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 $\mathcal{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d^{\pi^\star}$

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)

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

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

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

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)

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

Today: Hybrid Setting

Algorithm: Maximum Entropy Inverse Reinforcement Learning

Running Example: Human trajectory forecasting

[Kitani, et al, ECCV 12]

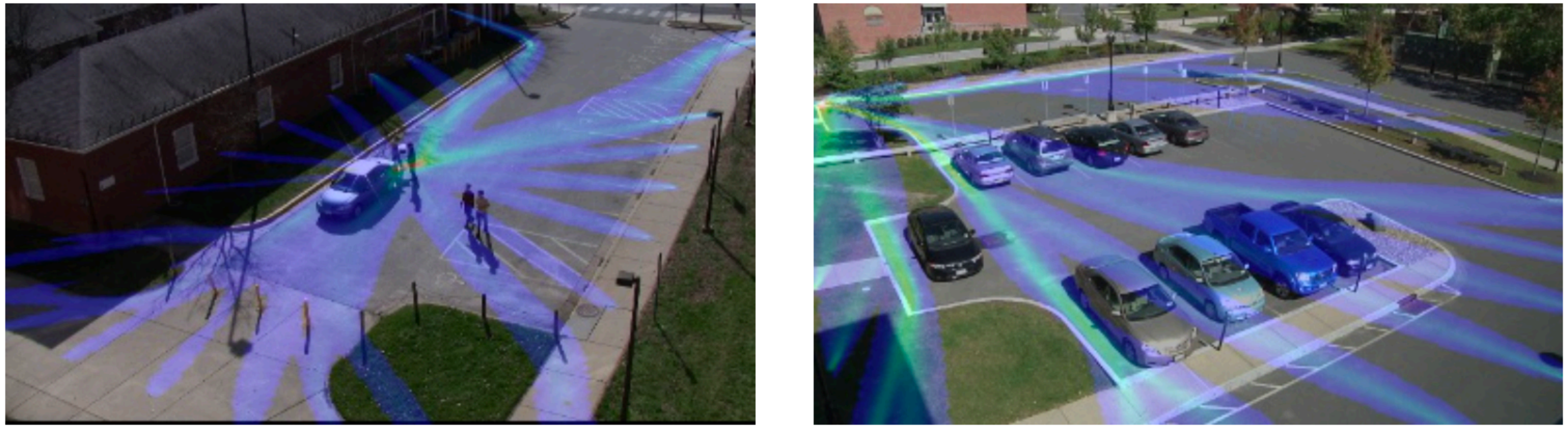


Fig. 1. Given a single pedestrian detection, our proposed approach forecasts plausible paths and destinations from noisy vision-input

Running Example: Human trajectory forecasting

[Kitani, et al, ECCV 12]

