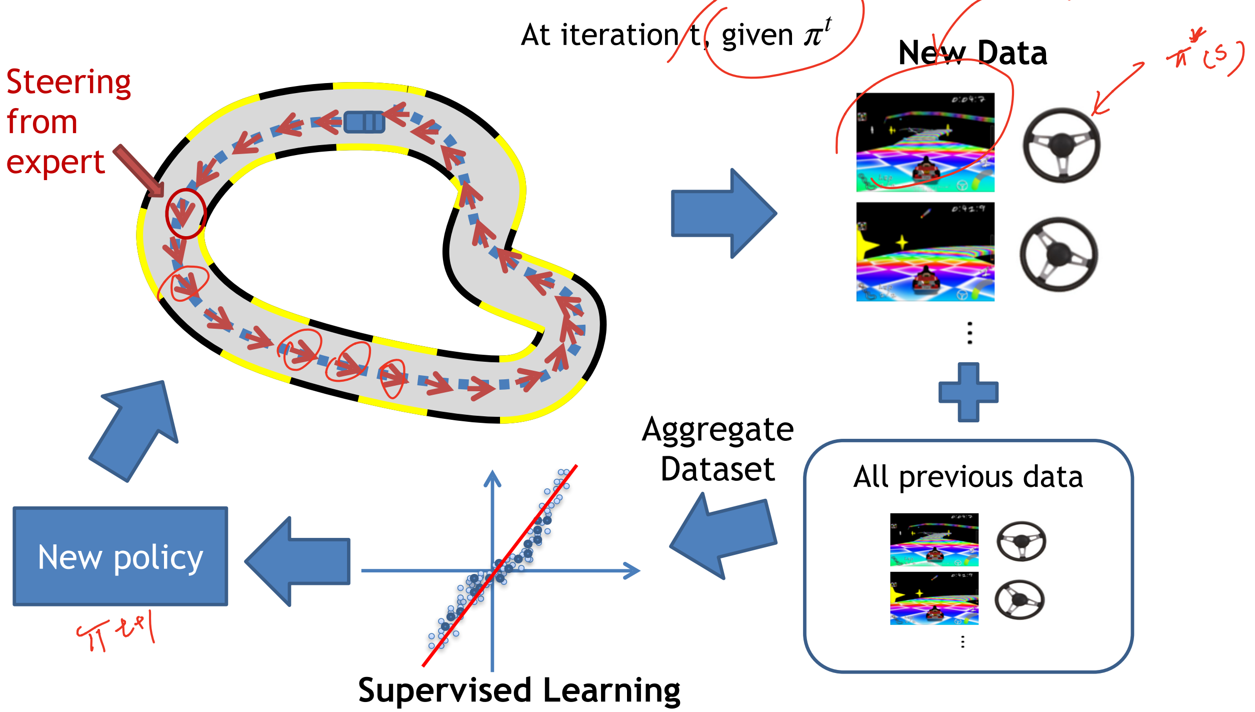
Interactive Imitation Learning
(continue)

Recap

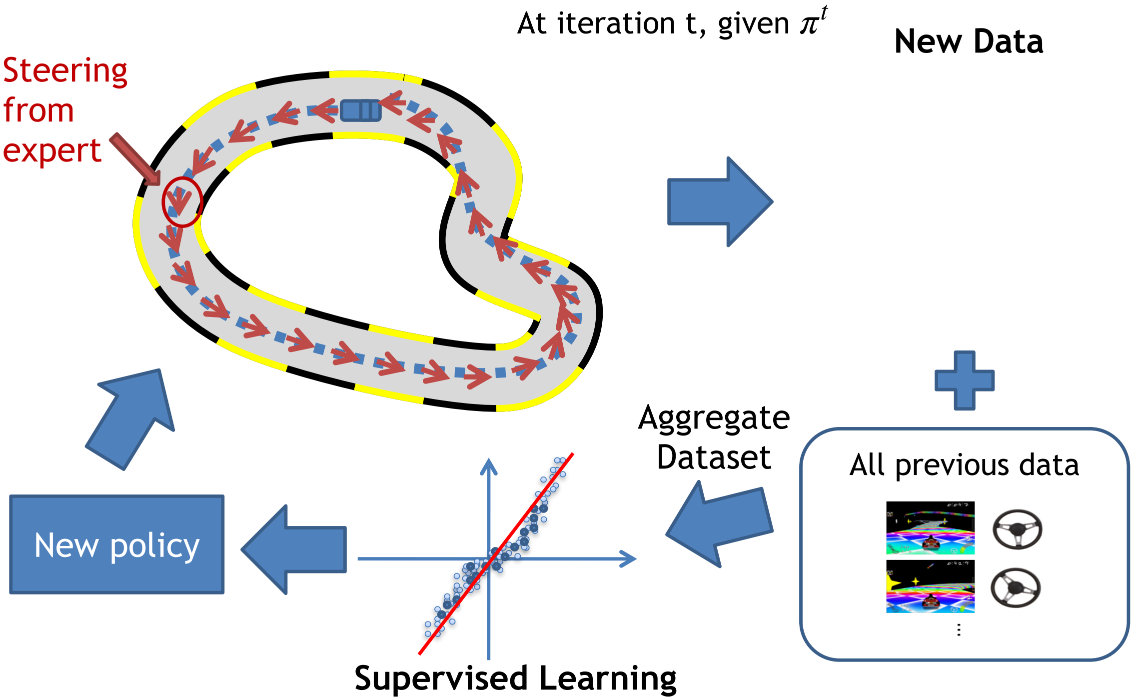
Interactive Imitation Learning Setting

Key assumption:
we can query expert $\pi^\star$ at any time and any state during training

DAgger Revisit



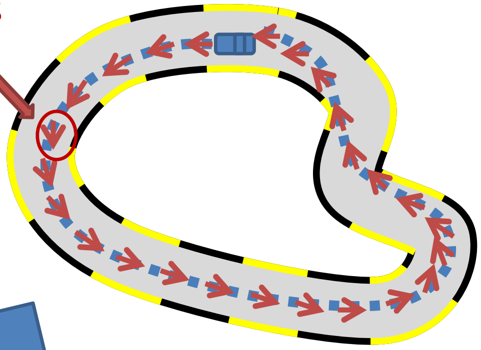
DAgger Revisit



DAgger Revisit

At iteration t, given π^t

Steering from expert



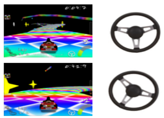
New Data

$\mathbb{1}\{\pi(s) \neq \pi^\star(s)\}$

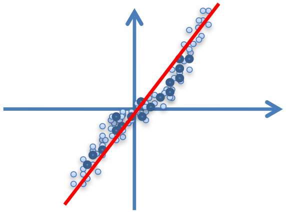
$$\ell_t(\pi) = \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[c\left(\pi, s, \pi^\star(s)\right)\right]$$

