Maximum Entropy IRL
(continue)

Recap:

Constraint Optimization

$$\min_x f(x) \quad s.t.,\ g_1(x) = 0, g_2(x) = 0, \ldots, g_d(x) = 0$$

Recap:

Constraint Optimization

$$\min_x f(x) \quad s.t., g_1(x) = 0, g_2(x) = 0, \ldots, g_d(x) = 0$$

Introduce Lagrange multipliers $w \in \mathbb{R}^d$, we have:

$$\min_x \max_{w \in \mathbb{R}^d} \underbrace{f(x) + w^\top g(x)}_{:= \ell(x,w)}, \qquad (g(x) := [g_1(x), \ldots, g_d(x)]^\top)$$

Recap:

Constraint Optimization

$$\min_x f(x) \quad s.t., g_1(x) = 0, g_2(x) = 0, \ldots, g_d(x) = 0$$

Introduce Lagrange multipliers $w \in \mathbb{R}^d$, we have:

$$\min_x \max_{w \in \mathbb{R}^d} \underbrace{f(x) + w^\top g(x)}_{:=\ell(x,w)}, \quad (g(x) := [g_1(x), \ldots, g_d(x)]^\top)$$

We can optimize the dual $\max_{w \in \mathbb{R}^d} \min_x \ell(x, w)$ instead using iterative approach:

Recap:

Constraint Optimization

$$\min_x f(x) \quad s.t., g_1(x) = 0, g_2(x) = 0, \ldots, g_d(x) = 0$$

Introduce Lagrange multipliers $w \in \mathbb{R}^d$, we have:

$$\min_x \max_{w \in \mathbb{R}^d} \underbrace{f(x) + w^\top g(x)}_{:=\ell(x,w)}, \qquad (g(x) := [g_1(x), \ldots, g_d(x)]^\top)$$

We can optimize the dual $\max_{w \in \mathbb{R}^d} \min_x \ell(x, w)$ instead using iterative approach:

$$x^t = \arg\min_x \ell(x, w^t),$$

Best Response

Gradien Ascent

$$w^{t+1} = w_t + \eta \nabla_w \ell(x^t, w)\Big|_{w=w^t}$$

Recap on Inverse RL setting:

Recap on Inverse RL setting:

Q: we want to find a policy π such that $\mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)$

(Note **linear cost assumption** implies π is as good as $\pi^\star$)

But there are potentially many such policies…

$c(s,a) = (\theta^\star)^\top \phi(s,a)$

A

$\mathbb{E}_{s,a\sim\pi}[c(s,a)] = \mathbb{E}_{s,a\sim\pi^\star}[c(s,a)]$

Recap on Inverse RL setting:

Q: we want to find a policy π such that $\mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)$

(Note **linear cost assumption** implies π is as good as $\pi^\star$)

But there are potentially many such policies…

The principle of Maximum Entropy:

Find a policy π that maximizes some entropy while subject to the constraint:

Recap on Inverse RL setting:

Q: we want to find a policy π such that $\mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)$

(Note **linear cost assumption** implies π is as good as $\pi^\star$)

But there are potentially many such policies…

match

d_μ^π & $d_\mu^{\pi^\star}$

$\min_\pi \underline{\| d_\mu^\pi(\cdot,\cdot) - d_\mu^{\pi^\star}(\cdot,\cdot) \|_{TV}}$

Δ

The principle of Maximum Entropy:

Find a policy π that maximizes some entropy while subject to the constraint:

$$\max_\pi \mathbb{E}_{s\sim d_\mu^\pi}\left[\text{entropy}\left(\pi(\cdot\,|\,s)\right)\right]$$

$$s.t, \mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)$$

Recap on Inverse RL setting:

Q: we want to find a policy π such that $\mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)$

(Note **linear cost assumption** implies π is as good as $\pi^\star$)

But there are potentially many such policies…

The principle of Maximum Entropy:

Find a policy π that maximizes some entropy while subject to the constraint:

$$\max_\pi \mathbb{E}_{s\sim d_\mu^\pi}\left[\text{entropy}\left(\pi(\cdot\,|\,s)\right)\right]$$

$$s.t, \mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)$$

This can be estimated using expert data:

$$\sum_{i=1}^{N} \phi(s_i^\star, a_i^\star)/N$$

Plan for Today:

1. The Iterative Algorithm framework

2. How to compute best response: Soft Value Iteration (DP again)

3. The MaxEnt-IRL algorithm

Setting

Finite horizon MDP $\mathscr{M} = \{S, A, H, c, P, \mu, \pi^\star\}$

Setting

Finite horizon MDP $\mathscr{M} = \{S, A, H, c, P, \mu, \pi^\star\}$

We have a dataset $\mathscr{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d_\mu^{\pi^\star}$

Setting

Finite horizon MDP $\mathscr{M} = \{S, A, H, c, P, \mu, \pi^\star\}$

We have a dataset $\mathscr{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d_\mu^{\pi^\star}$

Key Assumption on cost:

$c(s,a) = \langle \theta^\star, \phi(s,a) \rangle$, **linear w.r.t feature** $\phi(s,a) \in \mathbb{R}^d$

found π s.t.

$$\mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s,a) = \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a)$$

1 $\Rightarrow$ $\theta^{\star T} \left[\mathbb{E}_\pi \phi(s,a) \right] = (\theta^\star)^T \left(\mathbb{E}_{\pi^\star} \phi(s,a) \right)$

2 $\Rightarrow$

$$\mathbb{E}_{s,a \sim d_\mu^\pi} \underbrace{(\theta^\star)^T \phi(s,a)}_{c(s,a)} = \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \underbrace{(\theta^\star)^T \phi(s,a)}_{c(s,a)}$$

Notation on Distributions

$\mathbb{P}_h^\pi(s, a; \mu)$: probability of visiting (s, a) at time step h following π

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

$$d_\mu^\pi(s) = \sum_a d_\mu^\pi(s, a)$$: average state distribution

A

Maximum Entropy Inverse RL:

Let's simplify the objective $\max_{\pi} \mathbb{E}_{s\sim d_\mu^\pi}\left[\underbrace{\text{entropy}(\pi(\cdot\,|\,s))}_{:= -\mathbb{E}_{a\sim\pi(\cdot|s)} \ln \pi(a|s)}\right]$:

Maximum Entropy Inverse RL:

Let's simplify the objective $\max_{\pi} \mathbb{E}_{s\sim d_\mu^\pi}\left[\text{entropy}(\pi(\cdot \mid s))\right]$:

$$\mathbb{E}_{s\sim d_\mu^\pi}\left[\text{entropy}(\pi(\cdot \mid s))\right] = -\mathbb{E}_{s\sim d_\mu^\pi}\underbrace{\mathbb{E}_{a\sim\pi(\cdot\mid s)} \ln \pi(a \mid s)}_{\text{Entropy}} = -\mathbb{E}_{s,a\sim d_\mu^\pi} \ln \pi(a \mid s)$$

Maximum Entropy Inverse RL:

Let's simplify the objective $\max_{\pi} \mathbb{E}_{s \sim d_\mu^\pi} \left[\text{entropy}(\pi(\cdot \mid s)) \right]$:

$$\mathbb{E}_{s \sim d_\mu^\pi} \left[\text{entropy}(\pi(\cdot \mid s)) \right] = -\mathbb{E}_{s \sim d_\mu^\pi} \mathbb{E}_{a \sim \pi(\cdot \mid s)} \ln \pi(a \mid s) = -\mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \mid s)$$

4

$$\arg\max_{\pi} \mathbb{E}_{s \sim d_\mu^\pi} \left[\text{entropy}(\pi(\cdot \mid s)) \right] = \arg\min_{\pi} \underline{\mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \mid s)}$$

Maximum Entropy Inverse RL formulation

We arrive at the following constraint optimization problem:

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_{\mu}^{\pi}} \ln \pi(a \mid s)$$

$$s.t, \mathbb{E}_{s,a\sim d_{\mu}^{\pi}}\phi(s,a) = \mathbb{E}_{s,a\sim d_{\mu}^{\pi^\star}}\phi(s,a)$$

Maximum Entropy Inverse RL formulation

We arrive at the following constraint optimization problem:

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi} \ln \pi(a \mid s)$$

$$s.t, \mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a) \in \mathbb{R}^d$$

Introduce the Lagrange multiplier $w \in \mathbb{R}^d$ (we have d many constraints), consider the max-min dual version:

Maximum Entropy Inverse RL formulation

We arrive at the following constraint optimization problem:

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \ln \pi(a \mid s)$$

$$s.t, \mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \phi(s,a) = \mathbb{E}_{s,a\sim d^{\pi^\star}_{\mu}} \phi(s,a)$$

d many constraints

$\phi \in \mathbb{R}^d$

Introduce the Lagrange multiplier $w \in \mathbb{R}^d$ (we have d many constraints), consider the max-min dual version:

$$\max_{w\in\mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \ln \pi(a \mid s) + w^\top \left(\mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}_{\mu}} \phi(s,a) \right)}_{:=\ell(\pi,w)}$$

Maximum Entropy Inverse RL formulation

We arrive at the following constraint optimization problem:

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi} \ln \pi(a\,|\,s)$$

$$s.t, \mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)$$

Introduce the Lagrange multiplier $w \in \mathbb{R}^d$ (we have d many constraints), consider the max-min dual version:

$$\max_{w\in\mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a\sim d_\mu^\pi} \ln \pi(a\,|\,s) + w^\top\left(\mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)\right)}_{:=\ell(\pi,w)}$$

Let's solve it by the iterative procedure!

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \,|\, s) + w^\top \left(\mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a) \right)}_{:= \ell(\pi, w)}$$

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \mid s) + w^\top \left(\mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a) \right)}_{:= \ell(\pi, w)}$$

Initialize $w^0 \in \mathbb{R}^d$

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \mid s) + w^\top \left(\mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a) \right)}_{:= \ell(\pi, w)}$$

Initialize $w^0 \in \mathbb{R}^d$

For $t = 0 \to T-1$

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \mid s) + w^\top \left(\mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a) \right)}_{:= \ell(\pi, w)}$$

Initialize $w^0 \in \mathbb{R}^d$

For $t = 0 \to T - 1$

$$\pi^t = \arg\min_{\pi} \mathbb{E}_{s,a \sim d_\mu^\pi} \left[(w^t)^\top \phi(x,a) + \ln \pi(a \mid s) \right]$$

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \mid s) + w^\top \left(\mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a) \right)}_{:= \ell(\pi, w)}$$

$$\min_{\pi} \ell(\pi, w^t) \iff \min_{\pi} \left[\mathbb{E}_{s,a \sim d_\mu^\pi} [(w^t)^\top \phi(s,a)] + \ln \pi(a|s) \right]$$

Initialize $w^0 \in \mathbb{R}^d$

For $t = 0 \to T - 1$

$$\pi^t = \arg\min_{\pi} \mathbb{E}_{s,a \sim d_\mu^\pi} \left[(w^t)^\top \phi(s, a) + \ln \pi(a \mid s) \right]$$

(# best response: $\pi^t = \arg\min_{\pi} \ell(\pi, w^t)$)

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w\in\mathbb{R}^d}\min_{\pi}\underbrace{\mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\ln\pi(a\,|\,s)+w^{\top}\left(\mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\phi(s,a)-\mathbb{E}_{s,a\sim d^{\pi^\star}_{\mu}}\phi(s,a)\right)}_{:=\ell(\pi,w)}$$

Initialize $w^0\in\mathbb{R}^d$

For $t=0\to T-1$

$$\pi^t=\arg\min_{\pi}\mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\left[(w^t)^{\top}\phi(x,a)+\ln\pi(a\,|\,s)\right]$$ (# best response: $\pi^t=\arg\min_{\pi}\ell(\pi,w^t)$)

$$w^{t+1}=w^t+\eta\underbrace{\left(\mathbb{E}_{s,a\sim d^{\pi^t}_{\mu}}\phi(s,a)-\mathbb{E}_{s,a\sim d^{\pi^\star}_{\mu}}\phi(s,a)\right)}_{:=\nabla_w\ell(\pi^t,w)|_{w=w^t}}$$

(# gradient update: $w^{t+1}=w^t+\eta\nabla_w\ell(\pi^t,w^t)$)

A

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \mid s) + w^\top \left(\mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a) \right)}_{:= \ell(\pi, w)}$$

Initialize $w^0 \in \mathbb{R}^d$

For $t = 0 \to T - 1$

$$\pi^t = \arg\min_{\pi} \mathbb{E}_{s,a \sim d_\mu^\pi} \left[(w^t)^\top \phi(x,a) + \ln \pi(a \mid s) \right]$$

