Maximum Entropy Inverse Reinforcement Learning

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

Announcements

Project Presentation (Dec 8th and 10th):

Please sign up time slots
(see Piazza post for more details)

Recap

Offline IL Setting:

Recap

Offline IL Setting:

Ground truth reward $r(s, a) \in [0,1]$ is unknown; assume expert is a near optimal policy $\pi^\star$

Recap

Offline IL Setting:

Ground truth reward $r(s, a) \in [0,1]$ is unknown; assume expert is a near optimal policy $\pi^\star$

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

Recap

Offline IL Setting:

Ground truth reward $r(s, a) \in [0,1]$ is unknown; assume expert is a near optimal policy $\pi^\star$

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

Offline IL Algorithm: Behavior Cloning (Maximum Likelihood)

Recap

Offline IL Setting:

Ground truth reward $r(s, a) \in [0,1]$ is unknown; assume expert is a near optimal policy $\pi^\star$

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

Offline IL Algorithm: Behavior Cloning (Maximum Likelihood)

$$\widehat{\pi} = \arg\max_{\pi \in \Pi} \sum_{i=1}^{M} \ln \pi(a_i^\star \mid s_i^\star)$$

Recap

Hybrid IL Setting:

Recap

Hybrid IL Setting:

We have a dataset $\mathscr{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d^{\pi^\star}$ **<u>and</u>** access to transition $P(\cdot \mid s, a), \forall s, a$

Recap

Hybrid IL Setting:

We have a dataset $\mathcal{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d^{\pi^\star}$ **and** access to transition $P(\cdot \mid s, a), \forall s, a$

Hybrid IL Algorithm: Distribution Matching
(Statistically efficient, but not computationally)

Start from TV: $\mathcal{F} = \{ f : \|f\|_\infty \le 1 \}$

$\forall \pi, \pi' \in \Pi$

$f_{\pi,\pi'} = \arg\max_{f \in \mathcal{F}} \mathbb{E}_{s,a \sim d_\pi} f(s,a) - \mathbb{E}_{s,a \sim d_{\pi'}} f(s,a)$

Recap

$\widetilde{\mathcal{F}} = \{ f_{\pi,\pi'} : \pi, \pi' \in \Pi \}$

Hybrid IL Setting:

We have a dataset $\mathcal{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d^{\pi^\star}$ **and** access to transition $P(\cdot \mid s, a), \forall s, a$

$\mathcal{F} \to \widetilde{\mathcal{F}} \quad |\widetilde{\mathcal{F}}| \le |\Pi|^2$

Hybrid IL Algorithm: Distribution Matching
(Statistically efficient, but not computationally)

$$\widehat{\pi} := \arg\min_{\pi \in \Pi} \left[\underbrace{\max_{f \in \widetilde{\mathcal{F}}} \left[\mathbb{E}_{s,a \sim d^\pi} f(s,a) - \frac{1}{M} \sum_{i=1}^M f(s_i^\star, a_i^\star) \right]}_{IPM(\widetilde{\mathcal{F}})} \right]$$

Today: Hybrid Setting

Algorithm: Maximum Entropy Inverse Reinforcement Learning

Running Example: Human trajectory forecasting

[Kitani, et al, ECCV 12]

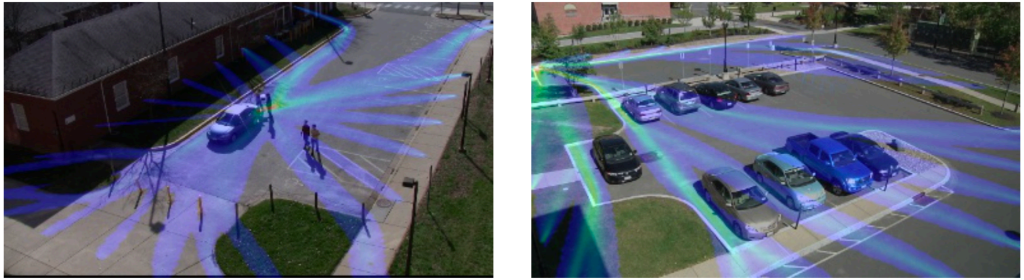


Fig. 1. Given a single pedestrian detection, our proposed approach forecasts plausible paths and destinations from noisy vision-input

Running Example: Human trajectory forecasting

[Kitani, et al, ECCV 12]

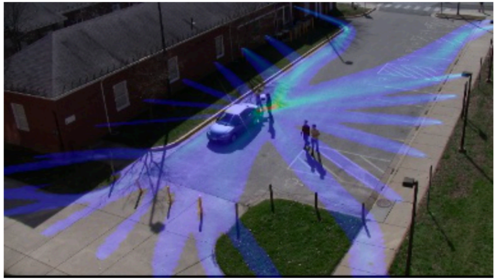
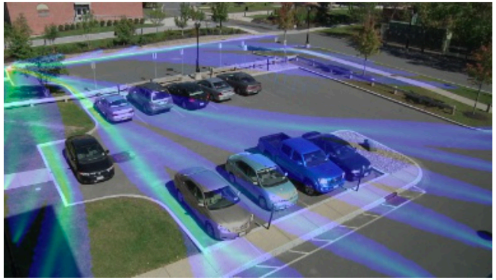