Aggregate Dataset

All previous data



⋮



Supervised Learning

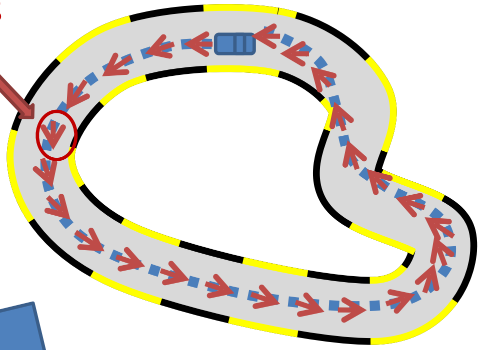
New policy

DAgger Revisit

At iteration t, given π^t

New Data

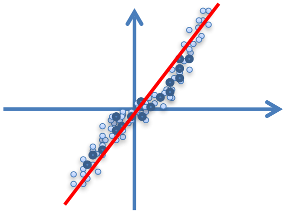
Steering from expert



$$\ell_t(\pi) = \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[c\left(\pi, s, \pi^\star(s)\right)\right]$$

All previous data

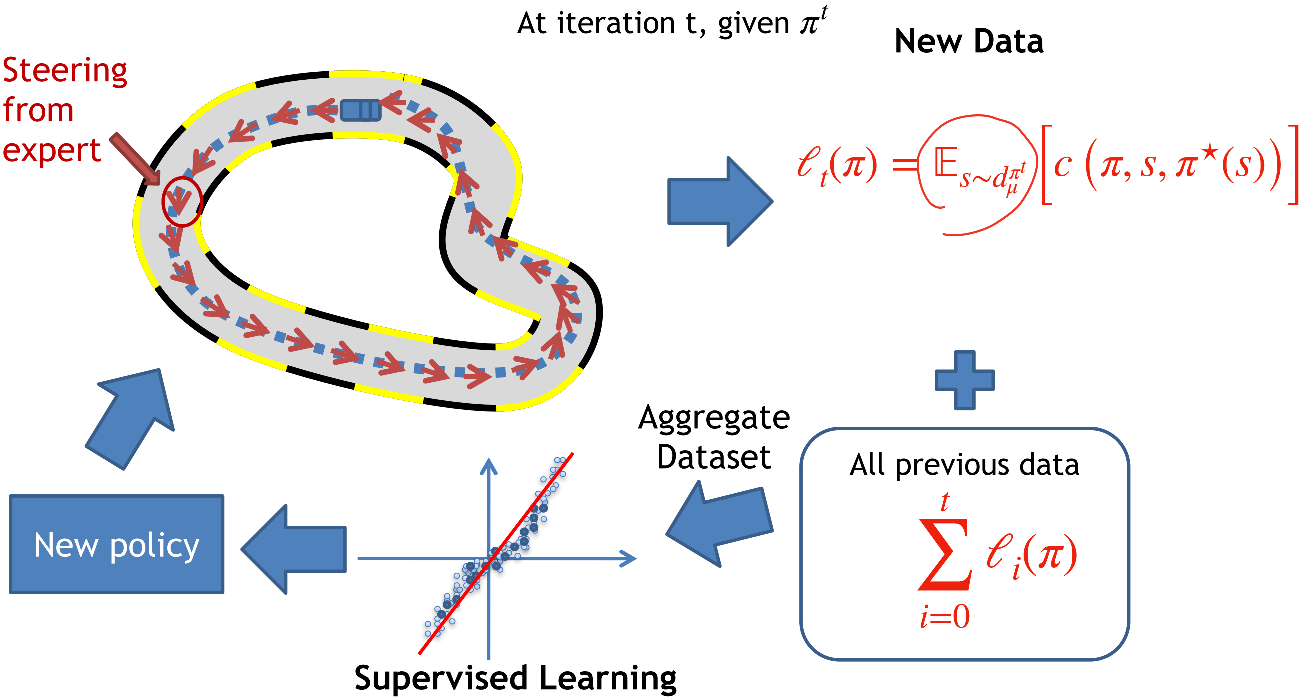
Aggregate Dataset



Supervised Learning

New policy

DAgger Revisit



DAgger Revisit

At iteration t, given π^t

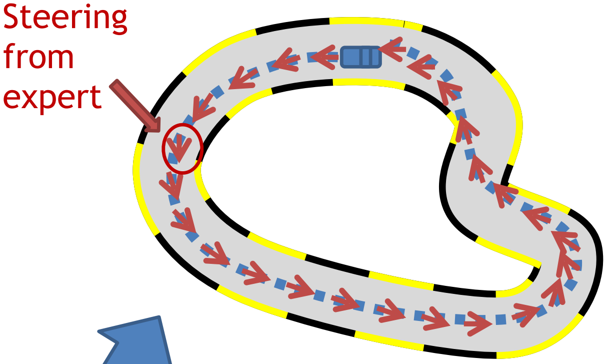
New Data

$$\ell_t(\pi) = \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[c\left(\pi, s, \pi^\star(s)\right)\right]$$

All previous data

$$\sum_{i=0}^{t} \ell_i(\pi)$$

Aggregate Dataset



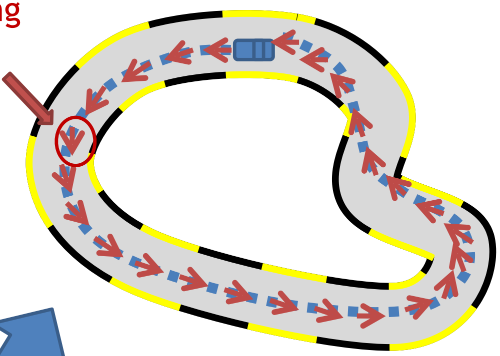
New policy

Supervised Learning

DAgger Revisit

At iteration t, given π^t

Steering from expert



New Data

$$\ell_t(\pi) = \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[c\left(\pi, s, \pi^\star(s)\right)\right]$$

All previous data

$$\sum_{i=0}^{t} \ell_i(\pi)$$

Aggregate Dataset

New policy

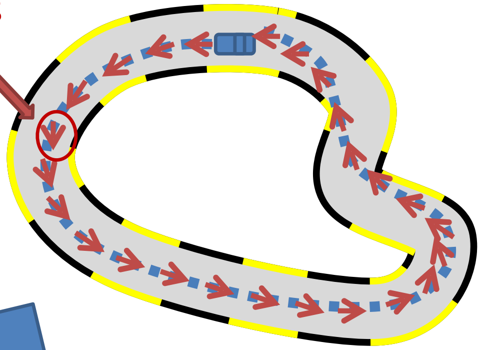
$$\pi^{t+1} = \arg\min_{\pi} \sum_{i=0}^{t} \ell_i(\pi) + \lambda R(\pi)$$