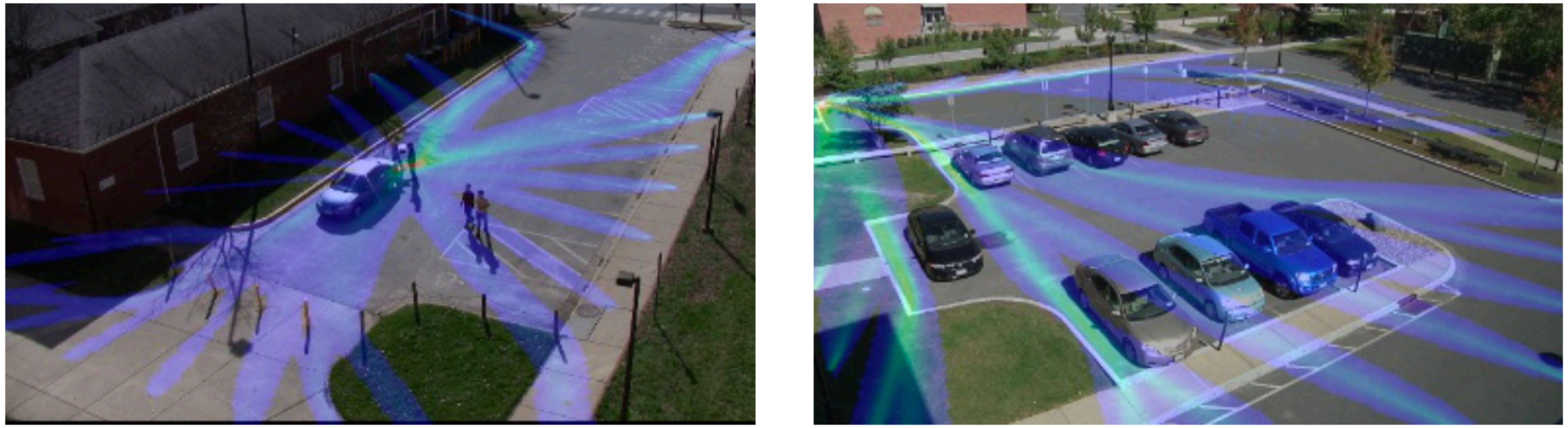


Fig. 1. Given a single pedestrian detection, our proposed approach forecasts plausible paths and destinations from noisy vision-input

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

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

Setting

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

(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 $\mathcal{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d^{\pi^\star}$

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

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

$\rho^\pi(\tau) := \mu_0(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)\ldots\pi(a_{H-1} \mid s_{H-1})P(s_H \mid s_{H-1}, a_{H-1})$:
Likelihood of the trajectory τ under π

Running Example: Define feature map

Key Assumption on cost:

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

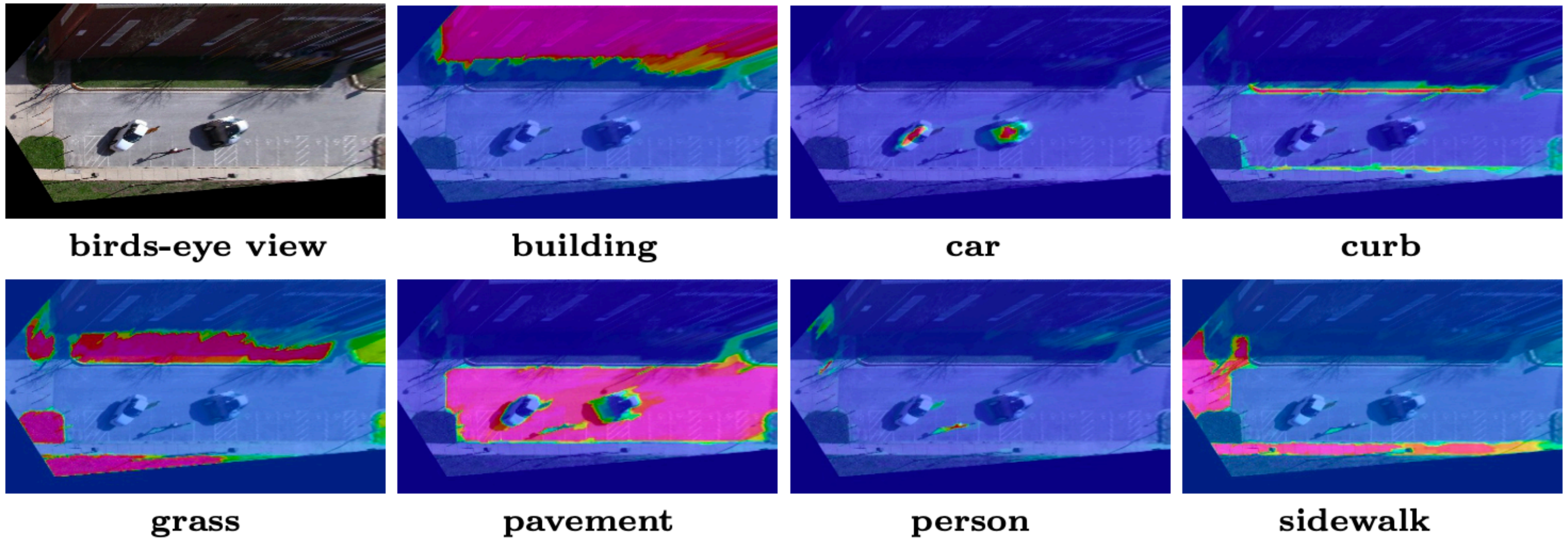


Fig. 4. Classifier feature response maps. Top left is the original image.