Fig. 1. Given a single pedestrian detection, our proposed approach forecasts plausible paths and destinations from noisy vision-input

High-level assumptions:

(1) Experts may have some cost function regarding walking in their mind

(2) Experts are (approximately) optimizing the cost function

Setting

Finite horizon MDP $\mathscr{M} = \{S, A, H, c, P, \mu_0, \pi^\star\}$

Setting

Finite horizon MDP $\mathscr{M} = \{S, A, H, c, P, \mu_0, \pi^\star\}$

(1) Ground truth cost $c(s, a)$ is unknown;

(2) assume expert is the optimal policy $\pi^\star$ of the cost c

(3) transition P is known

Setting

Finite horizon MDP $\mathscr{M} = \{S, A, H, c, P, \mu_0, \pi^\star\}$

(1) Ground truth cost $c(s, a)$ is unknown;

(2) assume expert is the optimal policy $\pi^\star$ of the cost c

(3) transition P is known

hybrid setting

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

Setting

Finite horizon MDP $\mathscr{M} = \{S, A, H, c, P, \mu_0, \pi^\star\}$

(1) Ground truth cost $c(s, a)$ is unknown;
(2) assume expert is the optimal policy $\pi^\star$ of the cost c
(3) transition P is known

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

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

Notation on Distributions

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

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

$\tau = \{ s_0\, a_0\, s_1\, a_1 \ldots\, s_{H-1}, a_{H-1}, s_H \}$

$$\rho^\pi(\tau) := \mu_0(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)\ldots\pi(a_{H-1} \mid s_{H-1})P(s_H \mid s_{H-1}, a_{H-1})$$:

Likelihood of the trajectory τ under π

Δ

Running Example: Define feature map

Key Assumption on cost:

$c(s, a) = \langle \theta^\star, \phi(s, a) \rangle$, **linear wrt feature** $\phi(s, a)$

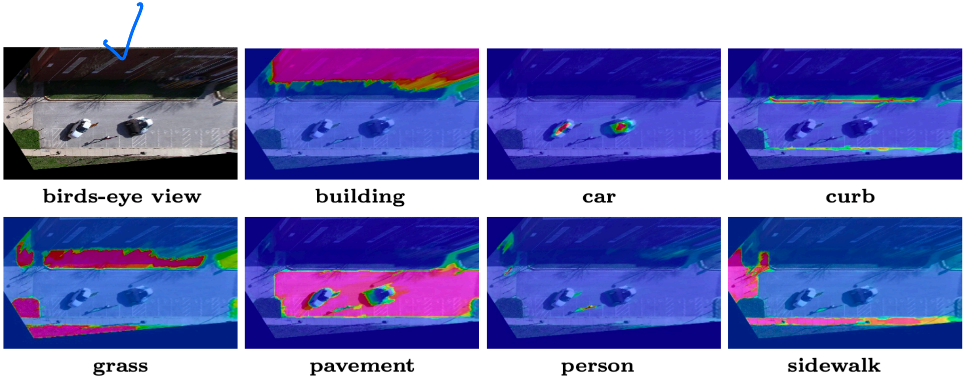


Fig. 4. Classifier feature response maps. Top left is the original image.

Running Example: Define feature map

Key Assumption on cost:

$c(s, a) = \langle \theta^\star, \phi(s, a) \rangle$, **linear wrt feature** $\phi(s, a)$

State s: pixel or a group of neighboring pixels in image)

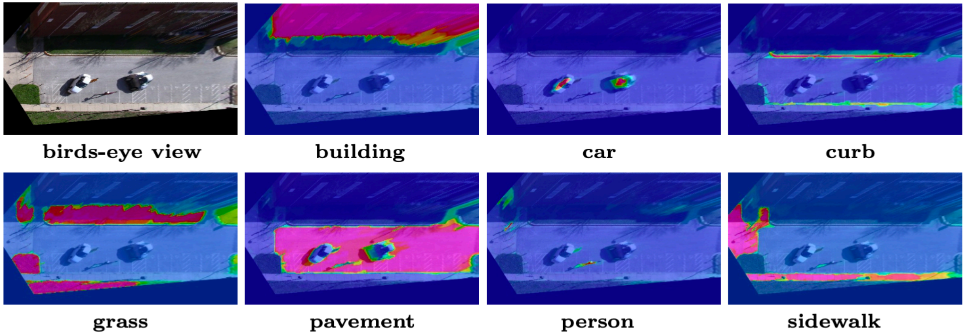


Fig. 4. Classifier feature response maps. Top left is the original image.