Supervised Learning

DAgger Revisit

At iteration t, given π^t

Steering from expert



New Data

$$\ell_t(\pi) = \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[c\left(\pi, s, \pi^\star(s)\right)\right]$$

+

All previous data

$$\sum_{i=0}^{t} \ell_i(\pi)$$

Aggregate Dataset

New policy

$$\pi^{t+1} = \arg\min_{\pi} \sum_{i=0}^{t} \ell_i(\pi) + \lambda R(\pi)$$

Supervised Learning

Data Aggregation = Follow-the-Regularized-Leader Online Learner

Recap on the Follow-the-Regularized Leader Guarantee:

At the end of iteration t, learner has seen $\underbrace{\ell_0, \ldots \ell_{t-1}, \ell_t}$, learner updates to a new decision:

$$\textbf{FTL: } \theta_{t+1} = \min_{\theta \in \Theta} \sum_{i=0}^{t} \ell_i(\theta) + \lambda R(\theta)$$

Recap on the Follow-the-Regularized Leader Guarantee:

At the end of iteration t, learner has seen $\ell_0, \ldots \ell_{t-1}, \ell_t$, learner updates to a new decision:

$$\textbf{FTL: } \theta_{t+1} = \min_{\theta \in \Theta} \sum_{i=0}^{t} \ell_i(\theta) + \lambda R(\theta)$$

Data Aggregation

Theorem (FTL) (optional): if Θ is convex, and ℓ_t is convex for all t, and $R(\theta)$ is strongly convex, then for regret of FTL, we have:

$$\frac{1}{T}\left[\sum_{t=0}^{T-1} \ell_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=0}^{T-1} \ell_t(\theta)\right] = O\left(1/\sqrt{T}\right)$$

$T \to \infty$

Best in Hindsight

Today's Plan

1. Finish DAgger's Analysis

2. Intro to Maximum Entropy Inverse RL
(We have offline demonstrations, but learner can interact with the environments)

DAgger Analysis: A reduction to no-regret online learning

infinite horizon MDP

(assume discrete action space—**in fact let's assume 2 actions**, so we do binary classification)

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

DAgger Analysis: A reduction to no-regret online learning

infinite horizon MDP

(assume discrete action space—**in fact let's assume 2 actions**, so we do binary classification)

$$\mathscr{M} = \{S, A, \gamma, r, P, \mu\}$$

$A = \{-1, +1\}$

Function approximation:

Decision set $\Pi := \{\pi : S \mapsto A\}$ (assume $\pi^\star \in \Pi$)

Realizability

$\Pi \rightarrow$ Binary classifier

DAgger Analysis: A reduction to no-regret online learning

infinite horizon MDP
(assume discrete action space—**in fact let's assume 2 actions**, so we do binary classification)

$$\mathcal{M} = \left\{S, A, \gamma, r, P, \mu\right\}$$

Function approximation:

Decision set $\Pi := \{\pi : S \mapsto A\}$ (assume $\pi^\star \in \Pi$)

Classification algorithm (oracle) $\mathcal{A}$: ← SVM

label

Given a binary-class data distribution ρ, where $\{x, y\} \sim \rho, y \in \{-1,1\}$

feature

$$\widehat{\pi} = \mathcal{A}\left(\Pi, \rho\right) := \arg\min_{\pi\in\Pi} \underline{\mathbb{E}_{x,y\sim\rho}\left[c\left(\pi, x, y\right)\right]}$$

DAgger Analysis: A reduction to no-regret online learning

infinite horizon MDP
(assume discrete action space—**in fact let's assume 2 actions**, so we do binary classification)

$$\mathscr{M} = \left\{S, A, \gamma, r, P, \mu\right\}$$

Function approximation:

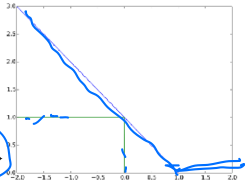
Decision set $\Pi := \{\pi : S \mapsto A\}$ (assume $\pi^\star \in \Pi$)

Classification algorithm (oracle) $\mathscr{A}$:

Given a binary-class data distribution ρ, where $\{x, y\} \sim \rho, y \in \{-1,1\}$

$$\hat{\pi} = \mathscr{A}\left(\Pi, \rho\right) := \arg\min_{\pi \in \Pi} \mathbb{E}_{x,y\sim\rho}\left[c\left(\pi, x, y\right)\right]$$

$$c(\pi, x, y) = \max\{0, 1 - \pi(x) \cdot y\}$$



DAgger Analysis: A reduction to no-regret online learning

Decision set Π (assume $\pi^\star \in \Pi$)

Δ

MDP and Expert

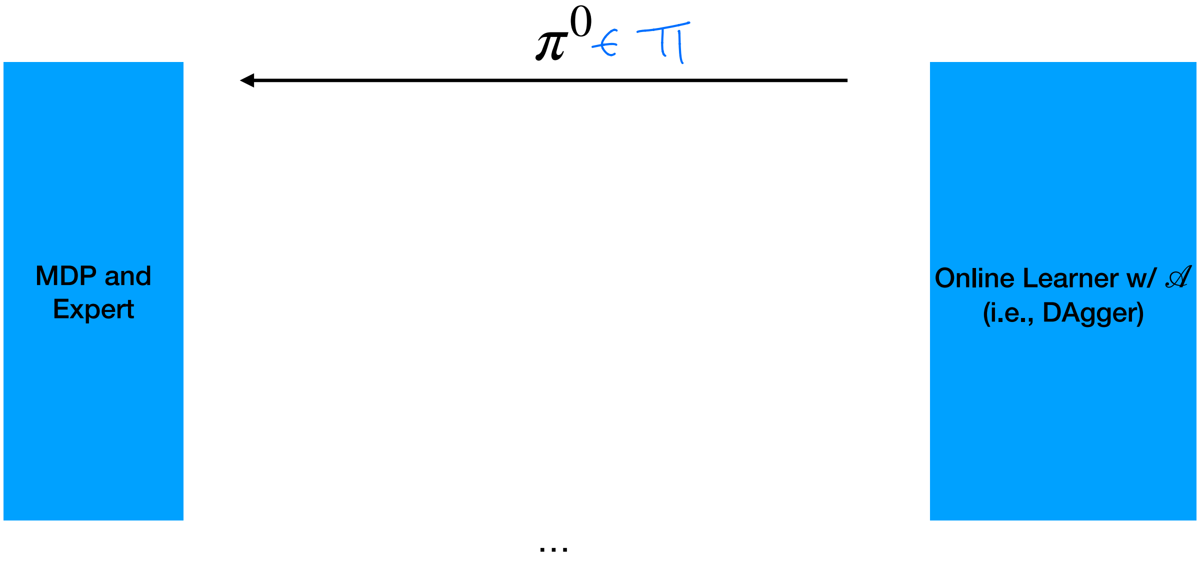
Online Learner w/ $\mathscr{A}$ (i.e., DAgger)

