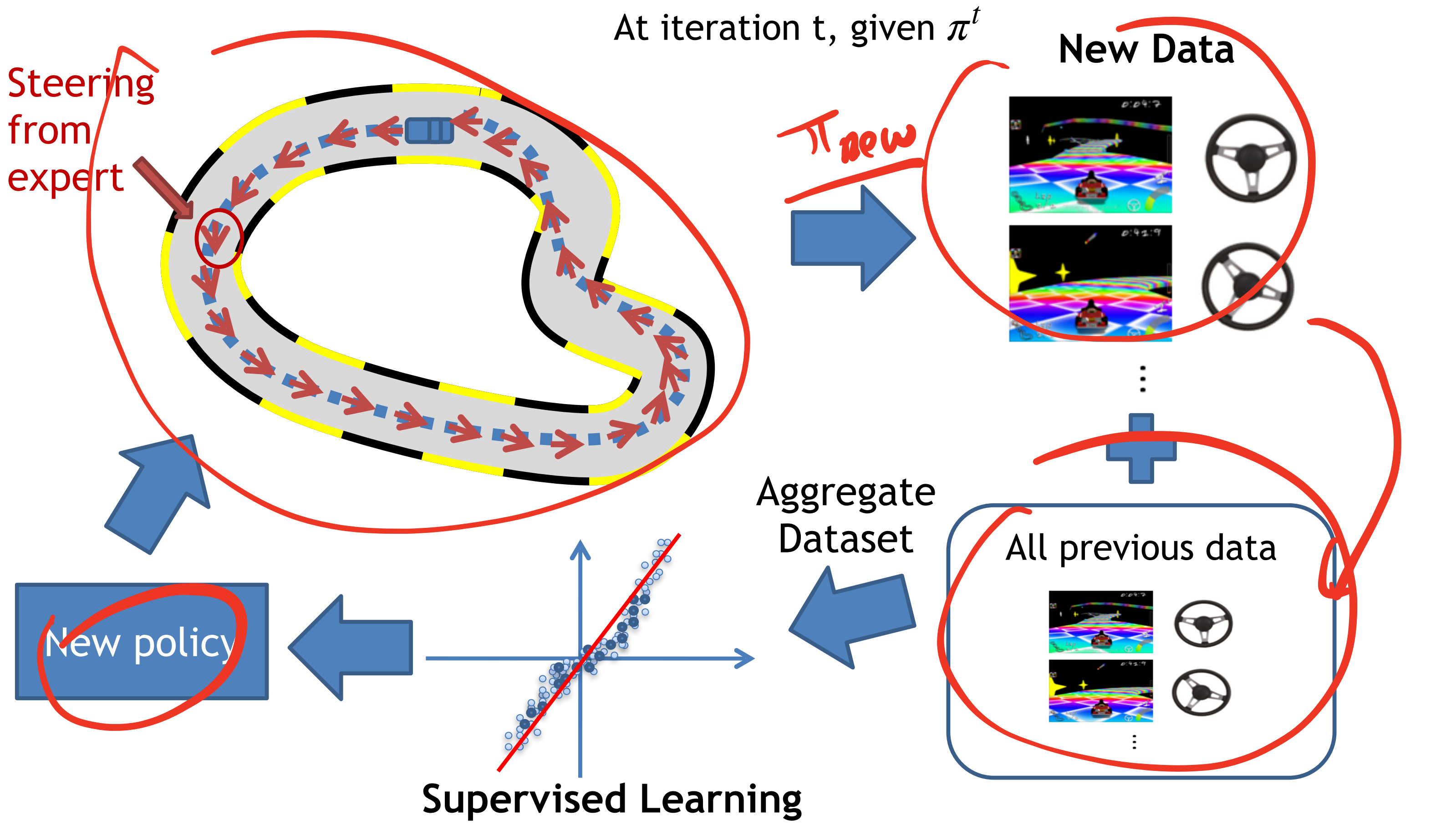
Interactive Imitation Learning (continue)

Recap

Interactive Imitation Learning Setting

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

DAgger Revisit



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 $\ell_0, \ldots \ell_{t-1}, \ell_t$, learner updates to a new decision:

$$\textbf{FTL: } \theta_{t+1} = \min_{\theta \in \Theta} \sum_{i=0}^{t} \ell_i(\theta) + \lambda R(\theta) \qquad \|\theta\|_2^2$$

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:

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

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:

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

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 policy is a binary classifier)

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

DAgger Analysis: A reduction to no-regret online learning

infinite horizon MDP

(assume discrete action space—**in fact let's assume 2 actions**, so policy is a binary classifier)

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

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

Classification:

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

$$\hat{\pi} = \arg\min_{\pi} \sum_{x,y} \left[\ell(\pi, x, y) \right]$$

loss

DAgger Analysis: A reduction to no-regret online learning

infinite horizon MDP

(assume discrete action space—**in fact let's assume 2 actions**, so policy is a binary classifier)

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

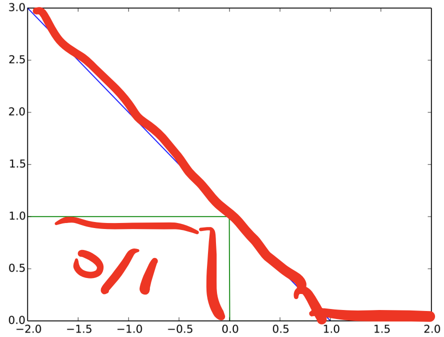
Classification:

SUM

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

$$\hat{\pi} = \arg\min_{\pi} \sum_{x,y} \left[\ell\left(\pi, x, y\right)\right]$$