Running Example: Define feature map

Key Assumption on cost:

$c(s,a) = \langle \theta^\star, \phi(s,a) \rangle$, **linear wrt feature** $\phi(s,a)$

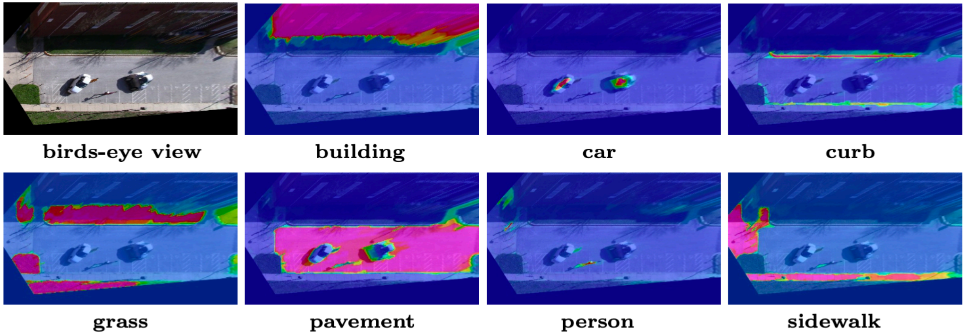


Fig. 4. Classifier feature response maps. Top left is the original image.

State s: pixel or a group of neighboring pixels in image)

$$\phi(s,a) = \begin{bmatrix} \mathbb{P}(\text{pixels being building}) \\ \mathbb{P}(\text{pixels being grass}) \\ \mathbb{P}(\text{pixels being sidewalk}) \\ \mathbb{P}(\text{pixels being car}) \\ \cdots \end{bmatrix}$$

Running Example: Define feature map

Key Assumption on cost:

$c(s,a) = \langle \theta^\star, \phi(s,a) \rangle$, **linear wrt feature** $\phi(s,a)$

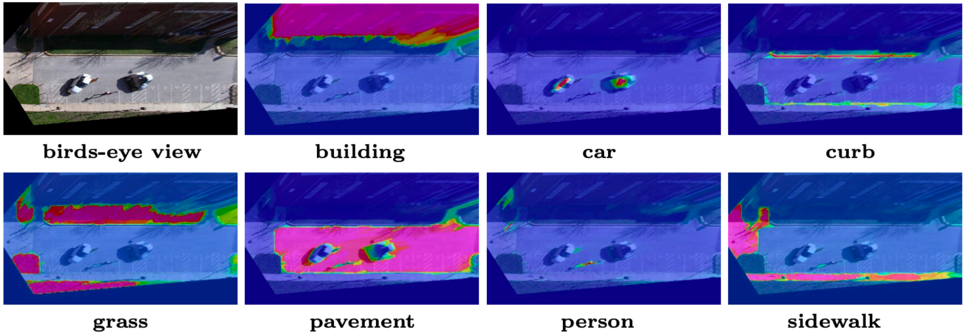


Fig. 4. Classifier feature response maps. Top left is the original image.

State s: pixel or a group of neighboring pixels in image)

$$\phi(s,a) = \begin{bmatrix} \mathbb{P}(\text{pixels being building}) \\ \mathbb{P}(\text{pixels being grass}) \\ \mathbb{P}(\text{pixels being sidewalk}) \\ \mathbb{P}(\text{pixels being car}) \\ \cdots \end{bmatrix}$$

Maybe colliding with cars or buildings has **high** cost, but walking on sideway or grass has **low** cost

Detour: Principle of Maximum Entropy

We want to find a distribution whose mean and covariance matrix equal to μ, Σ,
but there are infinitely many such distributions…

Detour: Principle of Maximum Entropy

We want to find a distribution whose mean and covariance matrix equal to μ, Σ,
but there are infinitely many such distributions…

Principle of Maximum Entropy:
Entropy Maximization subject to Moment Matching constraints

Detour: Principle of Maximum Entropy

We want to find a distribution whose mean and covariance matrix equal to μ, Σ,
but there are infinitely many such distributions…

Principle of Maximum Entropy:
Entropy Maximization subject to Moment Matching constraints

$$\max_{Q \in \Delta(X)} \text{entropy}(Q), \quad \text{s.t.,} \quad \mathbb{E}_{x \sim Q}[x] = \mu, \quad \mathbb{E}_{x \sim Q}[xx^\top] = \Sigma + \mu\mu^\top$$

Δ

Q^* ??