…

Total loss so far:

DAgger Analysis: A reduction to no-regret online learning

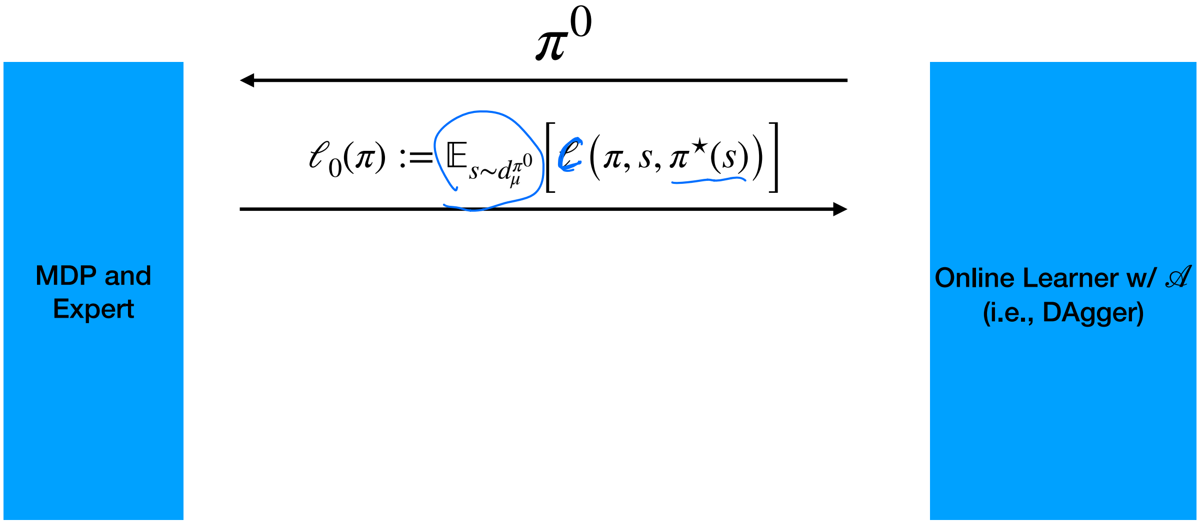
Decision set Π (assume $\pi^\star \in \Pi$)



Total loss so far:

DAgger Analysis: A reduction to no-regret online learning

Decision set Π (assume $\pi^\star \in \Pi$)

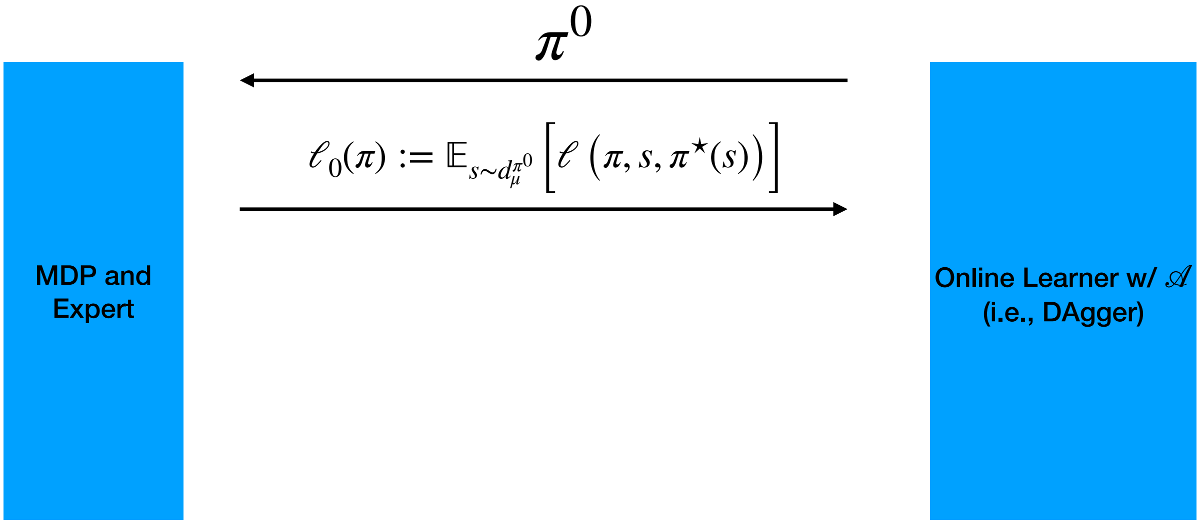


…

Total loss so far:

DAgger Analysis: A reduction to no-regret online learning

Decision set Π (assume $\pi^\star \in \Pi$)



…

Total loss so far: $\ell_0(\pi^0)$

DAgger Analysis: A reduction to no-regret online learning

Decision set Π (assume $\pi^\star \in \Pi$)

MDP and Expert

π^0

$$\ell_0(\pi) := \mathbb{E}_{s \sim d_\mu^{\pi^0}} \left[\ell \left(\pi, s, \pi^\star(s) \right) \right]$$

π^1

Online Learner w/ $\mathscr{A}$ (i.e., DAgger)

…

Total loss so far: $\ell_0(\pi^0)$

DAgger Analysis: A reduction to no-regret online learning

Decision set Π (assume $\pi^\star \in \Pi$)

MDP and Expert

Online Learner w/ $\mathscr{A}$ (i.e., DAgger)

$$\pi^0$$

$$\ell_0(\pi) := \mathbb{E}_{s \sim d_\mu^{\pi^0}} \left[\ell \left(\pi, s, \pi^\star(s) \right) \right]$$

$$\pi^1$$

$$\ell_1(\pi) := \mathbb{E}_{s \sim d_\mu^{\pi^1}} \left[\ell \left(\pi, s, \underline{\pi^\star(s)} \right) \right]$$

...

Total loss so far: $\ell_0(\pi^0)$

DAgger Analysis: A reduction to no-regret online learning

Decision set Π (assume $\pi^\star \in \Pi$)

MDP and Expert

π^0

$$\ell_0(\pi) := \mathbb{E}_{s \sim d_\mu^{\pi^0}} \left[\ell\left(\pi, s, \pi^\star(s)\right) \right]$$

π^1

$$\ell_1(\pi) := \mathbb{E}_{s \sim d_\mu^{\pi^1}} \left[\ell\left(\pi, s, \pi^\star(s)\right) \right]$$

