Maximum Entropy IRL
(continue)
Recap:

**Constraint Optimization**

$$\min_{x} f(x) \quad s.t., \quad g_{1}(x) = 0, g_{2}(x) = 0, \ldots, g_{d}(x) = 0$$
Recap:

**Constraint Optimization**

\[
\min_{x} f(x) \quad \text{s.t.} \quad g_1(x) = 0, g_2(x) = 0, \ldots, g_d(x) = 0
\]

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

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

\[ := \ell(x, w) \]
Recap:

**Constraint Optimization**

\[
\min_x f(x) \quad \text{s.t.} \quad g_1(x) = 0, g_2(x) = 0, \ldots, g_d(x) = 0
\]

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

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

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

**Constraint Optimization**

\[
\min_x f(x) \quad \text{s.t.} \quad g_1(x) = 0, g_2(x) = 0, \ldots, g_d(x) = 0
\]

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

\[
\min_x \max_{w \in \mathbb{R}^d} \left( f(x) + w^T g(x), (g(x) := [g_1(x), \ldots, g_d(x)]^T) \right)
\]

\( \equiv \ell(x, w) \)

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

\[
\begin{align*}
\text{Besse Response} & \quad x^t = \arg \min_x \ell(x, w^t), \\
\text{Gradient Ascent} & \quad w^{t+1} = w_t + \eta \nabla_w \ell(x^t, w) \bigg|_{w=w^t}
\end{align*}
\]
Recap on Inverse RL setting:
Recap on Inverse RL setting:

Q: we want to find a policy \( \pi \) such that

\[
\mathbb{E}_{s,a \sim d_\pi} \phi(s,a) = \mathbb{E}_{s,a \sim d_\pi^*} \phi(s,a)
\]

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

But there are potentially many such policies…
Recap on Inverse RL setting:

Q: we want to find a policy $\pi$ such that
\[ \mathbb{E}_{s,a \sim d_\pi} \phi(s, a) = \mathbb{E}_{s,a \sim d_{\pi^*}} \phi(s, a) \]
(Note linear cost assumption implies $\pi$ is as good as $\pi^*$)
But there are potentially many such policies...

The principle of Maximum Entropy:
Find a policy $\pi$ that maximizes some entropy while subject to the constraint:
Recap on Inverse RL setting:

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

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

But there are potentially many such policies…

The principle of Maximum Entropy:

Find a policy $\pi$ that maximizes some entropy while subject to the constraint:

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

$s \cdot t, \mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s, a) = \mathbb{E}_{s,a \sim d_\mu^{\pi^*}} \phi(s, a)$
Recap on Inverse RL setting:

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

(Note linear cost assumption implies $\pi$ is as good as $\pi^*$)
But there are potentially many such policies...

The principle of Maximum Entropy:
Find a policy $\pi$ that maximizes some entropy while subject to the constraint:

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

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

This can be estimated using expert data:
$$\sum_{i=1}^{N} \phi(s_i^*, a_i^*)/N$$
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 $\mathcal{M} = \{S, A, H, c, P, \mu, \pi^*\}$
Setting

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

We have a dataset $\mathcal{D} = (s_i^*, a_i^*)_{i=1}^M \sim \mathcal{D}_\mu^{\pi^*}$
Setting

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

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

**Key Assumption on cost:**

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

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

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

\[
d_{\mu}^{\pi}(s) = \sum_{a} d_{\mu}^{\pi}(s, a): \text{average state distribution}
\]
Maximum Entropy Inverse RL:

Let’s simplify the objective \(\max_{\pi} \mathbb{E}_{s \sim d_\mu} [\text{entropy}(\pi(\cdot|s))]:\)

\[= - \mathbb{E}_{a \sim \pi(\cdot|s)} \ln \pi(a|s)\]
Maximum Entropy Inverse RL:

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

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

Let’s simplify the objective \( \max \mathbb{E}_{s \sim d_\mu} \left[ \text{entropy}(\pi( \cdot | s)) \right] \):

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

\[
\arg \max \mathbb{E}_{s \sim d_\mu} \left[ \text{entropy}(\pi( \cdot | s)) \right] = \arg \min \mathbb{E}_{s,a \sim d_\mu} \ln \pi(a | s)
\]
Maximum Entropy Inverse RL formulation

We arrive at the following constraint optimization problem:

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

$$s.t., \mathbb{E}_{s,a \sim d_\mu} \phi(s, a) = \mathbb{E}_{s,a \sim d_\mu^*} \phi(s, a)$$
Maximum Entropy Inverse RL formulation

We arrive at the following constraint optimization problem:

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

subject to,

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

Introduce the Lagrange multiplier $\lambda \in \mathbb{R}^d$ (we have $d$ many constraints), consider the max-min dual version:
Maximum Entropy Inverse RL formulation

We arrive at the following constraint optimization problem:

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

subject to,

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

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

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a \mid s) + w^T \left( \mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s, a) - \mathbb{E}_{s,a \sim d_\mu^{\pi*}} \phi(s, a) \right)$$

$$:= \ell(\pi, w)$$
**Maximum Entropy Inverse RL formulation**

We arrive at the following constraint optimization problem:

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

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

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

\[
\max_{w \in \mathbb{R}^d} \min_{\pi} \mathbb{E}_{s,a \sim d_{\mu}^\pi} \ln \pi(a \mid s) + w^T \left( \mathbb{E}_{s,a \sim d_{\mu}^\pi} \phi(s, a) - \mathbb{E}_{s,a \sim d_{\mu}^{\pi*}} \phi(s, a) \right)
\]

\[:= \ell(\pi, w)\]

Let’s solve it by the iterative procedure!
Maximum Entropy Inverse RL Algorithm framework

\[
\max_{w \in \mathbb{R}^d} \min_{\pi} \mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a | s) + w^T \left( \mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s, a) - \mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s, a) \right)
\]

:= \ell(\pi, w)
Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_\pi \sum_{s,a \sim \mu} \ln \pi(a | s) + w^T \left( \mathbb{E}_{s,a \sim \mu} \phi(s,a) - \mathbb{E}_{s,a \sim \mu} \pi \phi(s,a) \right)$$

:= \ell(\pi,w)

Initialize $w^0 \in \mathbb{R}^d$
Maximum Entropy Inverse RL Algorithm framework

\[
\max_{w \in \mathbb{R}^d} \min_{\pi} \mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a | s) + w^T \left( \mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s, a) - \mathbb{E}_{s,a \sim d_\mu^*} \phi(s, a) \right)
\]

\[
:= \ell(\pi, w)
\]

Initialize \( w^0 \in \mathbb{R}^d \)

For \( t = 0 \rightarrow T - 1 \)

Maximum Entropy Inverse RL Algorithm framework

\[
\max_{w \in \mathbb{R}^d} \min_{\pi} \mathbb{E}_{s,a \sim d^\pi} \ln \pi(a | s) + w^T \left( \mathbb{E}_{s,a \sim d^\pi} \phi(s, a) - \mathbb{E}_{s,a \sim d^\pi} \phi(s, a) \right) \\
:= \ell(\pi, w)
\]

Initialize \( w^0 \in \mathbb{R}^d \)

For \( t = 0 \rightarrow T - 1 \)

\[
\pi^t = \arg \min_{\pi} \mathbb{E}_{s,a \sim d^\pi} \left[ (w^t)^T \phi(x, a) + \ln \pi(a | s) \right]
\]
Maximum Entropy Inverse RL Algorithm framework

\[
\max_{w \in \mathbb{R}^d} \min_{\pi} \mathbb{E}_{s,a \sim d_{\mu}} \ln \pi(a \mid s) + w^T \left( \mathbb{E}_{s,a \sim d_{\mu}} \phi(s,a) - \mathbb{E}_{s,a \sim d_{\mu}^*} \phi(s,a) \right)
\]

\[= \ell(\pi, w) \]

Initialize \( w^0 \in \mathbb{R}^d \)

For \( t = 0 \rightarrow T - 1 \)

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

(# best response: \( \pi^t = \arg \min_{\pi} \ell(\pi, w^t) \))
Maximum Entropy Inverse RL Algorithm framework

$$\max_{w \in \mathbb{R}^d} \min_{\pi} \mathbb{E}_{s,a \sim d_\mu} \ln \pi(a \mid s) + w^T \left( \mathbb{E}_{s,a \sim d_\mu} \phi(s, a) - \mathbb{E}_{s,a \sim d_\mu^*} \phi(s, a) \right)$$

$$:= \ell(\pi, w)$$

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

For $t = 0 \rightarrow T - 1$

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

$$w^{t+1} = w^t + \eta \left( \mathbb{E}_{s,a \sim d_\mu} \phi(s, a) - \mathbb{E}_{s,a \sim d_\mu^*} \phi(s, a) \right)$$  (# gradient update: $w^{t+1} = w^t + \eta \nabla_w \ell(\pi^t, w^t)$)
Maximum Entropy Inverse RL Algorithm framework

\[ \max_{\pi} \min_{w} \mathbb{E}_{s,a \sim d_\mu^t} \ln \pi(a \mid s) + w^T \left( \mathbb{E}_{s,a \sim d_\mu^t} \phi(s,a) - \mathbb{E}_{s,a \sim d_\mu^t} \phi(s,a) \right) =: \ell(\pi,w) \]