Detour: Principle of Maximum Entropy

We want to find a distribution whose mean and covariance matrix equal to μ, Σ,
but there are infinitely many such distributions…

Principle of Maximum Entropy:
Entropy Maximization subject to Moment Matching constraints

$$\max_{Q \in \Delta(X)} \text{entropy}(Q), \quad \text{s.t.,} \quad \mathbb{E}_{x\sim Q}[x] = \mu, \quad \mathbb{E}_{x\sim Q}[xx^\top] = \Sigma + \mu\mu^\top$$

Solution: $Q^\star = \mathcal{N}(\mu, \Sigma)$ exp(quadratic form)
(proof: use Lagrange multiplier)

Maximum Entropy Inverse RL:

Maximum Entropy Inverse RL:

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

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

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

But there are potentially many such policies…

$$\theta^\star \cdot \phi(s,a)$$

$$\theta^\star \cdot \left[\mathbb{E}_{s,a\sim d^{\pi}} \phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}} \phi(s,a) \right] = 0$$

Maximum Entropy Inverse RL:

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

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

But there are potentially many such policies…

Traj-distribution

Find a π whose ρ^π has the largest entropy,
subject to expected feature matching

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

Maximum Entropy Inverse RL:

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

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

But there are potentially many such policies…

Find a π whose ρ^π has the largest entropy, subject to expected feature matching $\mathbb{E}_{s,a\sim d^{\pi}}\phi(s,a) = \mathbb{E}_{s,a\sim d^{\pi^\star}}\phi(s,a)$

$$\max_{\pi} \text{entropy}[\rho^\pi]$$

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

Maximum Entropy Inverse RL:

Let's simplify the objective $\max_{\pi} \text{entropy}[\rho^{\pi}]$:

Maximum Entropy Inverse RL:

Let's simplify the objective $\max_{\pi} \text{entropy}[\rho^\pi]$:

Recall the definition of trajectory distribution:

$$\rho^\pi(\tau) = \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)\ldots$$

Maximum Entropy Inverse RL:

Let's simplify the objective $\max_{\pi} \text{entropy}[\rho^\pi]$:

Recall the definition of trajectory distribution:

$$\rho^\pi(\tau) = \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)\ldots$$

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

Maximum Entropy Inverse RL:

Let's simplify the objective $\max_{\pi} \text{entropy}[\rho^{\pi}]$:

Recall the definition of trajectory distribution:

$$\rho^{\pi}(\tau) = \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)\ldots$$

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

$$\arg\max_{\pi} \text{entroy}(\rho^{\pi}) = \arg\max_{\pi} -\sum_{\tau} \rho^{\pi}(\tau)\left[\sum_{h=0}^{H-1} \ln \pi(a_h \mid s_h)\right]$$

Maximum Entropy Inverse RL:

Let's simplify the objective $\max_{\pi} \text{entropy}[\rho^{\pi}]$:

Recall the definition of trajectory distribution:

$$\rho^{\pi}(\tau) = \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)\ldots$$

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

$$\arg\max_{\pi} \text{entroy}(\rho^{\pi}) = \arg\max_{\pi} -\sum_{\tau} \rho^{\pi}(\tau)\left[\sum_{h=0}^{H-1} \ln \pi(a_h \mid s_h)\right]$$

$$= \arg\max_{\pi} \underset{\Delta}{-} \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim \cancel{\rho^{\pi}}\, P_h^{\pi}} \ln \pi(a \mid s)$$

Maximum Entropy Inverse RL:

Reformulating the optimization program:

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

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

$\mathbb{E}_{s,a \sim d^{\pi}} c(s,a)$

feature-matching

Maximum Entropy Inverse RL:

Reformulating the optimization program:

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

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

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

Using Lagrange formulation (Lagrange multiplier θ), we get:

$$\min_{\pi} \mathbb{E}_{s,a\sim d^{\pi}} \ln \pi(a \mid s) + \max_{\theta} \left(\mathbb{E}_{s,a\sim d^{\pi}} \theta^\top \phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}} \theta^\top \phi(s,a) \right)$$

Maximum Entropy Inverse RL:

Reformulating the optimization program:

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

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

$\pi(\cdot|s) \in \Delta(A)$

Using Lagrange formulation (Lagrange multiplier θ), we get:

$$\min_{\pi} \mathbb{E}_{s,a\sim d^{\pi}} \ln \pi(a \mid s) + \max_{\theta} \left(\mathbb{E}_{s,a\sim d^{\pi}}\theta^\top\phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}}\theta^\top\phi(s,a) \right)$$

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

Using minimax theorem (John von Neumann), we can swap the order of min-max:

$$\max_{\theta} \min_{\pi} \left[\mathbb{E}_{s,a\sim d^{\pi}} \ln \pi(a \mid s) + \mathbb{E}_{s,a\sim d^{\pi}}\theta^\top\phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}}\theta^\top\phi(s,a) \right]$$

Maximum Entropy Inverse RL: Final Algorithm

We get the final formulation:

$$\max_{\theta} \min_{\pi} \left[\mathbb{E}_{s,a \sim d^{\pi}} \theta^{\top} \phi(s,a) - \mathbb{E}_{s,a \sim d^{\pi^{\star}}} \theta^{\top} \phi(s,a) + \underbrace{\mathbb{E}_{s,a \sim d^{\pi}} \ln \pi(a \mid s)}_{\text{Regularization}} \right]$$

fix θ

$\min_{\pi} E_{s,a \sim d^{\pi}} \theta^{\top} \phi(s,a)$ (1)

