CS 4/5789 27 Apr 2022 Prof Sarah Dean

Lecture 27: Max Entropy IRL

Recall the IRL setting: Finite Horizon MDP

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, H, \mu\}$$

where reward is unobserved and we have a dataset $\mathcal{D} = \{s_i^*, a_i^*\} \sim d_\mu^{\pi^*}$

Further assume linear reward function

$$r(s,a) = \theta_*^\top \varphi(s,a)$$

unknown ↗ ↖ known

Notice that $\mathbb{E}_{s,a \sim d_1} \varphi(s,a) = \mathbb{E}_{s,a\sim d_2} \varphi(s,a) \Rightarrow \mathbb{E}_{s,a\sim d_1} r(s,a) = \mathbb{E}_{s,a\sim d_2} r(s,a)$

(why? by linearity of expectation)

State/Action distributions:

$P_h^\pi(s,a;\mu)$: probability of visiting (s,a) at step h using π

$d_\mu^\pi(s,a) = \sum_{h=0}^{H-1} P_h^\pi(s,a;\mu)/H$ average state-action distribution

$d_\mu^\pi(s) = \sum_{a\in\mathcal{A}} d_\mu^\pi(s,a)$ average state distribution.

Max-Ent IRL Problem

last lecture we arrived at the constrained optimization problem:

$$\min_{\pi} \mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\left[\log(\pi(a|s))\right]$$

equivalent to maximizing entropy

ensuring consistency w/ expert

$$\text{s.t.} \quad \mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\left[\varphi(s,a)\right] = \mathbb{E}_{s,a\sim d^{\pi^*}_{\mu}}\left[\varphi(s,a)\right]$$

estimate from $\mathcal{D}$

d dimensional expression, $\varphi: \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$

Then using the Lagrange formulation:

$$\max_{w\in\mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\left[\log\pi(a|s)\right] + w^T\left(\mathbb{E}_{s,a\sim d^{\pi^*}_{\mu}}\left[\varphi(s,a)\right] - \mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\left[\varphi(s,a)\right]\right)}_{\mathcal{L}(\pi,w)}$$

Notice that we can write

$$\mathcal{L}(\pi,w) = \mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\left[\log\pi(a|s) - w^T\varphi(s,a)\right] + w^T\mathbb{E}_{s,a\sim d^{\pi^*}_{\mu}}\left[\varphi(s,a)\right]$$

2) Iterative Max-Ent IRL

Initialize $w_0 \in \mathbb{R}^d$

For $t = 0, \dots, T-1$

$$\pi^t = \arg\max_{\pi} \mathbb{E}_{s,a \sim d^{\pi}_{\mu}} \left[w_t^{\top} \phi(s,a) - \log \pi(a|s) \right]$$ (best response)

$$w_{t+1} = w_t + \eta \left(\mathbb{E}_{s,a \sim d^{\pi^*}_{\mu}} \phi(s,a) - \mathbb{E}_{s,a \sim d^{\pi^t}_{\mu}} \phi(s,a) \right)$$ (gradient update)

Return $\bar{\pi} = \text{Uniform}(\pi^0, \dots, \pi^{T-1})$

$$\bar{\pi}(a|s) = \frac{1}{T} \sum_{t=0}^{T-1} \pi^t(a|s)$$

Best Response step is like RL problem with reward $w_t^{\top}\phi(s,a)$ and addition policy dependent term $-\log \pi(a|s)$

3) Soft Value Iteration

We will solve this minimization problem with dynamic programming!

$$\underset{\pi}{\arg\max} \; \mathbb{E}_{s,a \sim d^{\pi}_{\mu}}\left[r(s,a) - \log \pi(a|s)\right] = \underset{\pi}{\arg\max} \; \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t,a_t) - \log \pi_t(a_t|s_t) \,\middle|\, \begin{matrix} s_{t+1} \sim P(s_t,a_t) \\ a_t \sim \pi_t(s_t) \end{matrix}\right]$$

Initialize $V_H^*(s) = 0$

For $h = H-1, \ldots, 0,$

1) $$Q_h^*(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} V_{h+1}^*(s')$$

2) $$\pi_h^*(\cdot|s) = \underset{\rho \in \Delta(\mathcal{A})}{\arg\min} \sum_{a \in \mathcal{A}} \rho(a) Q_h^*(s,a) + \sum_{a \in \mathcal{A}} \rho(a) \underbrace{\log(\rho(a))}_{\text{entropy}}$$

constraint $\sum_a \rho(a) = 1$

(derivation below)
$$= \frac{\exp(Q_h^*(s,a))}{\sum_{a' \in \mathcal{A}} \exp(Q_h^*(s,a'))}$$

3) $$V_h^*(s) = \mathbb{E}_{a \sim \pi_h^*(\cdot|s)}\left[-\log \pi_h^*(a|s) + Q_h^*(s,a)\right]$$

(derivation below)
$$= \log\left(\sum_{a \in \mathcal{A}} \exp(Q_h^*(s,a))\right)$$

Contrast π^*, V^* with the classic RL solution!

$$\underset{a}{\text{softmax}} \; Q_h^*(s,a) \quad \text{vs.} \quad \max_a Q_h^*(s,a)$$

Deriving the softmax policy:

$$\max_{\rho} \min_{w} \underbrace{\sum_a \rho_a\left(Q_h^*(s,a) - \rho_a \log \rho_a\right) + w\left(\sum_a \rho_a - 1\right)}_{\mathcal{L}(\rho, w)}$$

$$\nabla_\rho \mathcal{L} = \nabla_\rho\left[\sum_a \rho_a\left(Q_h^*(s,a) + w\right) - \rho_a \log \rho_a\right]$$

$$\overset{=0}{\Rightarrow} Q_h^*(s,a) + w - \log \rho_a - 1 = 0 \quad \forall a$$

$$\rho_a = \exp(Q_h^*(s,a)) \exp(-1+w)$$

$$\nabla_w \mathcal{L} = \sum_a \rho_a - 1 \quad \Rightarrow \quad \sum_a \rho_a = 1$$

Combining the equations,

$$\sum_a \exp(Q_h^*(s,a))(\exp(-1+w)) = 1$$

$$\Rightarrow \exp(-1+w) = \frac{1}{\sum_a \exp(Q_h^*(s,a))} \quad \checkmark$$

Deriving the value:

$$V_h^*(s) = \mathbb{E}_{a \sim \pi_h^*(\cdot|s)}\left[-\log \pi_h^*(a|s) + Q_h^*(s,a)\right]$$

$$= \mathbb{E}_{a \sim \pi_h^*(\cdot|s)}\left[\cancel{-Q_h^*(s,a)} + \log\left(\sum_{a' \in \mathcal{A}} \exp(Q_h^*(s,a'))\right) + \cancel{Q_h^*(s,a)}\right]$$

(independent of a) $$= \log\left(\sum_{a' \in \mathcal{A}} \exp(Q_h^*(s,a'))\right) \mathbb{E}_{a \sim \pi_h^*(\cdot|s)}[1] \quad \checkmark$$