Initialize \( w^0 \in \mathbb{R}^d \)

For \( t = 0 \rightarrow T - 1 \)

\[ \pi^t = \arg \min_{\pi} \mathbb{E}_{s,a \sim d_\mu^t} \left[ (w^t)^T \phi(x,a) + \ln \pi(a \mid s) \right] \quad \text{(# best response: } \pi^t = \arg \min_{\pi} \ell(\pi, w^t)) \]

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

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

Return \( \tilde{\pi} = \text{Uniform}(\pi^0, \ldots, \pi^{T-1}) \)
Maximum Entropy Inverse RL Algorithm framework

\[
\max_{w \in \mathbb{R}^d} \min_{\pi} \mathbb{E}_{s,a \sim d^\pi_\mu} \ln \pi(a \mid s) + w^T \left( \mathbb{E}_{s,a \sim d^\pi_\mu} \phi(s,a) - \mathbb{E}_{s,a \sim d^\pi_\mu} \phi(s,a) \right) \\
:= \ell(\pi, w)
\]

Initialize \( w^0 \in \mathbb{R}^d \)

For \( t = 0 \to T - 1 \)

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

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

\[
w^{t+1} = w^t + \eta \left( \mathbb{E}_{s,a \sim d^\pi_\mu} \phi(s,a) - \mathbb{E}_{s,a \sim d^\pi_\mu} \phi(s,a) \right)
\]

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

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

This is like an RL problem w/ cost \( c(s,a) := (w^t)^T \phi(s,a) \), but w/ an additional \( \ln \pi(a \mid s) \)
Maximum Entropy Inverse RL Algorithm framework

\[
\max_{w \in \mathbb{R}^d} \min_{\pi} \mathbb{E}_{s,a \sim d_\mu^\pi} \ln \pi(a | s) + w^\top \left( \mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s, a) - \mathbb{E}_{s,a \sim d_\mu^\pi} \phi(s, a) \right)
\]

\( := \ell(\pi, w) \)

Initialize \( w^0 \in \mathbb{R}^d \)

For \( t = 0 \rightarrow T - 1 \)

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

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

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

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

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

1. The Iterative Algorithm framework

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

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

3. The MaxEnt-IRL algorithm
Maximum Entropy RL: Soft Value Iteration

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

$$\iff \arg\min_{\pi} \mathbb{E}_{s,a \sim \rho} \left[ c(s, a) - \left\langle \frac{\text{Entropy} (\pi(\cdot \mid s))}{\Delta} \right\rangle \right]$$
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):

$$\bar{V}_h^y(s) = \min_{\pi_h \pi_{h+1} \cdots \pi_{H-1}} \mathbb{E} \left[ \sum_{t=h}^{H-1} c(S_t, a_t) + \ln \pi_h (a_t | S_t) \left| S_h = s \right. \right]_{a_t \sim \pi_t}$$
Maximum Entropy RL: Soft Value Iteration

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

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

\[ V_H^*(s) = 0 \quad \forall s \]
Maximum Entropy RL: Soft Value Iteration

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

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

$$V^*_h(s) = 0$$

$$Q^*_h(s, a) = c(s, a) + \mathbb{E}_{s' \sim \pi(s,a)} V^*_{h+1}(s')$$

$$\min_{\pi \in \Delta(A)} \left[ \sum_{a} \pi(a) \left[ c(s,a) + \ln \pi(a | s) + \mathbb{E}_{s' \sim \pi(s,a)} V^*_{h+1}(s') \right] \right]$$
Maximum Entropy RL: Soft Value Iteration

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

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

$$V_h^*(s) = 0$$

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

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

$$\sum_a \rho(a) = 1$$
Maximum Entropy RL: Soft Value Iteration

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

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

$$V^*_h(s) = 0$$

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

$$\pi^*_h(\cdot \mid s) = \arg\min_{\rho \in \Delta(A)} \left[ \sum_a \rho(a) Q^*_h(s, a) + \sum_a \rho(a) \ln \rho(a) \right]$$

Use Lagrange (we have a constraint here $\sum_a \rho(a) = 1$), we can show:
Maximum Entropy RL: Soft Value Iteration

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

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

$$V^*_H(s) = 0$$

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

For $s,a$ at $h$:

$$\pi^*_h(· | s) = \arg \min_{\rho \in A} \left[ \sum_a \rho(a) Q^*_h(s,a) + \sum_a \rho(a) \ln \rho(a) \right]$$