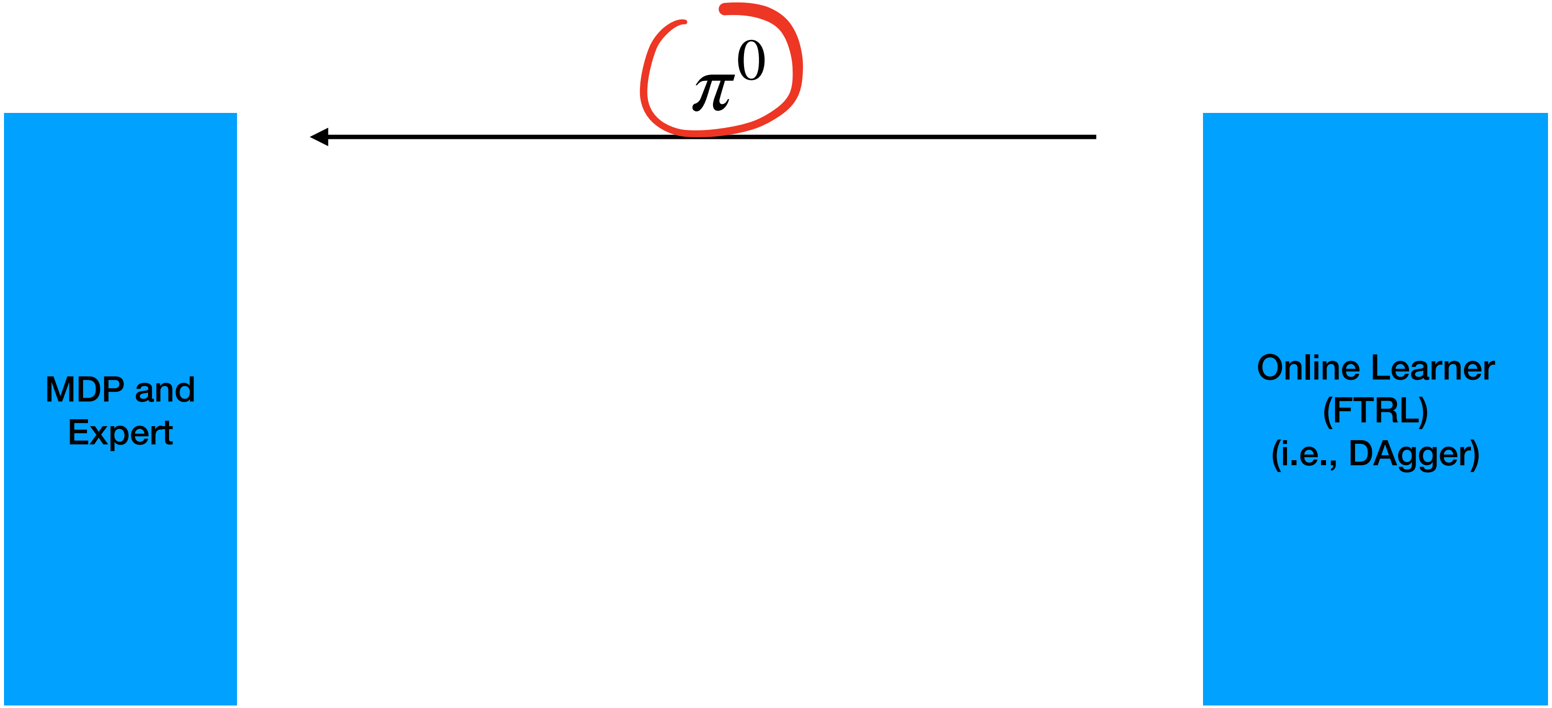
$$\ell(\pi, x, y) = \max\{0, 1 - \pi(x) \cdot y\}$$



DAgger Analysis: A reduction to no-regret online learning

Total loss so far:

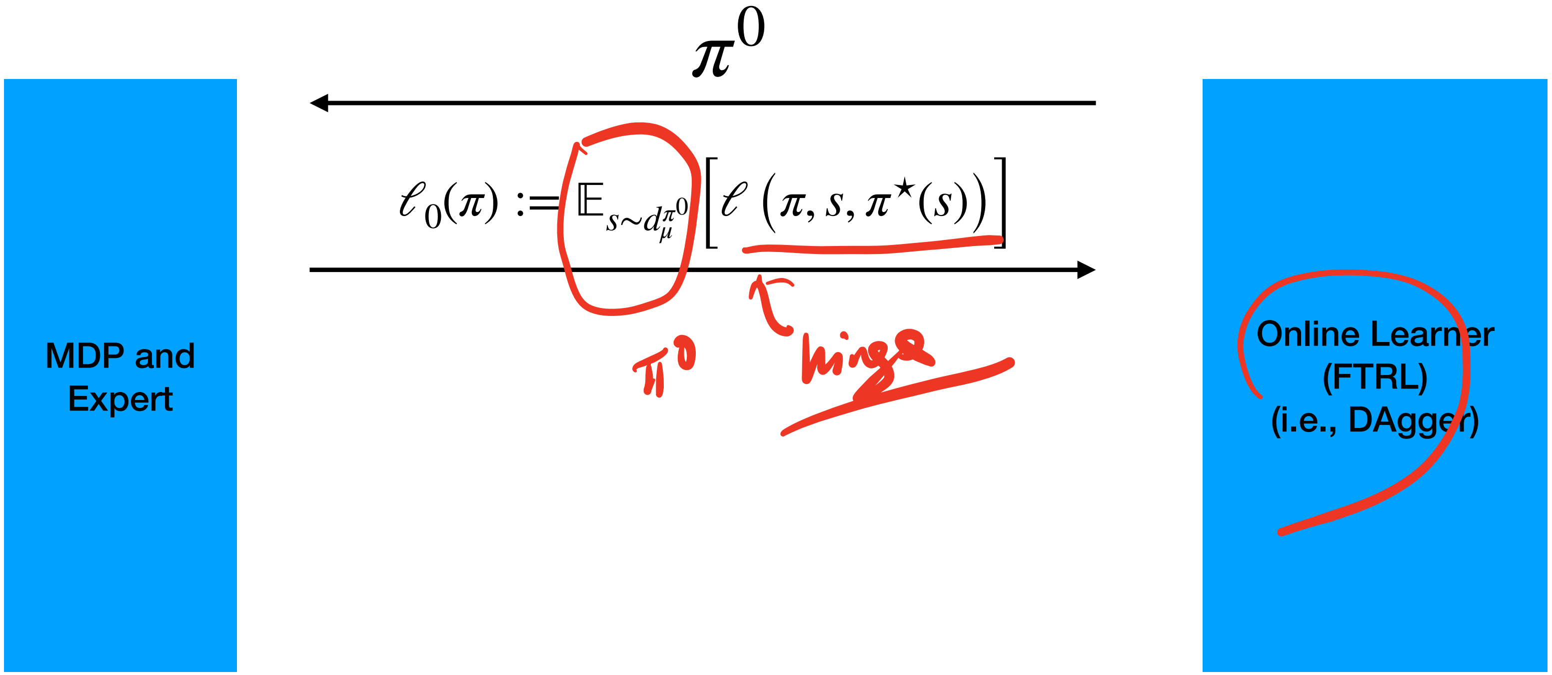
DAgger Analysis: A reduction to no-regret online learning



…

Total loss so far:

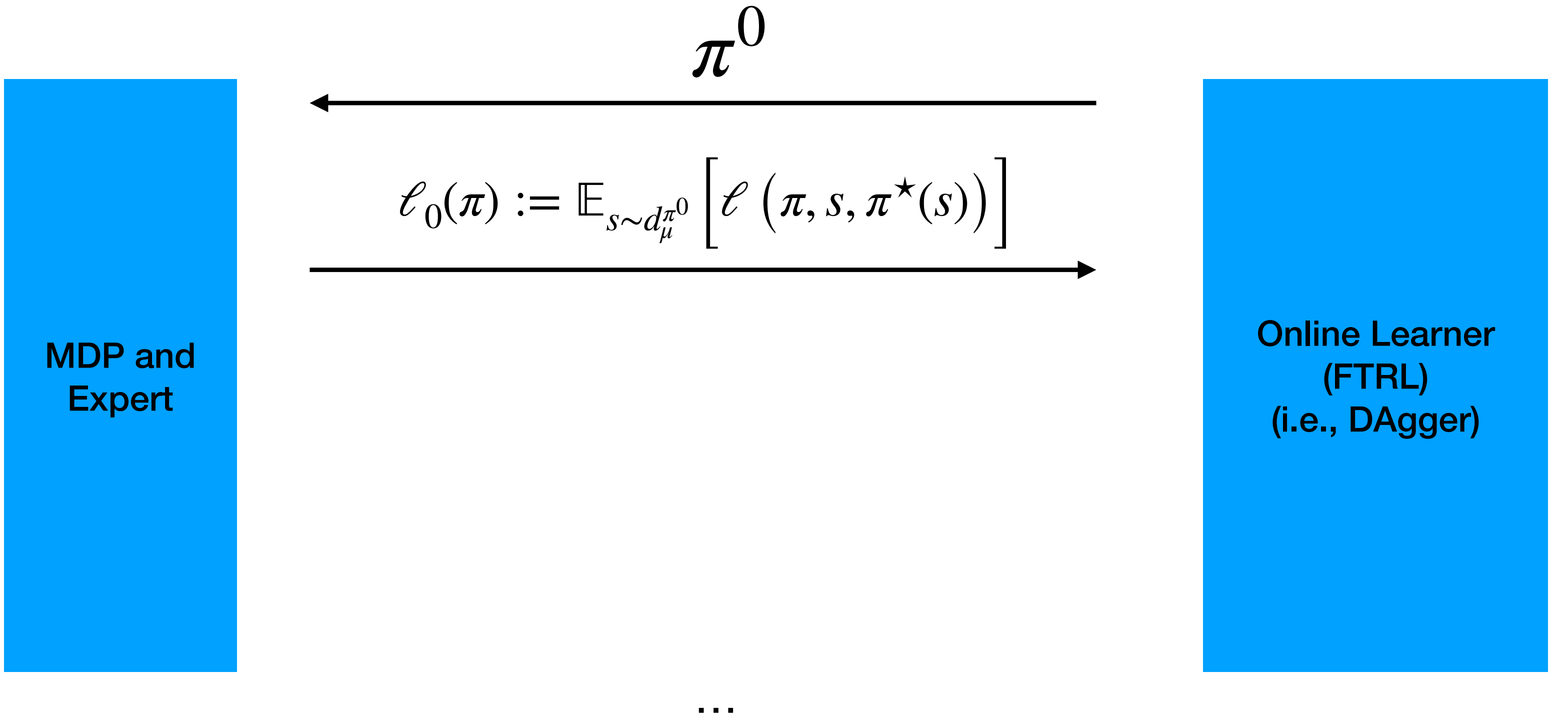
DAgger Analysis: A reduction to no-regret online learning



…

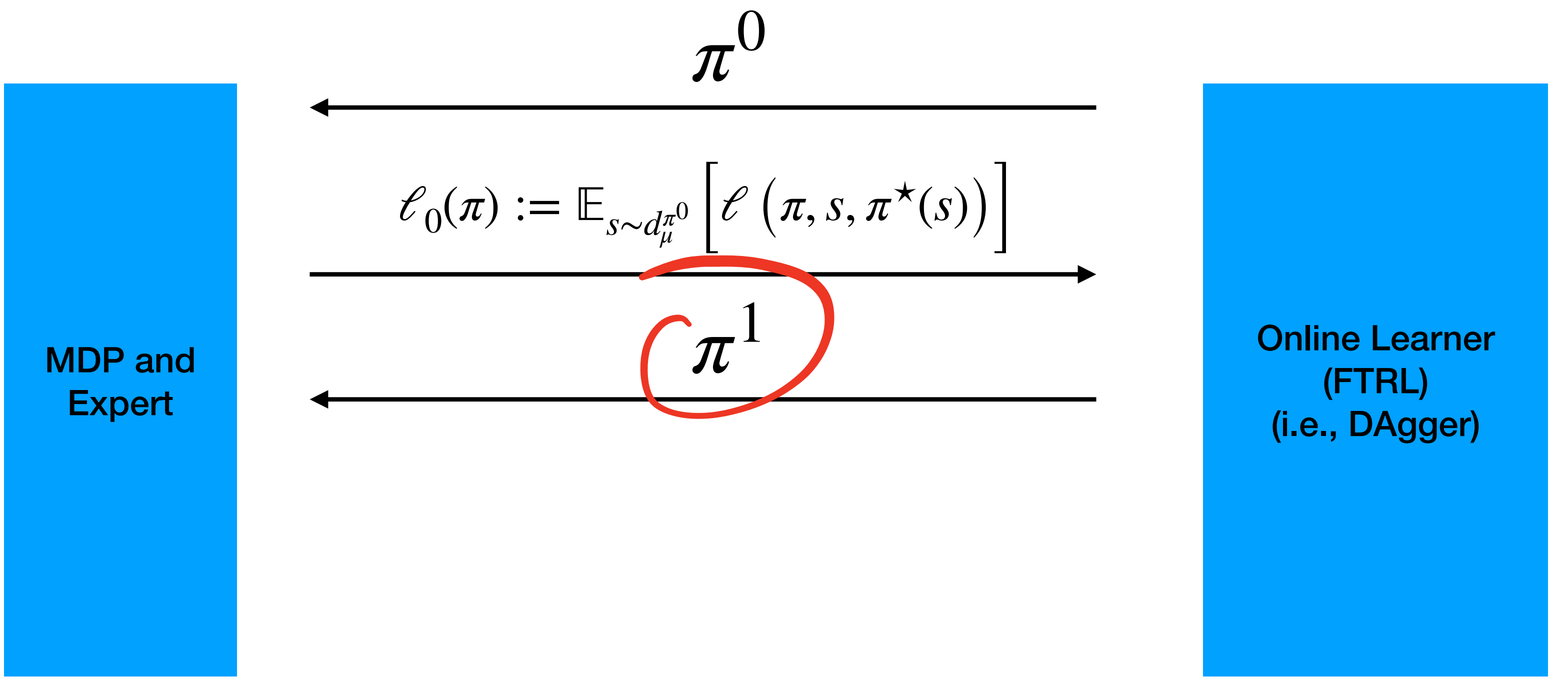
Total loss so far:

DAgger Analysis: A reduction to no-regret online learning



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

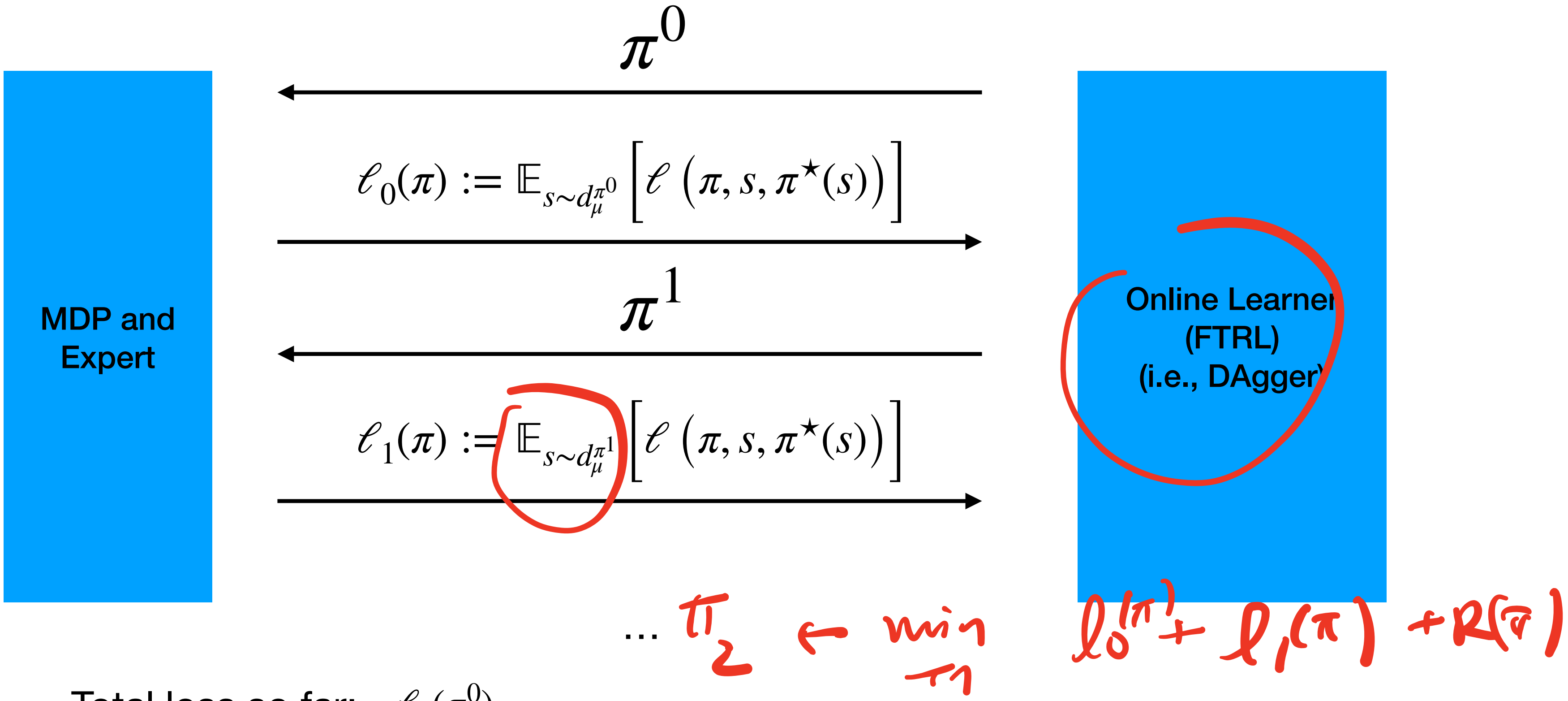
DAgger Analysis: A reduction to no-regret online learning