(# best response: $\pi^t = \arg\min_{\pi} \ell(\pi, w^t)$)

$$w^{t+1} = w^t + \eta \left(\mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a) \right)$$

(# gradient update: $w^{t+1} = w^t + \eta \nabla_w \ell(\pi^t, w^t)$)

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

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a\sim d_\mu^\pi} \ln \pi(a \mid s) + w^\top \left(\mathbb{E}_{s,a\sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}} \phi(s,a) \right)}_{:=\ell(\pi,w)}$$

Initialize $w^0 \in \mathbb{R}^d$

This is like an RL problem w/ cost $c(s,a) := (w^t)^\top \phi(s,a)$, but w/ an additional $\ln \pi(a \mid s)$

For $t = 0 \to T-1$

$$\pi^t = \arg\min_\pi \mathbb{E}_{s,a\sim d_\mu^\pi} \left[(w^t)^\top \phi(x,a) + \ln \pi(a \mid s) \right]$$

(# best response: $\pi^t = \arg\min_\pi \ell(\pi, w^t)$)

$$w^{t+1} = w^t + \eta \left(\mathbb{E}_{s,a\sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}} \phi(s,a) \right)$$

(# gradient update: $w^{t+1} = w^t + \eta \nabla_w \ell(\pi^t, w^t)$)

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

$\max_{w} \min_{\pi} \ell(\pi, w)$

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \mid s) + w^\top \left(\mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a) \right)}_{:= \ell(\pi, w)}$$

Initialize $w^0 \in \mathbb{R}^d$

This is like an RL problem w/ cost $c(s,a) := (w^t)^\top \phi(s,a)$, but w/ an additional $\ln \pi(a \mid s)$

For $t = 0 \to T-1$

$$\pi^t = \arg\min_{\pi} \mathbb{E}_{s,a \sim d_\mu^\pi} \left[(w^t)^\top \phi(\overset{s}{\cancel{\pi}}, a) + \ln \pi(a \mid s) \right]$$

(# best response: $\pi^t = \arg\min_{\pi} \ell(\pi, w^t)$)

$$w^{t+1} = w^t + \eta \left(\mathbb{E}_{s,a \sim d_\mu^{\pi^t}} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^\star}} \phi(s,a) \right)$$

(# gradient update: $w^{t+1} = w^t + \eta \nabla_w \ell(\pi^t, w^t)$)

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

Plan for Today:

1. The Iterative Algorithm framework

2. How to compute best response: Soft Value Iteration (DP again)

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi}\left[c(x,a) + \ln \pi(a\,|\,s)\right]$$

cost function; in iter t: $c(s,a) = (w^t)^T \phi(s,a)$

3. The MaxEnt-IRL algorithm

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi}\left[c(s,a) + \ln \pi(a\,|\,s)\right]$$

$$\Leftrightarrow \arg\min_{\pi} \underbrace{\mathbb{E}_{s\sim d_\mu^\pi}\left[c(s,a)\right.}\ - \underbrace{\text{Entropy}(\pi(\cdot|s))}_{\Delta} \left.\right]$$

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \left[c(s,a) + \ln \pi(a \mid s) \right]$$

Soft Value Iteration for finite horizon MDP (Dynamic Programming again):

$$V^{*}_{h}(s) = \min_{\pi_h, \pi_{h+1}, \ldots \pi_{H-1}} \mathbb{E}\left[\sum_{t=h}^{H-1} c(s_t, a_t) + \ln \pi_t(a_t \mid s_t) \,\middle|\, s_h = s, a_t \sim \pi_t \right]$$

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi}\left[c(s,a) + \ln \pi(a\,|\,s)\right]$$

Soft Value Iteration for finite horizon MDP (Dynamic Programming again):

$$V_H^\star(s) = 0 \quad , \ \forall s$$

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \left[c(s,a) + \ln \pi(a \mid s) \right]$$

Soft Value Iteration for finite horizon MDP (Dynamic Programming again):

$$V^{\star}_{H}(s) = 0$$

$$Q^{\star}_{h}(s,a) = c(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)} V^{\star}_{h+1}(s')$$

what's the $\pi^{\star}_{h}(\cdot|s)$?? $\in \Delta(A)$

$$\min_{\rho\in\Delta(A)} \Big[\sum_{a} \rho(a) \big[c(s,a) + \ln \rho(a) + \mathbb{E}_{s'\sim P(s,a)} V^{\star}_{h+1}(s') \big] \Big]$$