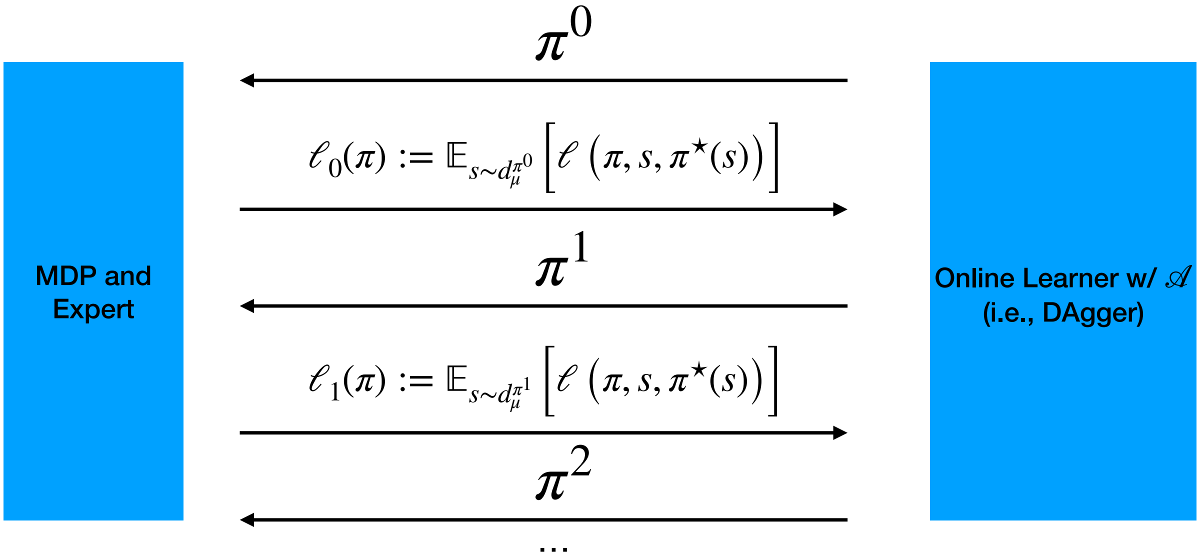
Online Learner w/ $\mathscr{A}$ (i.e., DAgger)

...

Total loss so far: $\ell_0(\pi^0) + \ell_1(\pi^1)$

DAgger Analysis: A reduction to no-regret online learning

Decision set Π (assume $\pi^\star \in \Pi$)



Total loss so far: $\ell_0(\pi^0) \quad + \ell_1(\pi^1)$

DAgger Analysis: A reduction to no-regret online learning

Interactive setting

Decision set Π (assume $\pi^\star \in \Pi$)

MDP and Expert

Online Learner w/ $\mathscr{A}$ (i.e., DAgger)

$$\pi^0$$

$$\ell_0(\pi) := \mathbb{E}_{s \sim d_\mu^{\pi^0}} \left[\ell\left(\pi, s, \pi^\star(s)\right) \right]$$

$$\pi^1$$

$$\ell_1(\pi) := \mathbb{E}_{s \sim d_\mu^{\pi^1}} \left[\ell\left(\pi, s, \pi^\star(s)\right) \right]$$

$$\pi^2$$

$$\cdots$$

Total loss so far: $\ell_0(\pi^0) \quad + \ell_1(\pi^1) \ + \ell_2(\pi^2) + \ldots$

DAgger Analysis: A reduction to no-regret online learning

After in total T many iterations, we have the following regret for DAgger:

$$\text{Avg-Regret}_T = \frac{1}{T}\left[\underbrace{\sum_{t=0}^{T-1} \ell_t(\pi^t)}_{\text{Total-loss Dagger suffered}} - \underbrace{\min_{\pi\in\Pi} \sum_{t=0}^{T-1} \ell_t(\pi)}\right] \le \underbrace{O\left(\frac{1}{\sqrt{T}}\right)}_{\epsilon_{reg}}$$

Goal: Turn ϵ_{reg} to $V^{\pi} - V^{\pi^*}$

DAgger Analysis: A reduction to no-regret online learning

After in total T many iterations, we have the following regret for DAgger:

$$\text{Avg-Regret}_T = \frac{1}{T}\left[\sum_{t=0}^{T-1} \ell_t(\pi^t) - \min_{\pi\in\Pi}\sum_{t=0}^{T-1}\ell_t(\pi)\right] \leq \underbrace{O\left(\frac{1}{\sqrt{T}}\right)}_{\epsilon_{reg}}$$

Recall we assume $\pi^\star \in \Pi$, we must have:

$$\underline{\min_{\pi\in\Pi}\sum_{t=0}^{T-1}\ell_t(\pi)} \leq \underline{\sum_{t=0}^{T-1}\ell_t(\pi^\star)} = 0$$

$$\mathbb{1}\{\pi(s) \neq \pi^*(s)\}$$

$$\ell_t(\pi) = \mathbb{E}_{s\sim d^{\pi^t}_{\mu}}\left[c(\pi, s, \pi^*(s))\right]$$

$$\ell_t(\pi^*) = 0$$

DAgger Analysis: A reduction to no-regret online learning

$\frac{1}{T}\sum \ell_t(\pi^t) \leq \min_{\pi\in\Pi} \sum \ell_t(\pi) + \epsilon_{reg} \leq 0 + \epsilon_{reg}$

After in total T many iterations, we have the following regret for DAgger:

$$\text{Avg-Regret}_T = \frac{1}{T}\left[\sum_{t=0}^{T-1} \ell_t(\pi^t) - \min_{\pi\in\Pi}\sum_{t=0}^{T-1}\ell_t(\pi)\right] \leq \underbrace{O\left(\frac{1}{\sqrt{T}}\right)}_{\epsilon_{reg}}$$

$\{\pi, \pi^*\}$

$\min_{\hat{\pi}\in\{\pi,\pi^*\}} \ell(\hat{\pi}) \leq \ell(\pi^*)$

Recall we assume $\pi^\star \in \Pi$, we must have:

$$\min_{\pi\in\Pi}\sum_{t=0}^{T-1}\ell_t(\pi) \leq \sum_{t=0}^{T-1}\ell_t(\pi^\star) = 0$$

$\frac{1}{T}\sum_{t=0}^{T-1} \ell_t(\pi^t)$

$\leq 0 + \epsilon_{reg}$

min $\leq$ Avg

Which implies that:

$$\min_{t\in\{0\ldots T-1\}} \ell_t(\pi^t) \leq \frac{1}{T}\sum_{t=0}^{T-1}\ell_t(\pi^t) \leq \epsilon_{reg}$$

DAgger Analysis: A reduction to no-regret online learning

Summary so far: we know that there must exists $t \in \{0,\ldots,T-1\}$, such that:

$$\ell_t\left(\pi^t\right) \le \epsilon_{reg}$$

$$\left(\min_{t \in \{0,\ldots,T-1\}} \ell_t(\pi^t) \le \epsilon_{reg} \right)$$

DAgger Analysis: A reduction to no-regret online learning

Summary so far: we know that there must exists $t \in \{0, \ldots, T-1\}$, such that:

$$\ell_t\left(\pi^t\right) \leq \epsilon_{reg}$$

Recall the definition of $\ell_t(\pi^t)$

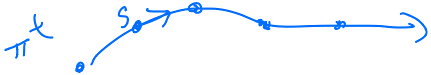
$$\ell_t\left(\pi^t\right) = \underline{\mathbb{E}_{s \sim d_\mu^{\pi^t}}}\left[c\left(\pi^{t}, s, \pi^\star(s)\right)\right] \leq \epsilon_{reg}$$