Running Example: Define feature map

Key Assumption on cost:

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

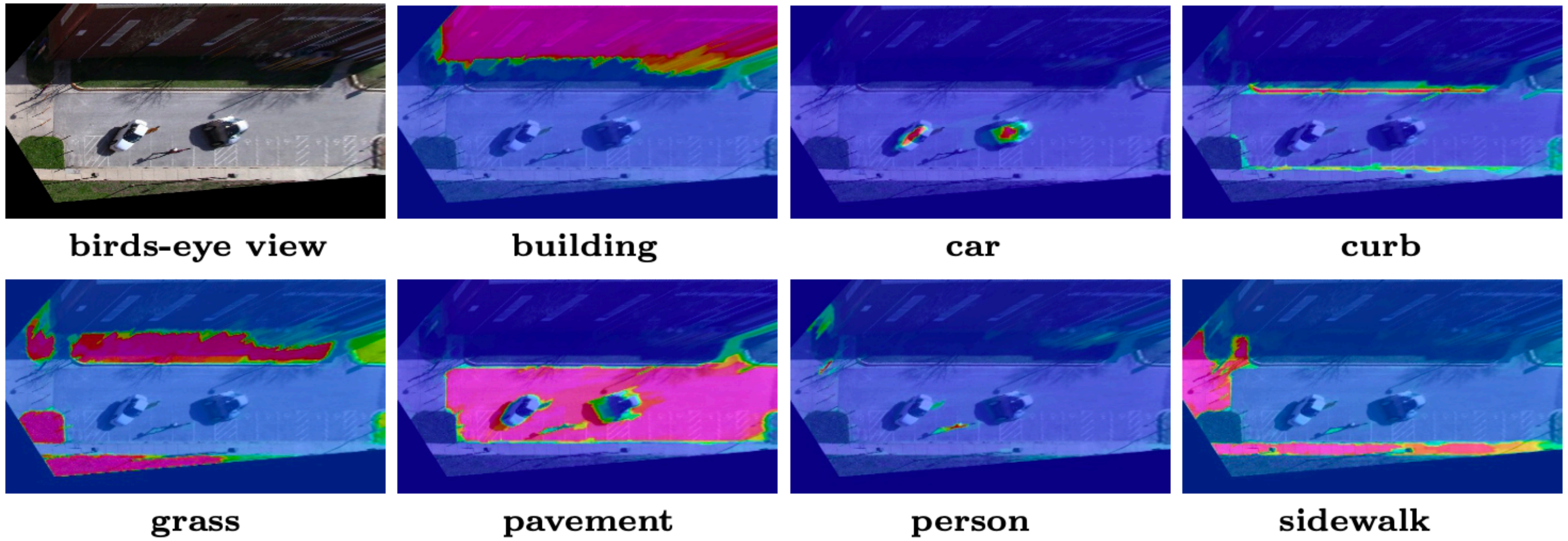


Fig. 4. Classifier feature response maps. Top left is the original image.

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

Running Example: Define feature map

Key Assumption on cost:

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

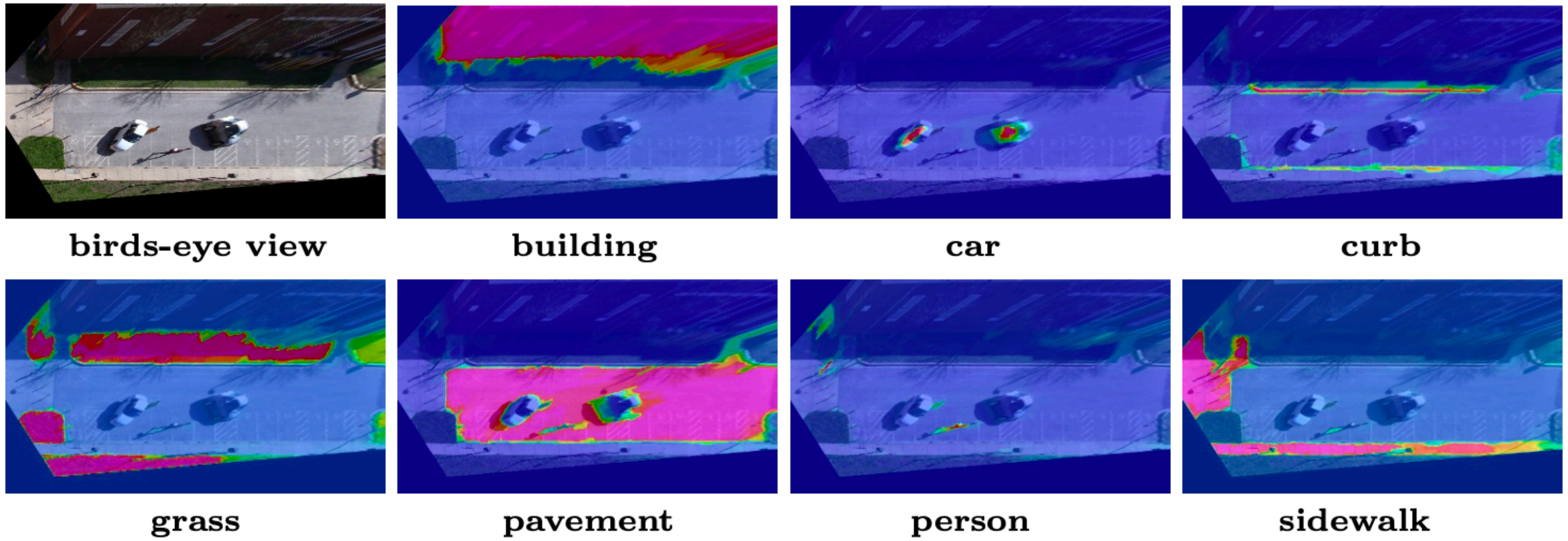


Fig. 4. Classifier feature response maps. Top left is the original image.

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

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

Running Example: Define feature map

Key Assumption on cost:
$c(s,a) = \langle \theta^\star, \phi(s,a) \rangle$, **linear wrt feature** $\phi(s,a)$