$E_{s,a \sim d^{\pi^{*}}} \theta^{\top} \phi(s,a)$ (2)

(1) < (2)

for θ^{*} (2) ≤ (1) (π^{*} is optimal under θ^{*})

Maximum Entropy Inverse RL: Final Algorithm

incremental update on θ

Best Response on π

We get the final formulation:

$$\max_{\theta}\min_{\pi}\left[\mathbb{E}_{s,a\sim d^{\pi}}\theta^{\top}\phi(s,a)-\mathbb{E}_{s,a\sim d^{\pi^\star}}\theta^{\top}\phi(s,a)+\mathbb{E}_{s,a\sim d^{\pi}}\ln\pi(a\,|\,s)\right]$$

Algorithm: gradient ascent on θ (w/ fixed π),

and exact computation (e.g, planning, VI) for π (w/ fixed θ

Maximum Entropy Inverse RL: Final Algorithm

We get the final formulation:

$$\max_{\theta} \min_{\pi} \left[\mathbb{E}_{s,a\sim d^{\pi}}\theta^{\top}\phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^{\star}}}\theta^{\top}\phi(s,a) + \mathbb{E}_{s,a\sim d^{\pi}} \ln \pi(a \mid s)\right]$$

Algorithm: gradient ascent on θ (w/ fixed π),

and exact computation (e.g, planning, VI) for π (w/ fixed θ

$\max_{\theta} \min_{\pi} [$

Initializing θ_0:

For $t = 0, \ldots,$

min Best Response to θ_t

$$\pi_t = \arg \cancel{\max_{\pi}} \mathbb{E}_{s,a\sim d^{\pi}}\left[\theta_t^{\top}\phi(s,a) + \pi(a \mid s)\right]$$

$$\theta_{t+1} = \theta_t + \eta \left(\underbrace{\mathbb{E}_{s,a\sim d^{\pi_t}}\phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^{\star}}}\phi(s,a)}_{\text{Gradient of } \theta \text{ wrt } \pi_t}\right)$$

Return θ_T

Maximum Entropy Inverse RL: Final Algorithm

We get the final formulation:

$$\max_{\theta} \min_{\pi} \left[\mathbb{E}_{s,a\sim d^{\pi}} \theta^{\top}\phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}} \theta^{\top}\phi(s,a) + \mathbb{E}_{s,a\sim d^{\pi}} \ln \pi(a \mid s)\right]$$

Algorithm: gradient ascent on θ (w/ fixed π),
and exact computation (e.g, planning, VI) for π (w/ fixed θ

(hybrid setting: P is known)

Initializing θ_0:

For $t = 0, \ldots,$

(Maximum Entropy RL)

$$\pi_t = \arg\max_{\pi} \mathbb{E}_{s,a\sim d^{\pi}} \left[\theta_t^{\top}\phi(s,a) + \pi(a \mid s)\right]$$

$$\theta_{t+1} = \theta_t + \eta \left(\mathbb{E}_{s,a\sim d^{\pi_t}}\phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}}\phi(s,a)\right)$$

Return θ_T

Maximum Entropy Inverse RL: Final Algorithm

We get the final formulation:

$$\max_{\theta} \min_{\pi} \left[\mathbb{E}_{s,a\sim d^{\pi}} \theta^{\top} \phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}} \theta^{\top} \phi(s,a) + \mathbb{E}_{s,a\sim d^{\pi}} \ln \pi(a\,|\,s) \right]$$

Algorithm: gradient ascent on θ (w/ fixed π),
and exact computation (e.g, planning, VI) for π (w/ fixed θ

Initializing θ_0:

For $t = 0, \ldots,$

we have $\underline{\underline{P}}$ (planing)

(Maximum Entropy RL)

$$\pi_t = \arg\max_{\pi} \mathbb{E}_{s,a\sim d^{\pi}} \left[\theta_t^{\top} \phi(s,a) + \pi(a\,|\,s) \right]$$

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

$$\theta_{t+1} = \theta_t + \eta \left(\mathbb{E}_{s,a\sim d^{\pi_t}} \phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}} \phi(s,a) \right)$$

(Gradient equal to the difference of expected features)

Return θ_T

Maximum Entropy RL: Soft ~~Policy~~ Value Iteration

Maximum Entropy RL: what we do when our "cost" depends on policy π?

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

Maximum Entropy RL: Soft Policy Iteration

Maximum Entropy RL: what we do when our "cost" depends on policy π?

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

Value

Soft ~~Policy~~ Iteration:

Maximum Entropy RL: Soft Policy Iteration

Maximum Entropy RL: what we do when our "cost" depends on policy π?