$Q^{\star}_{h}(s,a)$

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi}\left[c(s,a) + \ln \pi(a \mid s)\right]$$

Soft Value Iteration for finite horizon MDP (Dynamic Programming again):

$$V_H^\star(s) = 0$$

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

$$\pi_h^\star(\cdot \mid s) = \arg\min_{\rho\in\Delta(A)} \left[\sum_a \rho(a) \underline{Q_h^\star(s,a)} + \underline{\sum_a \rho(a) \ln \rho(a)}\right]$$

$\Rightarrow$ - Entropy (ρ)

s.t. $\sum_a \rho(a) = 1$

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \left[c(s,a) + \ln \pi(a \mid s) \right]$$

Soft Value Iteration for finite horizon MDP (Dynamic Programming again):

$$V^{\star}_{H}(s) = 0$$

$$Q^{\star}_{h}(s,a) = c(s,a) + \mathbb{E}_{s'\sim P(\cdot\mid s,a)} V^{\star}_{h+1}(s')$$

$$\pi^{\star}_{h}(\cdot \mid s) = \arg\min_{\rho\in\Delta(A)} \left[\sum_{a} \rho(a) Q^{\star}_{h}(s,a) + \sum_{a} \rho(a) \ln \rho(a) \right]$$

Use Lagrange (we have a constraint here $\sum_{a} \rho(a) = 1$), we can show:

Maximum Entropy RL: Soft Value Iteration

$$\min_{\rho} \max_{w} \sum_a \rho(a) Q_h^\star(s,a) + \sum_a \rho(a)\ln\rho(a)$$

$$\underbrace{+\, w\left[\sum_a \rho(a) - 1\right]}_{\ell(\rho, w)}$$

L-m

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi}\left[c(s,a) + \ln \pi(a\,|\,s)\right]$$

Soft Value Iteration for finite horizon MDP (Dynamic Programming again):

$$\min_{\rho}\max_{w} \ell(\rho, w)$$

$$V_H^\star(s) = 0$$

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

$$\Rightarrow \nabla_\rho \ell(\rho, w) = 0$$

$$\nabla_w \ell(\rho, w) = 0$$

For s, at h:

$$\pi_h^\star(\cdot\,|\,s) = \arg\min_{\rho\in\Delta(A)}\left[\sum_a \rho(a) Q_h^\star(s,a) + \sum_a \rho(a)\ln\rho(a)\right]$$

$\in \Delta(A)$

$\Rightarrow$ solve for ρ

Use Lagrange (we have a constraint here $\sum_a \rho(a) = 1$), we can show:

$$\pi_h^\star(a\,|\,s) = \frac{\exp(-Q_h^\star(s,a))}{\sum_{a'}\exp(-Q_h^\star(s,a'))}$$

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\left[c(s,a) + \ln \pi(a \mid s)\right]$$

Soft Value Iteration for finite horizon MDP (Dynamic Programming again):

$$V^{\star}_{H}(s) = 0$$

$$Q^{\star}_{h}(s,a) = c(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)} V^{\star}_{h+1}(s')$$

② $Q^{*}_{h}(s,a) = Q^{*}_{h}(s,a')$

$\pi^{*}_{h}(\cdot \mid s) =$ Uniform (a, a')

$$\pi^{\star}_{h}(\cdot \mid s) = \arg\min_{\rho\in\Delta(A)} \left[\sum_{a} \rho(a) Q^{\star}_{h}(s,a) + \sum_{a} \rho(a) \ln \rho(a)\right]$$

Use Lagrange (we have a constraint here $\sum_{a} \rho(a) = 1$), we can show:

① $Q^{*}_{h}(s,a) = 0$

$Q^{*}_{h}(s,a') = +\infty$

$\Rightarrow \pi^{*}_{h}(a \mid s) = 1$

$$\pi^{\star}_{h}(a \mid s) = \frac{\exp(-Q^{\star}_{h}(s,a))}{\sum_{a'} \exp(-Q^{\star}_{h}(s,a'))}$$

classic RL

(contrast this to $\arg\min_{a} Q^{\star}(s,a)$)