…

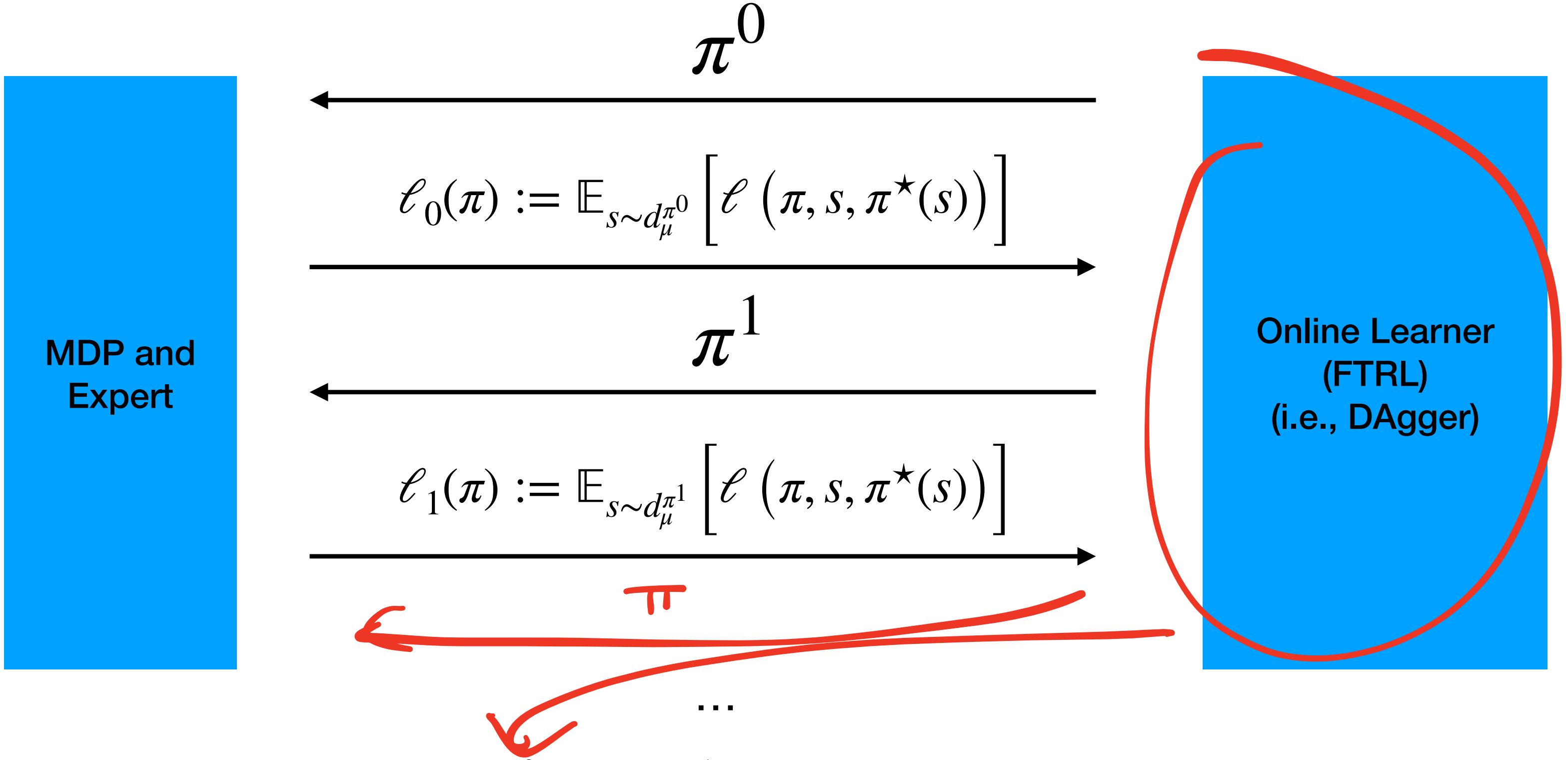
Total loss so far: $\ell_0(\pi^0)$

DAgger Analysis: A reduction to no-regret online learning



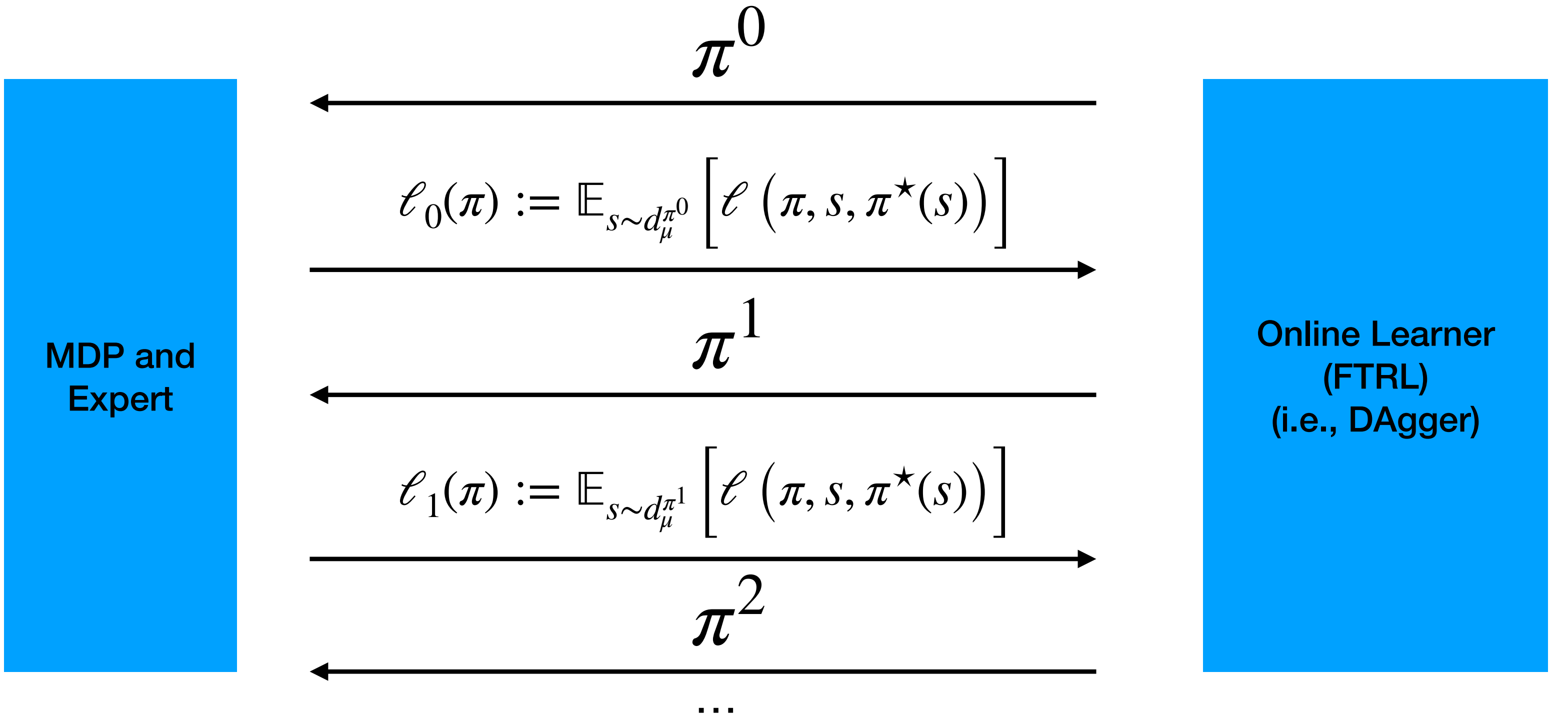
Total loss so far: $\ell_0(\pi^0)$

DAgger Analysis: A reduction to no-regret online learning



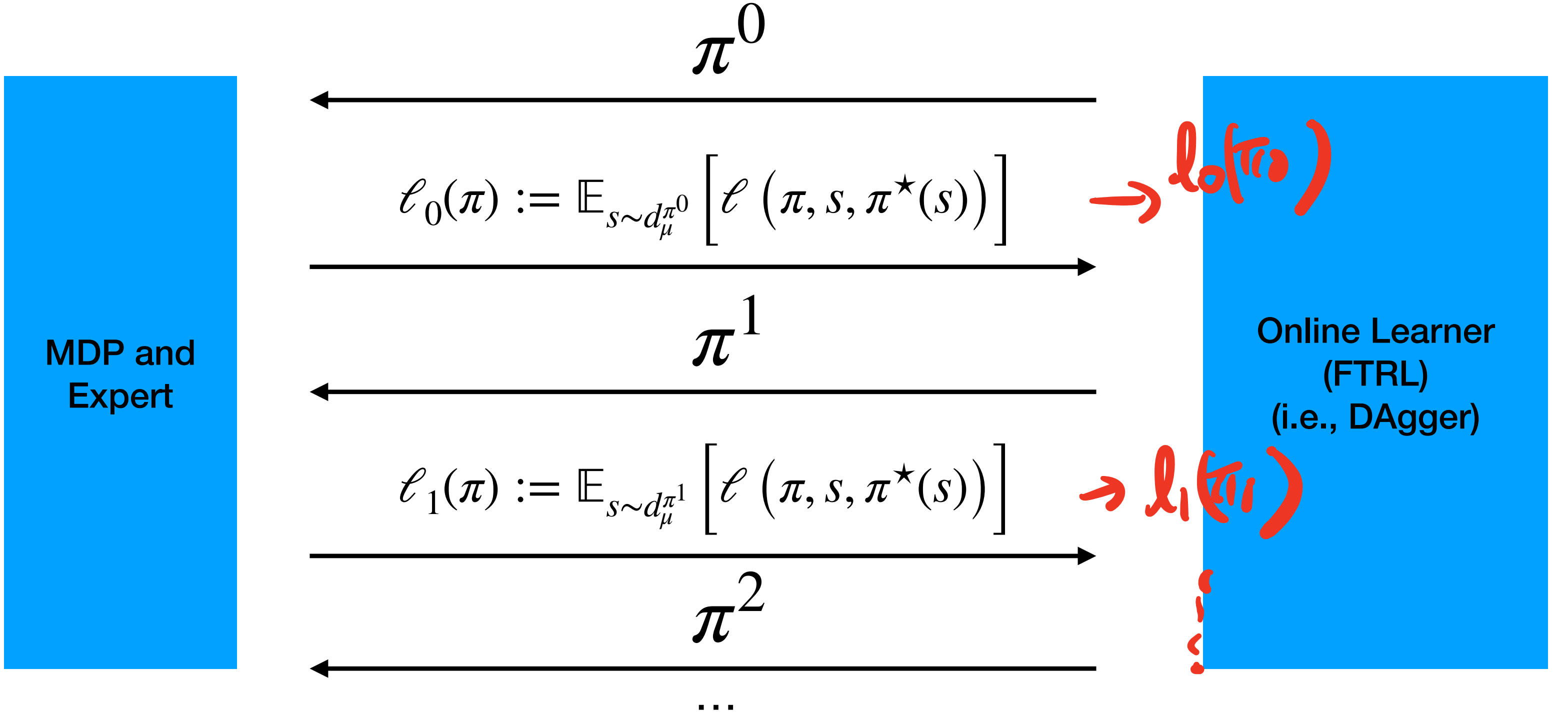
Total loss so far: $\ell_0(\pi^0) \quad + \ell_1(\pi^1)$