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

$\pi^{\star}_{H-1}(s) = \arg\min_{a \in A} Q^{\star}(s,a)$

Soft Policy Iteration:

$$Q^{\star}_{H-1}(s,a) = c(s,a) \quad \pi^{\star}_{H-1}(a\,|\,s) \propto \exp\left(-Q^{\star}_{H-1}(s,a)\right) \propto \exp(-A^{\star}_{H-1}(s,a))$$

$V^{\pi}_{H-1}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[c(s,a) + \ln \pi(a|s)\right]$

$\min_{\pi} V^{\pi}_{H-1}(s)$

Maximum Entropy RL: Soft Policy Iteration

Maximum Entropy RL: what we do when our "cost" depends on policy π?

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

Soft Policy Iteration:

$$Q^{\star}_{H-1}(s,a) = c(s,a) \quad \pi^{\star}_{H-1}(a \mid s) \propto \exp\left(-Q^{\star}_{H-1}(s,a)\right) \propto \exp(-A^{\star}_{H-1}(s,a))$$

$$V^{\star}_{H-1}(s) = \mathbb{E}_{a\sim\pi^{\star}_{H-1}(a|s)} \left[\ln \pi^{\star}_{H-1}(a \mid s) + Q^{\star}_{H-1}(s,a) \right] = - \ln \left(\sum_{a} \exp\left(-Q^{\star}_{H-1}(s,a)\right) \right)$$

$$V^{*}_{H-1}(s) = \min_{a\in\mathcal{A}} Q^{*}_{H-1}(s,a)$$

Maximum Entropy RL: Soft Policy Iteration

Maximum Entropy RL: what we do when our “cost” depends on policy π?

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

Soft Policy Iteration:

$$Q^{\star}_{H-1}(s,a) = c(s,a) \quad \pi^{\star}_{H-1}(a \mid s) \propto \exp\left(-Q^{\star}_{H-1}(s,a)\right) \propto \exp(-A^{\star}_{H-1}(s,a))$$

$$V^{\star}_{H-1}(s) = \mathbb{E}_{a\sim\pi^{\star}_{H-1}(a|s)}\left[\ln \pi^{\star}_{H-1}(a \mid s) + Q^{\star}_{H-1}(s,a)\right] = -\ln\left(\sum_{a} \exp\left(-Q^{\star}_{H-1}(s,a)\right)\right)$$

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

Maximum Entropy RL: Soft Policy Iteration

Maximum Entropy RL: what we do when our "cost" depends on policy π?

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

Soft Policy Iteration:

$$Q^{\star}_{H-1}(s,a) = c(s,a) \quad \pi^{\star}_{H-1}(a \mid s) \propto \exp\left(-Q^{\star}_{H-1}(s,a)\right) \propto \exp(-A^{\star}_{H-1}(s,a))$$

$$V^{\star}_{H-1}(s) = \mathbb{E}_{a\sim\pi^{\star}_{H-1}(a|s)}\left[\ln \pi^{\star}_{H-1}(a \mid s) + Q^{\star}_{H-1}(s,a)\right] = -\ln\left(\sum_{a} \exp\left(-Q^{\star}_{H-1}(s,a)\right)\right)$$

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

$$\pi^{\star}_{h}(a \mid s) \propto \exp(-Q^{\star}_{h}(s,a)) \propto \exp(-A^{\star}_{h}(s,a))$$

Maximum Entropy RL: Soft Policy Iteration

Maximum Entropy RL: what we do when our “cost” depends on policy π?

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

Soft Policy Iteration:

$$Q^{\star}_{H-1}(s,a) = c(s,a) \quad \pi^{\star}_{H-1}(a\,|\,s) \propto \exp\left(-Q^{\star}_{H-1}(s,a)\right) \propto \exp(-A^{\star}_{H-1}(s,a))$$

$$V^{\star}_{H-1}(s) = \mathbb{E}_{a\sim\pi^{\star}_{H-1}(a|s)}\left[\ln \pi^{\star}_{H-1}(a\,|\,s) + Q^{\star}_{H-1}(s,a)\right] = -\ln\left(\sum_a \exp\left(-Q^{\star}_{H-1}(s,a)\right)\right)$$

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

$$\pi^{\star}_{h}(a\,|\,s) \propto \exp(-Q^{\star}_{h}(s,a)) \propto \exp(-A^{\star}_{h}(s,a))$$

$$V^{\star}_{h}(s) = -\ln\left(\sum_a \exp(-Q^{\star}_{h}(s,a))\right)$$

Maximum Entropy RL: Soft Policy Iteration

Maximum Entropy RL: what we do when our "cost" depends on policy π?