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi}\left[c(s,a) + \ln \pi(a \mid s)\right]$$

$V^\star_{h+1}$

Soft Value Iteration for finite horizon MDP (continue)

$$Q_h^\star(s,a) = c(s,a) + \mathbb{E}_{s'\sim P(\cdot\mid s,a)} V_{h+1}^\star(s')$$

$$\pi_h^\star(a \mid s) = \frac{\exp(-Q_h^\star(s,a))}{\sum_{a'} \exp(-Q_h^\star(s,a'))}$$

$\Rightarrow V^\star_h(s)$

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \left[c(s,a) + \ln \pi(a \mid s)\right]$$

Soft Value Iteration for finite horizon MDP (continue)

$$Q^{\star}_h(s,a) = c(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)} V^{\star}_{h+1}(s')$$

$$\pi^{\star}_h(a \mid s) = \frac{\exp(-Q^{\star}_h(s,a))}{\sum_{a'} \exp(-Q^{\star}_h(s,a'))}$$

$$V^{\star}_h(s) = \mathbb{E}_{a\sim\pi^{\star}_h(\cdot|s)} \left[\underline{\ln \pi^{\star}_h(a \mid s)} + \underline{Q^{\star}_h(s,a)}\right] = \underline{-\ln\left(\sum_a \exp\left(-Q^{\star}_h(s,a)\right)\right)}$$

Maximum Entropy RL: Soft Value Iteration

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

Soft Value Iteration for finite horizon MDP (continue)

$$Q^\star_h(s,a) = c(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)} V^\star_{h+1}(s')$$

$$\pi^\star_h(a\,|\,s) = \frac{\exp(-Q^\star_h(s,a))}{\sum_{a'}\exp(-Q^\star_h(s,a'))} \checkmark$$

$$V^\star_h(s) = \mathbb{E}_{a\sim\pi^\star_h(\cdot|s)}\left[\ln \pi^\star_h(a\,|\,s) + Q^\star_h(s,a)\right] = -\ln\left(\underbrace{\sum_a \exp\left(-Q^\star_h(s,a)\right)}\right)$$

(contrast this to $\min_a Q^\star(s,a)$)

classic RL:

$$\left(V^*_h(s) = \min_a Q^*_h(s,a)\right)$$

① $Q^*_h(s,a) = 0$

$Q^*_h(s,a') = +\infty$

$\Rightarrow V^*_h(s) = Q^*_h(s,a) = 0$

② $Q^*_h(s,a) = Q^*_h(s,a')$

Maximum Entropy RL: Soft Value Iteration

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi}\left[c(s,a) + \ln \pi(a\,|\,s)\right]$$

Soft Value Iteration for finite horizon MDP (continue)

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

$$\pi_h^\star(a\,|\,s) = \frac{\exp(-Q_h^\star(s,a))}{\sum_{a'} \exp(-Q_h^\star(s,a'))}$$

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

(contrast this to $\min_a Q^\star(s,a)$)

$$Q_{h-1}^\star(s,a) = c(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)} V_h^\star(s')$$

$$\cdots$$

Plan for Today:

1. The Iterative Algorithm framework

finite Horizon

$\pi^* = \{\pi_0^*, \dots, \pi_{H-1}^*\}$

2. How to compute best response: Soft Value Iteration (DP again)

$$\arg\min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi}\left[c(x,a) + \ln \pi(a\,|\,s)\right]$$

$$\Longleftrightarrow \min_{\pi} \mathbb{E}_{s,a\sim d_\mu^\pi}\left[c(s,a) - \text{Entropy}(\pi(\cdot|s))\right]$$

3. The MaxEnt-IRL algorithm

Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \underbrace{\mathbb{E}_{s,a\sim d_\mu^\pi} \ln \pi(a\,|\,s) + w^\top \left(\mathbb{E}_{s,a\sim d_\mu^\pi} \phi(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}} \phi(s,a) \right)}_{:=\ell(\pi,w)}$$

Initialize $w^0 \in \mathbb{R}^d$ $\{\pi_0 \cdots \pi_{H-1}\}$

For $t = 0 \to \cancel{T}\,\infty - 1$ DP

π^t = soft-VI $\left(c(s,a) := (w^t)^\top \phi(s,a)\right)$ (# best response: $\pi^t = \arg\min_\pi \ell(\pi, w^t)$)