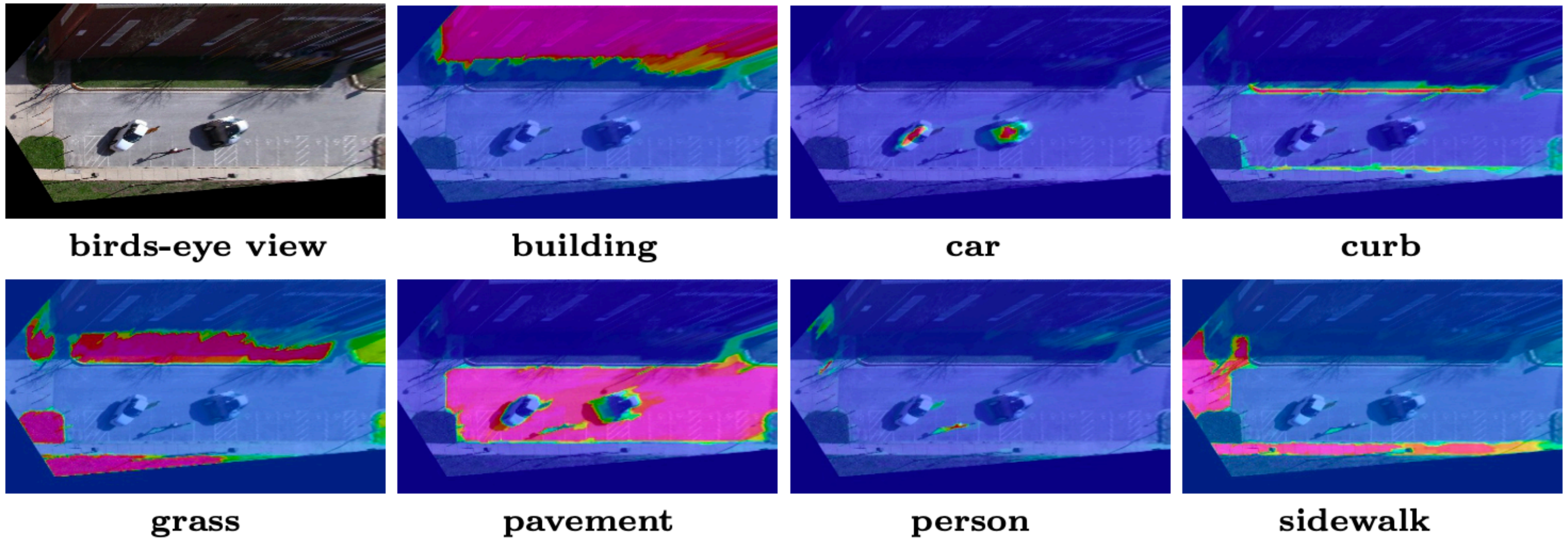


Fig. 4. Classifier feature response maps. Top left is the original image.

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

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

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

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

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

Maximum Entropy Inverse RL:

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…

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

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

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:

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

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

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:

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

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

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

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

Maximum Entropy Inverse RL:

Reformulating the optimization program:

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

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

Maximum Entropy Inverse RL:

Reformulating the optimization program:

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

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

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

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

Maximum Entropy Inverse RL:

Reformulating the optimization program:

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

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

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

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

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:

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

Maximum Entropy Inverse RL: Final Algorithm

We get the final formulation:

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

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

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 $t=0,\ldots,$

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

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

Return θ_T

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 $t = 0, \ldots,$

(Maximum Entropy RL)

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

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

Return θ_T

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)

Initializing θ_0:

For $t = 0, \ldots,$

(Maximum Entropy RL)

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

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

(Gradient equal to the difference of expected features)

Return θ_T

Maximum Entropy RL: Soft Value Iteration

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

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

Maximum Entropy RL: Soft Value Iteration

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

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

Soft Value Iteration:

Maximum Entropy RL: Soft Value Iteration

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

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

Soft Value Iteration:

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

Maximum Entropy RL: Soft Value Iteration

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

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

Soft Value Iteration:

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

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

Maximum Entropy RL: Soft Value Iteration

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

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

Soft Value Iteration:

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

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

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

Maximum Entropy RL: Soft Value Iteration

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

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

Soft Value 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 Value Iteration

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

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

Soft Value Iteration:

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

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

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

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

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

Maximum Entropy RL: Soft Value Iteration

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

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

Soft Value 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 $t = 0, \ldots,$

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

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

Return θ_T

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

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

Maximum Entropy RL: Calculate Trajectory Likelihood

Initializing θ_0:

For $t = 0, \ldots,$

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

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

Return θ_T

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

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

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?

Maximum Entropy RL: Calculate Trajectory Likelihood

Initializing θ_0:

For $t = 0, \ldots,$

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

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

Return θ_T

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

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

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?

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

Maximum Entropy RL: Calculate Trajectory Likelihood

Initializing θ_0:

For $t = 0, \ldots,$

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

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

Return θ_T

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

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

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?

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

Special case: deterministic MDP and state-dependent cost:

Maximum Entropy RL: Calculate Trajectory Likelihood

Initializing θ_0:

For $t = 0, \ldots,$

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

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

Return θ_T

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

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

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?