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

Soft Policy Iteration:

$$Q^{\star}_{H-1}(s,a) = c(s,a) \quad \pi^{\star}_{H-1}(a\,|\,s) \propto \exp\left(-Q^{\star}_{H-1}(s,a)\right) \propto \exp(-A^{\star}_{H-1}(s,a))$$

$$V^{\star}_{H-1}(s) = \mathbb{E}_{a\sim\pi^{\star}_{H-1}(a|s)}\left[\ln \pi^{\star}_{H-1}(a\,|\,s) + Q^{\star}_{H-1}(s,a)\right] = -\ln\left(\sum_a \exp\left(-Q^{\star}_{H-1}(s,a)\right)\right)$$

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

$$\pi^{\star}_{h}(a\,|\,s) \propto \exp(-Q^{\star}_{h}(s,a)) \propto \exp(-A^{\star}_{h}(s,a))$$

$$V^{\star}_{h}(s) = -\ln\left(\sum_a \exp(-Q^{\star}_{h}(s,a))\right)$$

Derivation: DP!

Maximum Entropy RL: Calculate Trajectory Likelihood

Initializing θ_0:

For $t = 0, \ldots,$

$$\pi_t = \arg\max_{\pi} \mathbb{E}_{s,a\sim d^{\pi}}\left[\theta_t^{\top}\phi(s,a) + \pi(a\,|\,s)\right]$$

$$\theta_{t+1} = \theta_t + \eta\left(\mathbb{E}_{s,a\sim d^{\pi_t}}\phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}}\phi(s,a)\right)$$

Return θ_T

$\widehat{\pi}$:= soft ~~P~~VI $\left(\theta_T^{\top}\phi(s,a)\right)$

$$\widehat{\pi}_h(a\,|\,s) \propto \exp\left(-Q_h^\star\left(s,a;\theta_T\right)\right)$$

$c(s,a) = \theta_T^{\top} \cdot \phi(s,a)$

Maximum Entropy RL: Calculate Trajectory Likelihood

Initializing θ_0:

For $t = 0, \ldots,$

$$\pi_t = \arg\max_{\pi} \mathbb{E}_{s,a\sim d^{\pi}} \left[\theta_t^{\top} \phi(s,a) + \pi(a|s)\right]$$

$$\theta_{t+1} = \theta_t + \eta \left(\mathbb{E}_{s,a\sim d^{\pi_t}} \phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}} \phi(s,a)\right)$$

Return θ_T

$\widehat{\pi}$:= soft PI $\left(\theta_T^{\top} \phi(s,a)\right)$

$$\widehat{\pi}_h(a|s) \propto \exp\left(-Q_h^\star\left(s,a;\theta_T\right)\right)$$

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?

Maximum Entropy RL: Calculate Trajectory Likelihood

Initializing θ_0:

For $t = 0, \ldots,$

$$\pi_t = \arg\max_{\pi} \mathbb{E}_{s,a\sim d^{\pi}} \left[\theta_t^{\top}\phi(s,a) + \pi(a \mid s)\right]$$

$$\theta_{t+1} = \theta_t + \eta \left(\mathbb{E}_{s,a\sim d^{\pi_t}}\phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}}\phi(s,a)\right)$$

Return θ_T

$\widehat{\pi}$:= soft PI $\left(\theta_T^{\top}\phi(s,a)\right)$

$$\widehat{\pi}_h(a \mid s) \propto \exp\left(-Q_h^\star\left(s,a;\theta_T\right)\right)$$

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?

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

Maximum Entropy RL: Calculate Trajectory Likelihood

Initializing θ_0:

For $t = 0, \ldots,$

$$\pi_t = \arg\max_{\pi} \mathbb{E}_{s,a \sim d^{\pi}} \left[\theta_t^{\top} \phi(s,a) + \pi(a \mid s)\right]$$

$$\theta_{t+1} = \theta_t + \eta \left(\mathbb{E}_{s,a \sim d^{\pi_t}} \phi(s,a) - \mathbb{E}_{s,a \sim d^{\pi^\star}} \phi(s,a)\right)$$

Return θ_T

$\hat{\pi} :=$ soft ~~PI~~ VI $\left(\theta_T^{\top} \phi(s,a)\right)$

$$\hat{\pi}_h(a \mid s) \propto \exp\left(-Q_h^\star\left(s, a; \theta_T\right)\right)$$

$$\min_{\pi} \mathbb{E}_{s,a \sim d^{\pi}} \left[c(s,a)\right] \Leftarrow \text{NPG}$$

$$\min_{\pi} \mathbb{E}_{s,a \sim d^{\pi}} \left[c(s,a)\right] + \beta\left(-\text{entropy}(\rho^{\pi})\right)$$

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?

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

Special case: deterministic MDP and state-dependent cost:

Maximum Entropy RL: Calculate Trajectory Likelihood

Initializing θ_0:

For $t = 0, \ldots,$

$$\pi_t = \arg\max_\pi \mathbb{E}_{s,a\sim d^\pi}\left[\theta_t^\top \phi(s,a) + \pi(a|s)\right]$$

$$\theta_{t+1} = \theta_t + \eta\left(\mathbb{E}_{s,a\sim d^{\pi_t}}\phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi^\star}}\phi(s,a)\right)$$

Return θ_T

$\hat{\pi}$:= soft PI $\left(\theta_T^\top \phi(s,a)\right)$

$$\hat{\pi}_h(a|s) \propto \exp\left(-Q_h^\star(s,a;\theta_T)\right)$$

$\theta_T \cdot \phi(s)$

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?