$\pi(H)$

$w^{t+1} = w^t + \eta \left(\mathbb{E}_{s,a\sim d_\mu^{\pi^t}} \phi(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}} \phi(s,a) \right)$

≈ 0

Return $\bar{\pi}$ = Uniform$(\pi^0, \ldots, \pi^{T-1})$ (# gradient update: $w^{t+1} = w^t + \eta \nabla_w \ell(\pi^t, w^t)$)

$\left\| \mathbb{E}_{s,a\sim d_\mu^{\pi^t}} \phi(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}} \phi(s,a) \right\|_2 \le \delta$

Maximum Entropy IRL: Calculate Trajectory Likelihood

Maximum Entropy IRL: Calculate Trajectory Likelihood

Given a trajectory $\tau = \{s_0, a_0, \ldots, s_{H-1}, a_{H-1}\}$

What's the likelihood of τ being generated by expert? $\bar{\pi}$

Maximum Entropy IRL: Calculate Trajectory Likelihood

Given a trajectory $\tau = \{s_0, a_0, \ldots, s_{H-1}, a_{H-1}\}$

What's the likelihood of τ being generated by expert?

$$\ln\left(\rho^{\bar{\pi}}(\tau)\right) = \sum_{h=0}^{H-1}\left[\underbrace{\ln P(s_{h+1} \mid s_h, a_h)}_{\text{known}} + \underbrace{\ln \bar{\pi}(a_h \mid s_h)}_{\text{our policy from MaxEnt-IRL}}\right]$$

$\mu(s_0)\bar{\pi}(a_0|s_0)P(s_1|s_0,a_0)\cdots$

Maximum Entropy IRL: Calculate Trajectory Likelihood

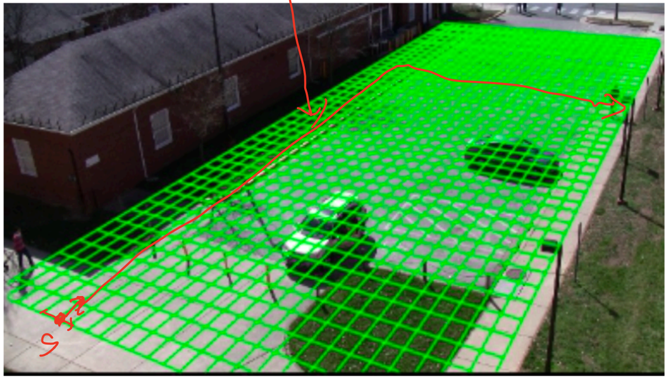
Given a trajectory $\tau = \{s_0, a_0, \ldots, s_{H-1}, a_{H-1}\}$

What's the likelihood of τ being generated by expert?

$$\ln\left(\rho^{\bar{\pi}}(\tau)\right) = \sum_{h=0}^{H-1}\left[\ln P(s_{h+1} \mid s_h, a_h) + \ln \bar{\pi}(a_h \mid s_h)\right]$$

$P(s_{h+1} \mid s_h, a_h)$ is deterministic

State space: grid,
action space: 4 actions



$\Rightarrow$

Summary for Today:

1. Maximum Entropy IRL framework

$$\max_{\pi} \mathbb{E}_{s\sim d_\mu^\pi}\left[\text{entropy}\left(\pi(\cdot\,|\,s)\right)\right]$$

$$s.t, \underbrace{\mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)}$$

feature Matching

Summary for Today:

1. Maximum Entropy IRL framework

$$\max_{\pi} \mathbb{E}_{s\sim d_\mu^\pi}\left[\text{entropy}\left(\pi(\cdot\,|\,s)\right)\right]$$

$$s.t,\ \mathbb{E}_{s,a\sim d_\mu^\pi}\phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^\star}}\phi(s,a)$$

← Best Response ← soft Value Iteration

2. Inside MaxEnt-IRL, we perform Maximum Entropy RL:

Note

$$\min_{\pi_0,\ldots,\pi_{H-1}} \mathbb{E}\left[\sum_{h=0}^{H-1}\left(\underline{c(s_h,a_h)} - \underline{\text{entropy}(\pi_h(\cdot\,|\,s_h))}\right) \,|\, s_0\sim\mu, a_h\sim\pi_h(\cdot\,|\,s_h)\right]$$