DAgger Analysis: A reduction to no-regret online learning

Summary so far: we know that there must exists $t \in \{0,\ldots,T-1\},$ such that:

$$\ell_t\left(\pi^t\right) \leq \epsilon_{reg}$$

$\pi^t(s) \approx \pi^\star(s)$



Recall the definition of $\ell_t(\pi^t)$

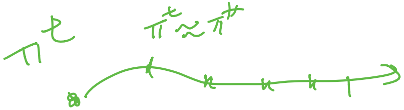
$$\ell_t\left(\pi^t\right) = \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[c\left(\pi^t, s, \pi^\star(s)\right)\right] \leq \epsilon_{reg}$$

π^t matches to $\pi^\star$ under its own state distribution!

DAgger Analysis: A reduction to no-regret online learning

Summary so far: we know that there must exists $t \in \{0,\ldots,T-1\}$, such that:

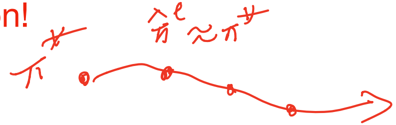
$$\ell_t\left(\pi^t\right) \leq \epsilon_{reg}$$



Recall the definition of $\ell_t(\pi^t)$

$$\ell_t\left(\pi^t\right) = \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[c\left(\pi^t, s, \pi^\star(s)\right)\right] \leq \epsilon_{reg}$$

π^t matches to $\pi^\star$ under its own state distribution!



Behavior-cloning

Recall BC, we had:

$\mathbb{E}_{s\sim d^{\pi^\star}}\left[c(\widehat{\pi}, s, \pi^\star(s))\right] \leq \epsilon$, i.e., we matched to $\pi^\star$ under $\pi^\star$'s distribution

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

$$V^{\pi^\star} - V^{\pi^t} \leq \frac{\max_{s,a} \left| A^{\pi^\star}(s,a) \right|}{\underline{1-\gamma}} \cdot \underline{\epsilon_{reg}} \leftarrow \frac{1}{\sqrt{T}}$$

This bound indicates that:

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

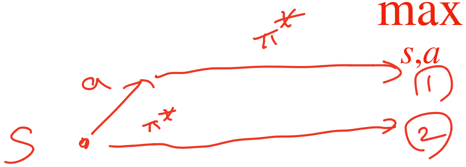
$$V^{\pi^\star} - V^{\pi^t} \leq \frac{\max_{s,a} \left| A^{\pi^\star}(s,a) \right|}{1-\gamma} \cdot \epsilon_{reg}$$

This bound indicates that:

We **avoid quadratic error** if expert $\pi^\star$ can quickly recover from a mistake

$$\max_{s,a} |A^{\pi^\star}(s,a)| \leq c \in \mathbb{R}^+$$

← small number

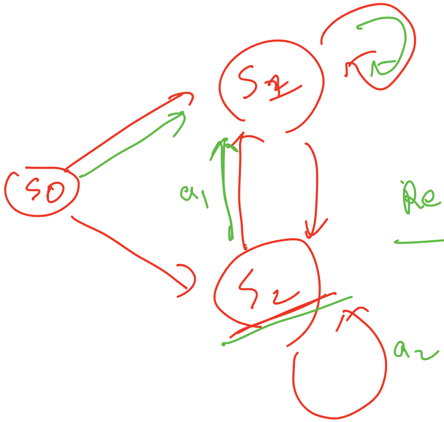


| 1 - 2 | is small

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

$$V^{\pi^\star} - V^{\pi^t} \leq \frac{\max_{s,a} \left| A^{\pi^\star}(s,a) \right|}{1-\gamma} \cdot \epsilon_{reg} \quad \leq \frac{c}{1-\gamma} \varepsilon_{reg}$$



Recover

$$Q^{\pi^\star}(s_2, a_2) - Q^{\pi^\star}(s_2, a_1) \approx 1$$

This bound indicates that:

We **avoid quadratic error** if expert $\pi^\star$ can quickly recover from a mistake

$$\max_{s,a} |A^{\pi^\star}(s,a)| \leq c \in \mathbb{R}^+$$

i.e., at s, taking a then following $\pi^\star$ is almost as good as following $\pi^\star$ directly

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

$$V^{\pi^\star} - V^{\pi^t} \leq \frac{\max_{s,a} \left| A^{\pi^\star}(s,a) \right|}{1-\gamma} \cdot \epsilon_{reg}$$

PDL

$$V^{\pi^t} - V^{\pi^\star} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}_{\mu}} \left[A^{\pi^\star}(s, \pi^t(s)) \right]$$

This bound indicates that:

We **avoid quadratic error** if expert $\pi^\star$ can quickly recover from a mistake

$$\max_{s,a} |A^{\pi^\star}(s,a)| \leq c \in \mathbb{R}^+$$

i.e., at s, taking a then following $\pi^\star$ is almost as good as following $\pi^\star$ directly

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

$$V^{\pi^\star} - V^{\pi^t} \leq \frac{\max_{s,a} \left| A^{\pi^\star}(s,a) \right|}{1-\gamma} \cdot \epsilon_{reg}$$

$f:$

$|f(x) - f(y)|$

$\leq \max_z |f(z)| \, 1(x \neq y)$

$$V^{\pi^t} - V^{\pi^\star} = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^\pi}\left[A^{\pi^\star}(s, \pi^t(s))\right]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^\pi}\left[A^{\pi^\star}(s, \pi^t(s)) - \underbrace{A^{\pi^\star}(s, \pi^\star(s))}_{=0}\right]$$

$A^{\pi^\star}(s, \pi^\star(s))$

$= Q^{\pi^\star}(s, \pi^\star(s))$

$- V^{\pi^\star}(s)$

$= 0$

$\left| A^{\pi^\star}(s, \underline{\pi^t(s)}) - A^{\pi^\star}(s, \underline{\pi^\star(s)}) \right|$

$\leq \max_{s,a} \left| A^{\pi^\star}(s,a) \right| 1\{\pi^t(s) \neq \pi^\star(s)\}$

This bound indicates that:

We **avoid quadratic error** if expert $\pi^\star$ can quickly recover from a mistake

$$\max_{s,a} |A^{\pi^\star}(s,a)| \leq c \in \mathbb{R}^+$$

i.e., at s, taking a then following $\pi^\star$ is almost as good as following $\pi^\star$ directly