DAgger Analysis: A reduction to no-regret online learning



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

DAgger Analysis: A reduction to no-regret online learning



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:

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

DAgger Analysis: A reduction to no-regret online learning

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

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

$= 0$

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

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

$\ell(\pi(s), \pi^\star(s))$

$1(\pi(s) \neq \pi^\star(s))$

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:

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

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:

$$\frac{1}{T}\sum_{t=1}^{T} \ell_t(\pi_t) \leq \epsilon_{reg} \rightarrow \quad \ell_t(\pi^t) \leq \epsilon_{reg}$$

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:

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

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

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

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:

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

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

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

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 $\ell_t(\pi^t)$

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

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

Recall BC, we had:

$\mathbb{E}_{s \sim d^{\pi^\star}}\left[\ell(\widehat{\pi}, s, \pi^\star(s))\right] \leq \epsilon$, 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:

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

max absolute value of Adv of π^*

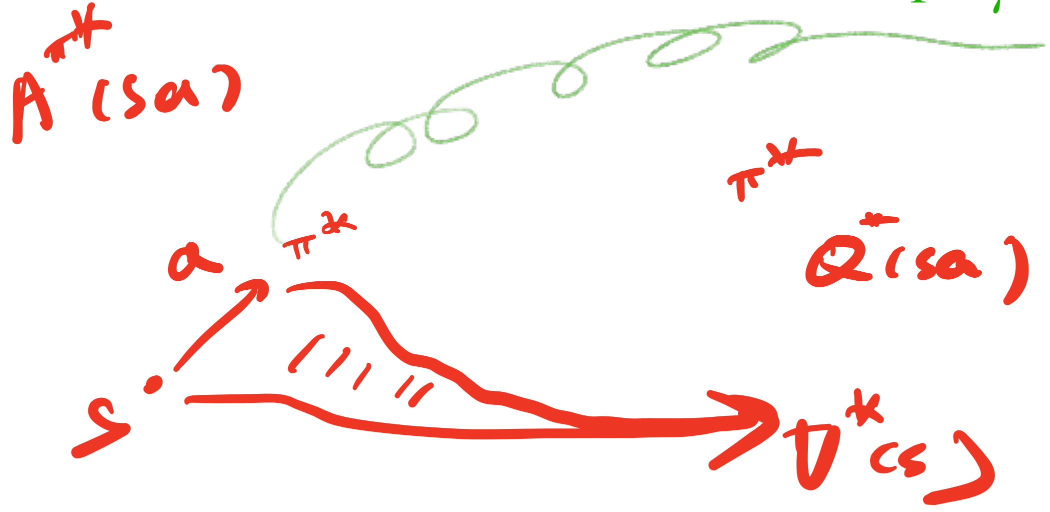
$A^{\pi^*} = Q^{\pi^*} - V^{\pi^*}$

This bound indicates that:

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

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



$$\left| Q^*(s,a) - V^*(s) \right| \leq c$$

This bound indicates that:

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

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

$$c \leq \frac{1}{1-\gamma}$$

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

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

i.e., at any state, $\pi^\star$ can quickly recover from your mistake (take action a)

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

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

$$A^{\pi^\star}(s, \pi^\star(s)) = 0$$

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 any state, $\pi^\star$ can quickly recover from your mistake (take action a)

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

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

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

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

// 0

Case 1, $\pi^t(s) = \pi^*(s)$

Case $\pi^t(s) \neq \pi^*(s)$

$A^{\pi^*}(s, \pi^t(s)) - A^{\pi^*}(s, \pi^*(s)) \geq -\max_{s,a} |A^{\pi^*}(s,a)|$

$= 0$

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 any state, $\pi^\star$ can quickly recover from your mistake (take action a)

Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration t, such that:

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

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

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

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

$$\mathbb{E}_{s\sim d^{\pi^t}} \mathbf{1}(\pi^\star(s) \neq \pi^t(s)) \approx \varepsilon_{reg}$$

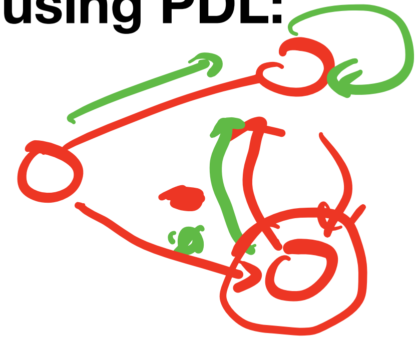
This bound indicates that:

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

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

i.e., at any state, $\pi^\star$ can quickly recover from your mistake (take action a)

Finally, turn things into the performance bound using PDL:



Theorem: There exists a iteration t, such that:

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

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

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

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

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

← FTRL

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 any state, $\pi^\star$ can quickly recover from your mistake (take action a)

Summary of DAgger

Summary of DAgger

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

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

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

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

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

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]