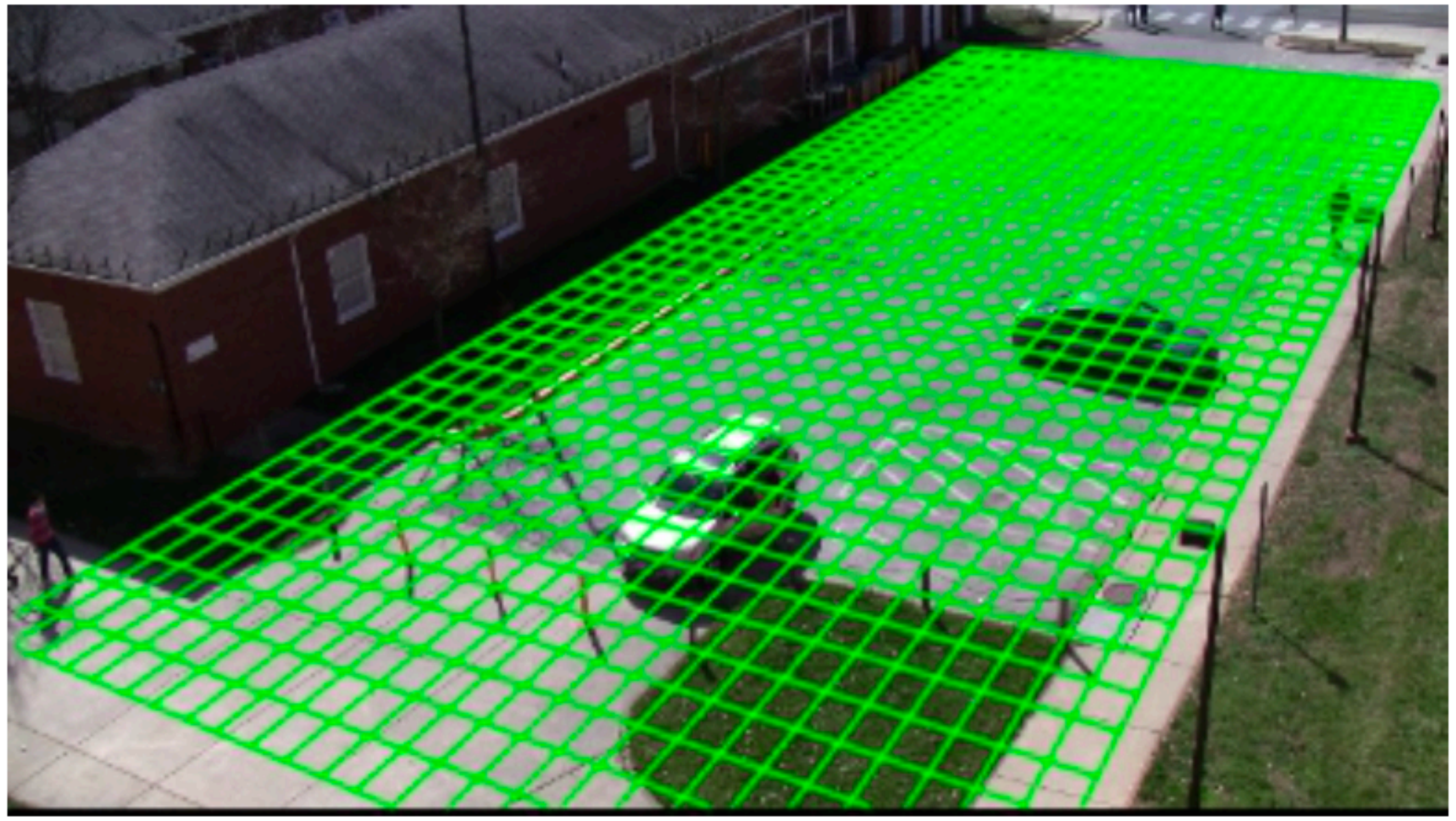
$$\ln\left(\rho^{\widehat{\pi}}(\tau)\right) = \sum_{h=0}^{H-1}\left[\ln P(s_{h+1} \mid s_h, a_h) + \ln \widehat{\pi}(a_h \mid s_h)\right]$$

Special case: deterministic MDP and state-dependent cost:

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