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

$$\pi^*_h(a \mid s) = \frac{\exp(-Q^*_h(s,a))}{\sum_{a'} \exp(-Q^*_h(s,a'))}$$
Maximum Entropy RL: Soft Value Iteration

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

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

$$V^*_h(s) = 0$$

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

$$\pi^*_h(\cdot | s) = \arg \min_{\rho \in \Delta(A)} \left[ \sum_a \rho(a) Q^*_h(s,a) + \sum_a \rho(a) \ln \rho(a) \right]$$

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

$$\pi^*_h(a | s) = \frac{\exp(-Q^*_h(s,a))}{\sum_a \exp(-Q^*_h(s,a'))}$$

(contrast this to $\arg \min_a Q^*(s,a)$)
Maximum Entropy RL: Soft Value Iteration

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

Soft Value Iteration for finite horizon MDP (continue)

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

\[
\pi_h^*(a | s) = \frac{\exp(-Q_h^*(s, a))}{\sum_{a'} \exp(-Q_h^*(s, a'))}
\]

\[
\Rightarrow V_h^*(s)
\]
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)

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

\[
\pi_h^*(a \mid s) = \frac{\exp(-Q_h^*(s, a))}{\sum_{a'} \exp(-Q_h^*(s, a'))}
\]

\[
V_h^*(s) = \mathbb{E}_{a \sim \pi_h^*(\cdot \mid s)} \left[ \ln \pi_h^*(a \mid s) + Q_h^*(s, a) \right] = -\ln \left( \sum_a \exp \left( -Q_h^*(s, a) \right) \right)
\]
Maximum Entropy RL: Soft Value Iteration

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

Soft Value Iteration for finite horizon MDP (continue)

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

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

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

(contrast this to)

$$\min_{\pi} \mathcal{R}_{\|}$$

$$V^{*}_{h}(s) = \min_{a} Q^{*}_{h}(s, a)$$
Maximum Entropy RL: Soft Value Iteration

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

Soft Value Iteration for finite horizon MDP (continue)

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

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

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

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

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

...
Plan for Today:

1. The Iterative Algorithm framework

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

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

3. The MaxEnt-IRL algorithm
Maximum Entropy Inverse RL Algorithm framework

\[
\max_{\pi} \min_{w \in \mathbb{R}^d} \mathbb{E}_{s,a \sim d^\pi} \ln \pi(a | s) + w^T \left( \mathbb{E}_{s,a \sim d^\pi} \phi(s, a) - \mathbb{E}_{s,a \sim d^w} \phi(s, a) \right)
\]

\[\ell(\pi, w) := \ldots\]

**Initialize** \(w^0 \in \mathbb{R}^d\)

**For** \(t = 0 \rightarrow T - 1\)

\[\pi^t = \text{soft-VI} \left( c(s, a) := (w^t)^T \phi(s, a) \right)\]

\[w^{t+1} = w^t + \eta \left( \mathbb{E}_{s,a \sim d^\pi} \phi(s, a) - \mathbb{E}_{s,a \sim d^w} \phi(s, a) \right)\]

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

\[\| \mathbb{E}_{s,a \sim d^\pi} \phi(s, a) - \mathbb{E}_{s,a \sim d^w} \phi(s, a) \|_2 \leq \varepsilon\]
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 \( \tau \) being generated by expert?
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 $\tau$ being generated by expert?

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

Known $\pi$ or policy from MaxEnt-IRL

$\mu(s_0) \bar{\pi}(a_0 | s_0) P(s_1 | s_0, a_0)$ -
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 $\tau$ being generated by expert?

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

State space: grid, action space: 4 actions

$P(s_{h+1} | s_h, a_h)$ is deterministic
Summary for Today:

1. Maximum Entropy IRL framework

\[
\max_{\pi} \mathbb{E}_{s \sim d^\pi_{\mu}} \left[ \text{entropy} \left( \pi(\cdot \mid s) \right) \right] \\
\text{s.t. } \mathbb{E}_{s,a \sim d^\pi_{\mu}} \phi(s,a) = \mathbb{E}_{s,a \sim d^{\pi*}_{\mu}} \phi(s,a)
\]
Summary for Today:

1. Maximum Entropy IRL framework

\[
\max_{\pi} \mathbb{E}_{s \sim d_\mu} \left[ \text{entropy} \left( \pi(\cdot | s) \right) \right]
\]
\[
s.t., \mathbb{E}_{s,a \sim d_\mu} \phi(s,a) = \mathbb{E}_{s,a \sim d_\mu^*} \phi(s,a)
\]

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

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