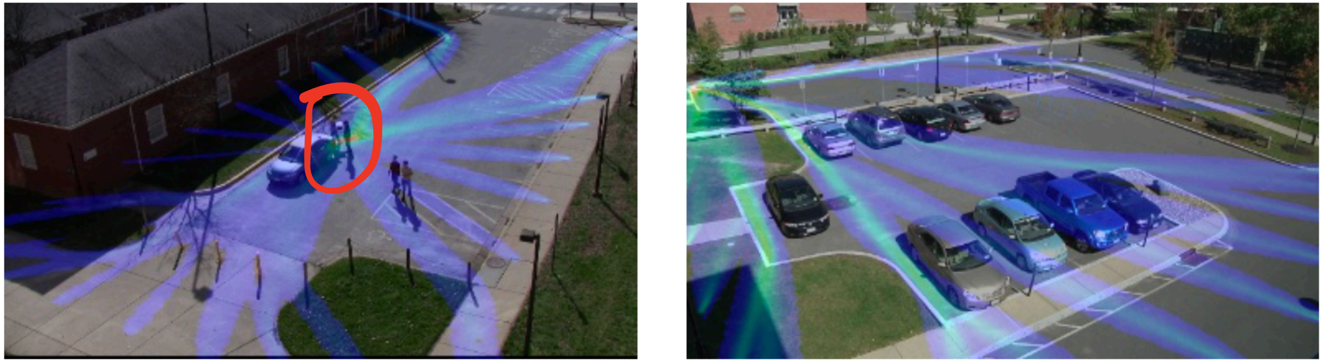


Fig. 1. Given a single pedestrian detection, our proposed approach forecasts plausible paths and destinations from noisy vision-input

Running Example: Human trajectory forecasting

[Kitani, et al, ECCV 12]

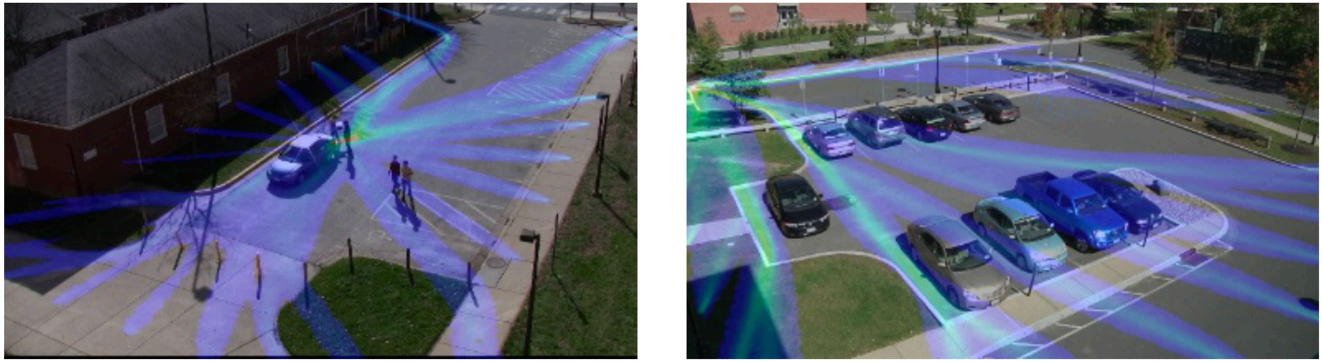


Fig. 1. Given a single pedestrian detection, our proposed approach forecasts plausible paths and destinations from noisy vision-input

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** $\phi(s, a)$

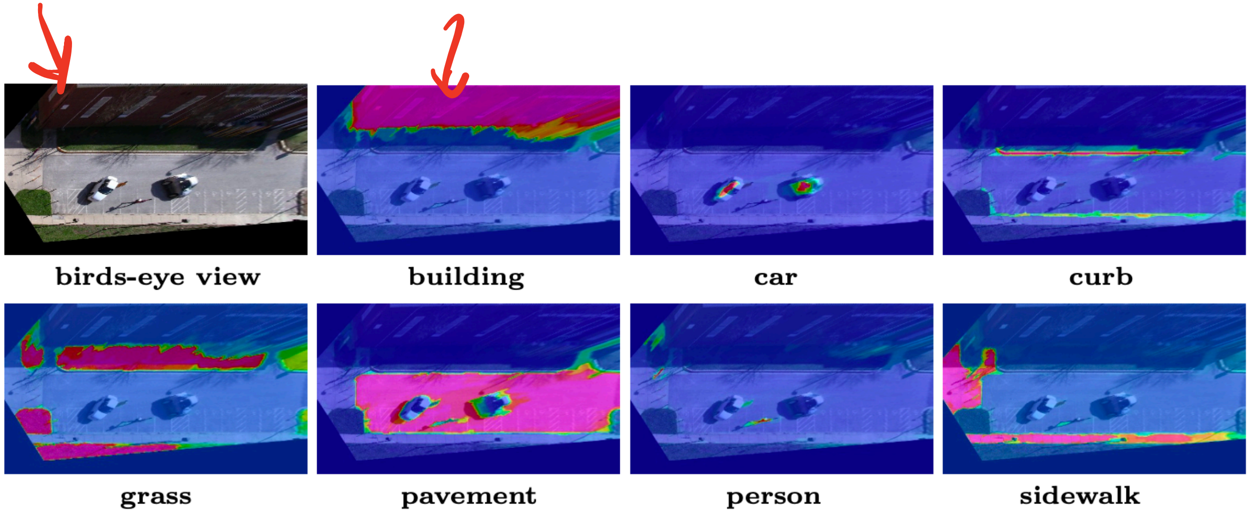


Fig. 4. Classifier feature response maps. Top left is the original image.

Running Example: Define feature map

Key Assumption on cost:

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

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

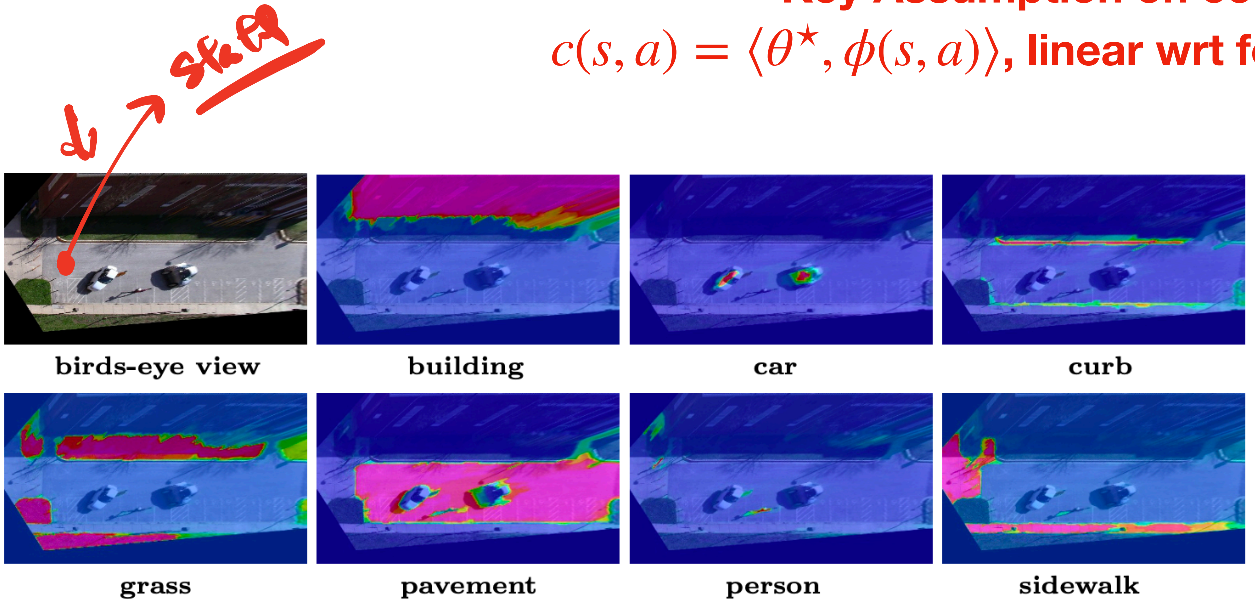


Fig. 4. Classifier feature response maps. Top left is the original image.

Running Example: Define feature map

Key Assumption on cost:

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

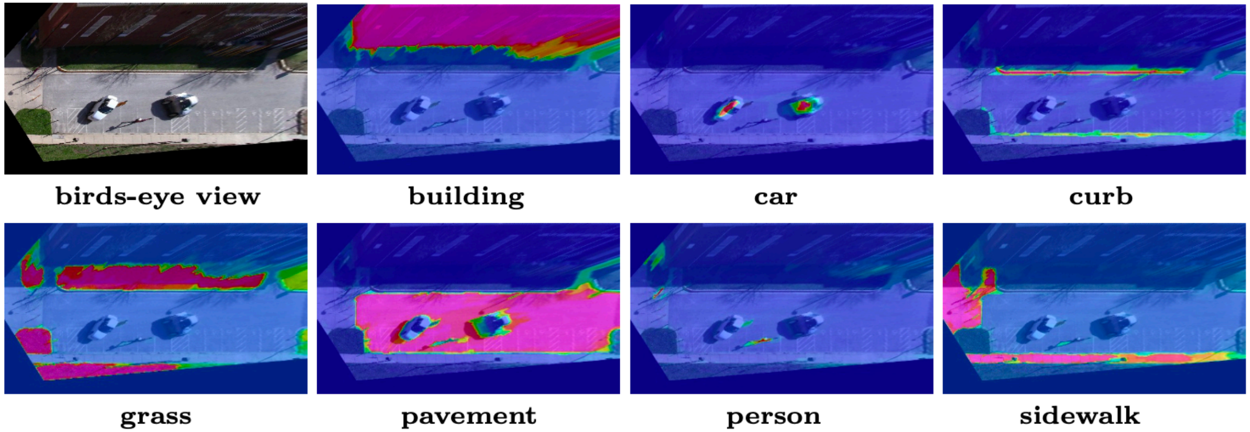


Fig. 4. Classifier feature response maps. Top left is the original image.

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

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

$\theta^{\star\top} \phi(s,a)$

Running Example: Define feature map

Key Assumption on cost:

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

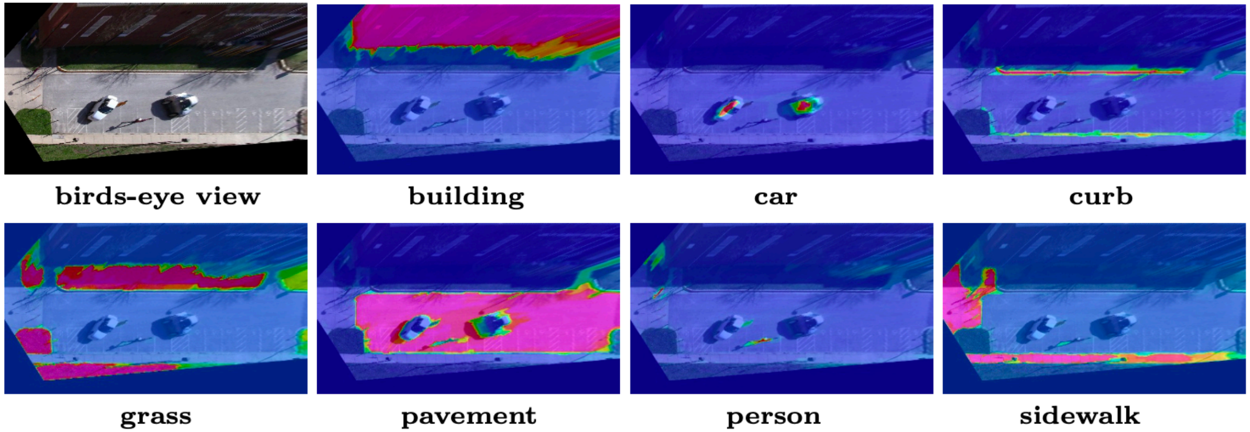


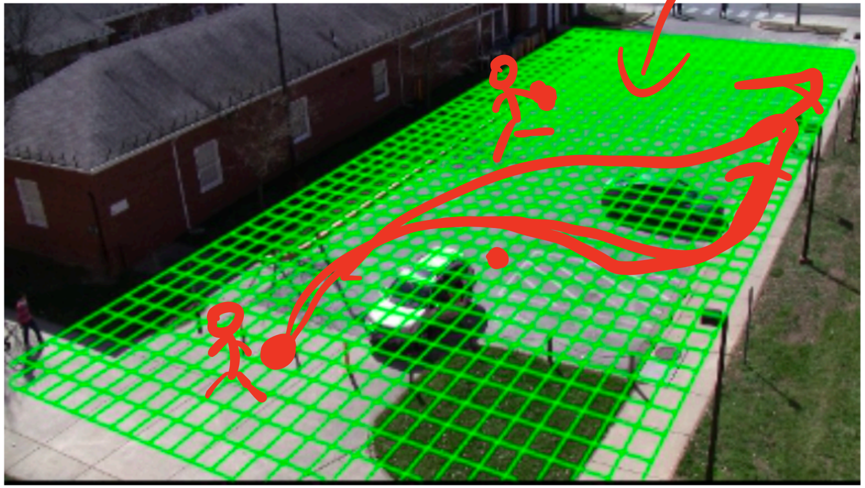
Fig. 4. Classifier feature response maps. Top left is the original image.

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

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