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

$$V^{\pi^\star} - V^{\pi^t} \leq \frac{\max_{s,a} \left| A^{\pi^\star}(s,a) \right|}{1-\gamma} \cdot \epsilon_{reg}$$

$$V^{\pi^t} - V^{\pi^\star} = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^\pi} \left[A^{\pi^\star}(s, \pi^t(s)) \right]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^\pi} \left[A^{\pi^\star}(s, \pi^t(s)) - A^{\pi^\star}(s, \pi^\star(s)) \right]$$

$\geq - \max_{s,a} \left| A^{\pi^\star}(s,a) \right| \mathbb{1}\left(\pi^t(s) \neq \pi^\star(s)\right)$

$$\geq \frac{-1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi^t}} \left[\max_{s,a} \left| A^{\pi^\star}(s,a) \right| \mathbf{1}\{\pi^t(s) \neq \pi^\star(s)\} \right]$$

from Dagger: $\ell_t(\pi^t) \leq \epsilon_{Reg}$

$\Rightarrow \mathbb{E}_{s\sim d_\mu^{\pi^t}} \mathbb{1}\left[\pi^t(s) \neq \pi^\star(s)\right] \leq \epsilon_{Reg}$

This bound indicates that:

We **avoid quadratic error** if expert $\pi^\star$ can quickly recover from a mistake

$$\max_{s,a} \left| A^{\pi^\star}(s,a) \right| \leq c \in \mathbb{R}^+$$

i.e., at s, taking a then following $\pi^\star$ is almost as good as following $\pi^\star$ directly

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

$$V^{\pi^\star} - V^{\pi^t} \leq \frac{\max_{s,a} \left| A^{\pi^\star}(s,a) \right|}{1-\gamma} \cdot \epsilon_{reg}$$

$$V^{\pi^t} - V^{\pi^\star} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^\pi} \left[A^{\pi^\star}(s, \pi^t(s)) \right]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^\pi} \left[A^{\pi^\star}(s, \pi^t(s)) - A^{\pi^\star}(s, \pi^\star(s)) \right]$$

$$\geq \frac{-1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[\max_{s,a} \left| A^{\pi^\star}(s,a) \right| \mathbf{1}\{\pi^t(s) \neq \pi^\star(s)\} \right] \geq -\frac{\epsilon_{reg}}{1-\gamma} \cdot \max_{sa} \left[A^{\pi^\star}(s,a) \right]$$

$$V^{\pi^\star} - V^{\pi^t} \leq \frac{\max_{s,a} \left| A^{\pi^\star}(s,a) \right|}{1-\gamma} \cdot \epsilon_{reg}$$

This bound indicates that:

We **avoid quadratic error** if expert $\pi^\star$ can quickly recover from a mistake

$$\max_{s,a} \left| A^{\pi^\star}(s,a) \right| \leq c \in \mathbb{R}^+$$

i.e., at s, taking a then following $\pi^\star$ is almost as good as following $\pi^\star$ directly

Summary of DAgger

Summary of DAgger

DAgger finds a policy $\widehat{\pi}$ such that it matches to $\pi^\star$ under $d_\mu^{\widehat{\pi}}$

$$\mathbb{E}_{s\sim d_\mu^{\widehat{\pi}}}\left[\mathbf{1}\{\widehat{\pi}(s) \neq \pi^\star(s)\}\right] \leq \epsilon_{reg} = O(1/\sqrt{T})$$

No-Regret
Argument
(FTRL)

Summary of DAgger

DAgger finds a policy $\widehat{\pi}$ such that it matches to $\pi^\star$ under $d_\mu^{\widehat{\pi}}$

$$\mathbb{E}_{s \sim d_\mu^{\widehat{\pi}}} \left[\mathbf{1}\{ \widehat{\pi}(s) \neq \pi^\star(s) \} \right] \leq \epsilon_{reg} = O(1/\sqrt{T})$$

If expert herself can quickly recover from a deviation, i.e., $\underbrace{|Q^{\pi^\star}(s,a) - V^{\pi^\star}(s)|}_{A^{\pi^\star}(s,a)}$ is small for all s,

$$V^{\pi^\star} - V^{\pi^t} \leq O\left(\frac{1}{1-\gamma} \cdot \epsilon_{reg} \right)$$

$$\mathbb{E}_{s \sim d^{\pi^\star}} \left[1(\widehat{\pi}(s) \neq \pi^\star(s)) \right] \leq \varepsilon$$

$$\mathbb{E}_{s \sim d_\mu^{\pi^\star}} \left[A^{\widehat{\pi}}(s, \pi^\star(s)) \right]$$

Today's Plan

1. Finish DAgger's Analysis

2. Intro to Maximum Entropy Inverse RL
(We have offline demonstrations, but learner can interact with the environments)

Review of the IL settings that we covered so far

1. Offline IL Setting:

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

No expert interaction, no real world interaction

Review of the IL settings that we covered so far

1. Offline IL Setting:

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

No expert interaction, no real world interaction

2. Interactive IL setting:

We have access to $\pi^\star$ during training

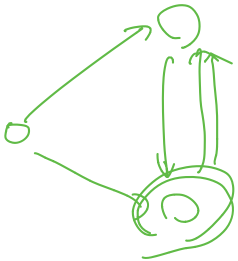
Interaction w/ expert and interaction w/ the world (i.e., we can try out our policies)

A new setting (more realistic maybe??)

Hybrid:

1. We have an offline dataset $\mathscr{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d^{\pi^\star}$ (e.g., a pre-collected demonstrations)

2. And we can interact with the world (e.g., try out our policy and see what happens)



Running Example: Human trajectory forecasting

[Kitani, et al, ECCV 12]

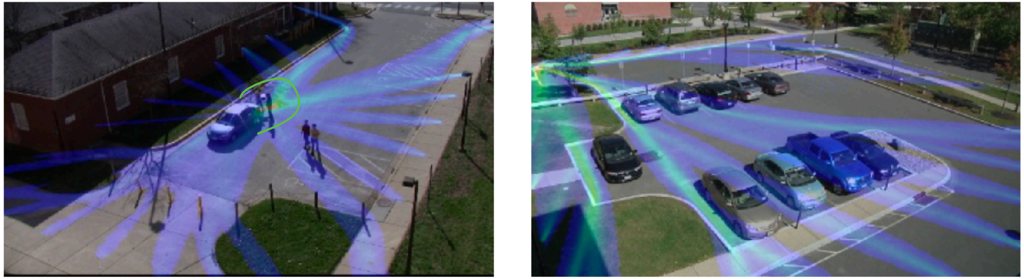


Fig. 1. Given a single pedestrian detection, our proposed approach forecasts plausible paths and destinations from noisy vision-input

Running Example: Human trajectory forecasting

[Kitani, et al, ECCV 12]

