V}_i^\pi(s_0)$$

Algorithm

- Given set of policies $\{\pi_1, \ldots, \pi_k\}$

  (1) Find α such that

$$\max \sum_{i=1}^{k} p\left(\hat{V}^{\pi^*}(s_0) - \hat{V}^{\pi_i}(s_0)\right) \qquad \text{s.t. } |\alpha_j| \leq 1 \qquad j = 1, \ldots, d$$

  (2) Find R

  (3) Find V^{π}

  (4) Find π_{k+1} by solving MDP

$p(x) = x$ if $x \geq 0$

$p(x) = 2x$ if $x < 0$

To penalize more if the condition is violated

- Repeat (1)-(4)
- Makes π^* the best compared to other possible policies over sampled trajectories

Abbeel, Pieter, and Andrew Y. Ng. "**Apprenticeship learning via inverse reinforcement learning**." In Proceedings of the International Conference on Machine Learning (ICML), 2004.

APPRENTICESHIP LEARNING

- MDP: (S, A, T, γ, D, R)
  - $T = \{P_{sa}\}$, transition probabilities
  - D: initial state distribution
  - $R : S \to \mathbb{R}$, reward function
- MDP\R: IRL problem
- $\phi : S \to [0, 1]^k$, features about states
- $R^*(s) = w^* \cdot \phi(s)$, true reward function $(\|w^*\|_1 \leq 1)$
- Value of π

$$\mathbb{E}_{s_0 \sim D}\left(V^{\pi}(s_0)\right) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t R(s_t)|\pi\right) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t w \cdot \phi(s_t)|\pi\right)$$
$$= w \cdot \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t \phi(s_t)|\pi\right) = w \cdot \mu(\pi)$$

  - $\mu(\pi)$: feature expectation

- π_E: expert's policy
- $\mu_E = \mu(\pi_E)$
- If $\|\mu(\tilde{\pi}) - \mu_E\|_2 \leq \epsilon$ for some policy $\tilde{\pi}$, then for any $w \in \mathbb{R}^k$ $(\|w\|_1 \leq 1)$,

$$\begin{aligned} \left| \mathbb{E}\left(\sum_t \gamma^t R(s_t)|\pi_E\right) - \mathbb{E}\left(\sum_t \gamma^t R(s_t)|\tilde{\pi}\right) \right| &= \left| w \cdot \mu(\tilde{\pi}) - w \cdot \mu_E \right| \\ \text{(Cauchy-Schwarz inequality)} \quad &\leq \|w\|_2 \|\mu(\tilde{\pi}) - \mu_E\|_2 \\ &\leq 1 \cdot \epsilon = \epsilon \end{aligned}$$

Algorithm

- Quadratic programming

$$\max_{t,w} \quad t$$

$$\text{subject to } w \cdot \mu_E \geq w \cdot \mu^{(j)} + t \qquad j = 0, 1, \ldots, i-1$$

$$\|w\|_2 \leq 1$$

- Expert does better by the largest margin compared to all previous policies

1. Randomly pick some policy $\pi^{(0)}$, compute (or approximate via Monte Carlo) $\mu^{(0)} = \mu(\pi^{(0)})$, and set $i = 1$.
2. Compute $t^{(i)} = \max_{w:\|w\|_2 \leq 1} \min_{j \in \{0..(i-1)\}} w^T(\mu_E - \mu^{(j)})$, and let $w^{(i)}$ be the value of w that attains this maximum.
3. If $t^{(i)} \leq \epsilon$, then terminate.
4. Using the RL algorithm, compute the optimal policy $\pi^{(i)}$ for the MDP using rewards $R = (w^{(i)})^T \phi$.
5. Compute (or estimate) $\mu^{(i)} = \mu(\pi^{(i)})$.
6. Set $i = i + 1$, and go back to step 2.