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

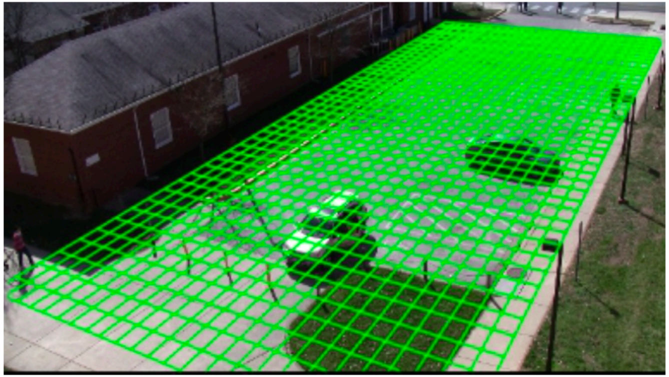
Special case: deterministic MDP and state-dependent cost:

$\theta^\star \phi(s)$

For a state trajectory, we have:

$$\rho^\pi(s_0, s_1, \ldots, s_H) \propto \exp\left(-\sum_h \theta_T^\top \phi(s_h)\right)$$

Running Example: Human Trajectory Forecasting



State space: grid,
action space: 4 actions

We predict that we are more likely to use sidewalk

$$\hat{c}(s) = \theta_T^T \phi(s)$$

$$p^\pi(\tau) \propto \exp\left(-\sum_{t=0}^{H-1} \hat{c}(s_t)\right)$$

MaxEnt - IRL

$$\max_{\pi} \text{Entropy}(\rho^{\pi})$$

$$\text{s.t} \quad \mathbb{E}_{s \text{ and } \pi} \phi(s,a) = \mathbb{E}_{s \text{ and } \pi^*} \phi(s,a)$$

Soft VI:

(SAC)

$$V_h^*(s) = -\ln\left[\sum_a \exp(-Q_h^*(s,a))\right]$$

$$\pi^*(a|s) \propto \exp(-Q_h^*(s,a))$$

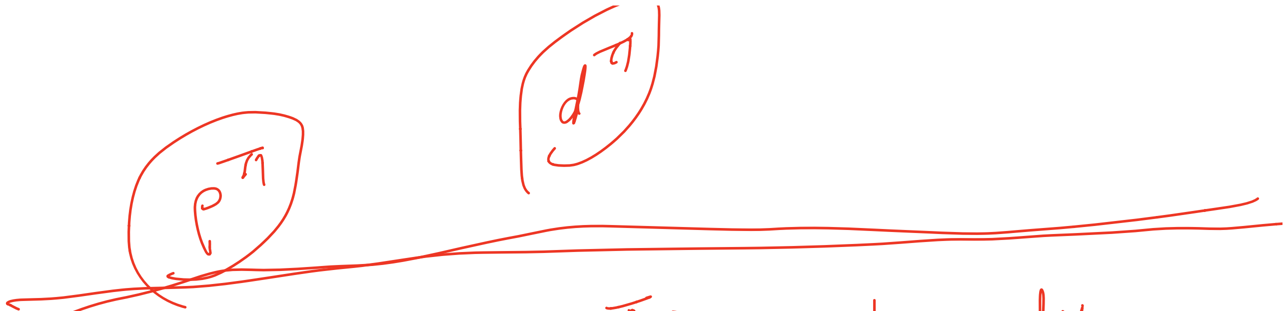
$$\left\| \mathbb{E}_{s,a \sim d^{\pi}} \phi(s,a) - \mathbb{E}_{s,a \sim d^{\pi^*}} \phi(s,a) \right\|_2^2$$

$$\|x\|_2 \iff \max_{\theta : \|\theta\|_2 \le 1} \theta^T x$$

$$\max_{\theta \in \Theta} \left[\mathbb{E}_{s,a \sim d^{\pi}} \theta^T \phi(s,a) - \mathbb{E}_{s,a \sim d^{\pi^*}} \theta^T \phi(s,a) \right]$$

unit-Ball

$$\mathcal{F} = \left\{ \theta^T \phi(s,a) : \theta \in \text{Unit-Ball} \right\}$$



Entropy(ρ^π) ← strongly convex function

Deterministic & $\phi(s)$

$$\rho_\theta(\tau) \propto \exp\left(-\sum_{h=0}^{H-1} \theta^T \cdot \phi(s)\right)$$

$\tau_1^*, \tau_2^*, \ldots \tau_N^* \leftarrow \rho^{\pi^*}$

$$\cancel{\max}\ \max_{\theta} \sum_{i=1}^{N} \ln\left(\rho_\theta(\tau_i^*)\right)$$

original - MaxEnt IRL