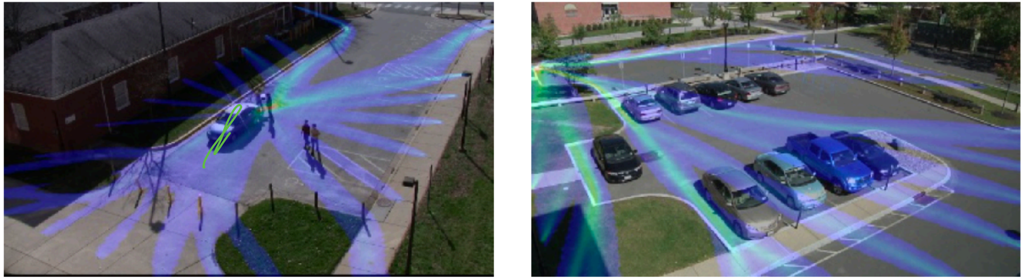


Fig. 1. Given a single pedestrian detection, our proposed approach forecasts plausible paths and destinations from noisy vision-input

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

Setting

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

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

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

(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** $\phi(s, a)$

Running Example: Define feature map

Key Assumption on cost:

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

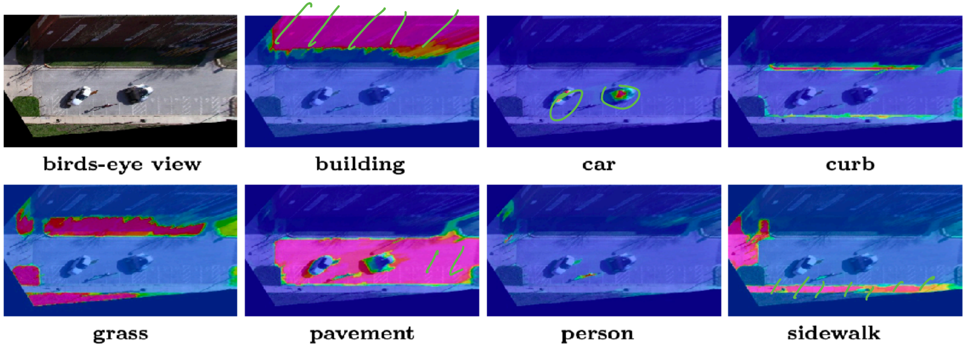


Fig. 4. Classifier feature response maps. Top left is the original image.

Running Example: Define feature map

Key Assumption on cost:

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

pixel

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

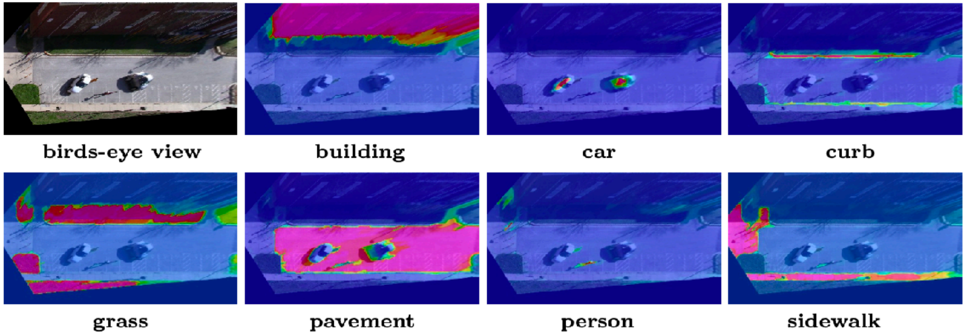


Fig. 4. Classifier feature response maps. Top left is the original image.

Running Example: Define feature map

Key Assumption on cost:

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

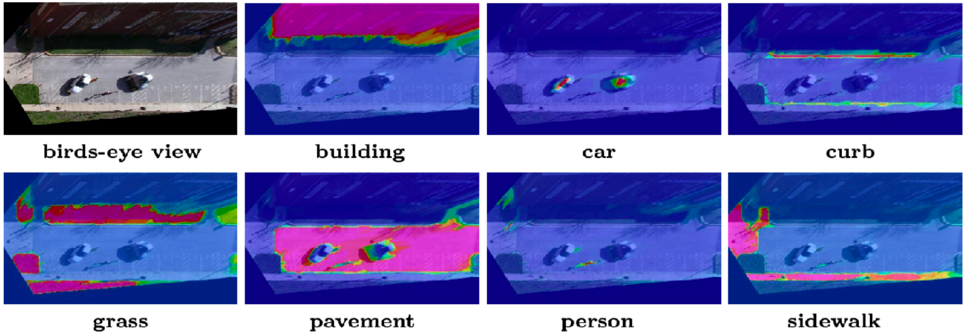


Fig. 4. Classifier feature response maps. Top left is the original image.

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

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

pixel

Running Example: Define feature map

Key Assumption on cost:

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

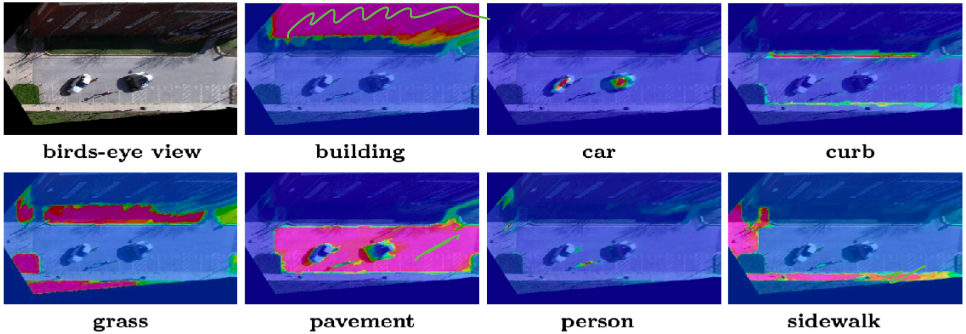


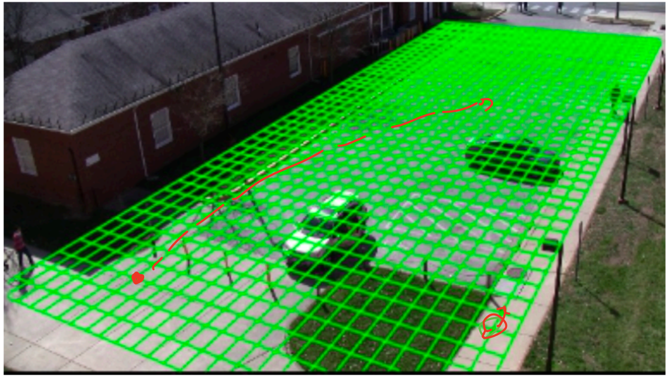
Fig. 4. Classifier feature response maps. Top left is the original image.

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

Running Example: Human Trajectory Forecasting



State space: grid,
action space: 4 actions

We predict that we are more likely to use sidewalk

We will talk about the algorithm (MaxEnt-IRL) behind it next week