Interactive Imitation Learning (continue)
Recap

Interactive Imitation Learning Setting

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

At iteration $t$, given $\pi^t$

Steering from expert

New Data

Aggregate Dataset

All previous data

Supervised Learning

New policy

$\pi^t$
DAgger Revisit

At iteration $t$, given $\pi^t$

New Data

Steering from expert

Aggregate Dataset

Supervised Learning

New policy

All previous data

...
At iteration $t$, given $\pi^t$

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

DAgger Revisit

Steering from expert

New policy

Supervised Learning

Aggregate Dataset

All previous data

New Data

$$\ell_t(\pi) = \mathbb{E}_{s \sim d_{\pi^t}} \left[ c \left( \pi, s, \pi^*(s) \right) \right]$$
At iteration $t$, given $\pi^t$:

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

**Steering from expert**

**New Data**

**Supervised Learning**

**New policy**

**Aggregate Dataset**

**All previous data**

**DAgger Revisit**
At iteration $t$, given $\pi^t$:

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

New Data

Steering from expert

Aggregate Dataset

All previous data

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

Supervised Learning

New policy
DAgger Revisit

At iteration $t$, given $\pi^t$

New Data

\[
\ell_t(\pi) = \mathbb{E}_{s \sim d_{\pi^t}} [c(\pi, s, \pi^*(s))]
\]

Aggregate Dataset

All previous data

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

New policy

Steering from expert

Supervised Learning
At iteration $t$, given $\pi^t$

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

Steering from expert

New policy

New Data

Aggregate Dataset

All previous data

Supervised Learning

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

New Data:

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

Supervised Learning:

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

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:

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

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

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

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

$t \to \infty$
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)

\[ M = \{ S, A, \gamma, r, P, \mu \} \]
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)

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

**Function approximation:**

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

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

Function approximation:

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

Classification algorithm (oracle) \( \mathcal{A} \): 

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

\[ \hat{\pi} = \mathcal{A} \left( \Pi, \rho \right) := \arg \min_{\pi \in \Pi} \mathbb{E}_{x, y \sim \rho} \left[ c \left( \pi, x, y \right) \right] \]
DAgger Analysis: A reduction to no-regret online learning

Assume a finite horizon MDP (infinite horizon MDP) (assume discrete action space—in fact let’s assume 2 actions, so we do binary classification)

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

Function approximation:

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

Classification algorithm (oracle) \( \mathcal{A} \):

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

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

\[ c(\pi, x, y) = \max\{0, 1 - \pi(x) \cdot y\} \]
DAgger Analysis: A reduction to no-regret online learning

Decision set $\Pi$ (assume $\pi^* \in \Pi$)

MDP and Expert

Online Learner w/ $A$ (i.e., DAgger)

Total loss so far:
DAgger Analysis: A reduction to no-regret online learning

Decision set $\Pi$ (assume $\pi^* \in \Pi$)

$\pi^0 \in \Pi$

MDP and Expert

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

Total loss so far:
DAgger Analysis: A reduction to no-regret online learning

Decision set $\Pi$ (assume $\pi^* \in \Pi$)

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

MDP and Expert

Online Learner w/ $A$ (i.e., DAgger)

... Total loss so far:
DAgger Analysis: A reduction to no-regret online learning

Decision set $\Pi$ (assume $\pi^* \in \Pi$)

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

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

MDP and Expert

Online Learner w/ $A$ (i.e., DAgger)
DAgger Analysis: A reduction to no-regret online learning

Decision set $\Pi$ (assume $\pi^* \in \Pi$)

$$\ell_0(\pi) := \mathbb{E}_{s \sim d^\pi_0} \left[ \ell(\pi, s, \pi^*(s)) \right]$$

Total loss so far: $\ell_0(\pi^0)$
DAgger Analysis: A reduction to no-regret online learning

Decision set $\Pi$ (assume $\pi^* \in \Pi$)

$\pi^0$

$\mathcal{L}_0(\pi) := \mathbb{E}_{s \sim d_\mu^0} \left[ \mathcal{L}(\pi, s, \pi^*(s)) \right]$

$\pi^1$

$\mathcal{L}_1(\pi) := \mathbb{E}_{s \sim d_\mu^1} \left[ \mathcal{L}(\pi, s, \pi^*(s)) \right]$

MDP and Expert

Online Learner w/ $A$ (i.e., DAgger)

Total loss so far: $\mathcal{L}_0(\pi^0)$
DAgger Analysis: A reduction to no-regret online learning

Decision set $\Pi$ (assume $\pi^* \in \Pi$)

$\pi^0$

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

$\pi^1$

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

Total loss so far: $\ell_0(\pi^0) + \ell_1(\pi^1)$
DAgger Analysis: A reduction to no-regret online learning

Decision set $\Pi$ (assume $\pi^* \in \Pi$)

MDP and Expert

Online Learner w/ $\mathcal{A}$ (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 $\Pi$ (assume $\pi^* \in \Pi$)

$\pi^0$

$\mathcal{L}_0(\pi) := \mathbb{E}_{s \sim d^0_\mu} \left[ \mathcal{L}(\pi, s, \pi^*(s)) \right]$

$\pi^1$

$\mathcal{L}_1(\pi) := \mathbb{E}_{s \sim d^1_\mu} \left[ \mathcal{L}(\pi, s, \pi^*(s)) \right]$

$\pi^2$

... 

Total loss so far: $\mathcal{L}_0(\pi^0) + \mathcal{L}_1(\pi^1) + \mathcal{L}_2(\pi^2) + ...$

Interactive setting
DAgger Analysis: A reduction to no-regret online learning

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

$$\text{Avg-Regret}_T = \frac{1}{T} \left[ \sum_{t=0}^{T-1} \mathcal{L}_t(\pi^t) - \min_{\pi \in \Pi} \sum_{t=0}^{T-1} \mathcal{L}_t(\pi) \right] \leq O\left(\frac{1}{\sqrt{T}}\right)$$

Goal: Turn $\epsilon_{\text{reg}}$ to $\sqrt{T} - V^{\pi^*}$
DAgger Analysis: A reduction to no-regret online learning

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

$$
\text{Avg-Regret}_T = \frac{1}{T} \left[ \sum_{t=0}^{T-1} \mathcal{L}_t(\pi^t) - \min_{\pi \in \Pi} \sum_{t=0}^{T-1} \mathcal{L}_t(\pi) \right] \leq O\left( \frac{1}{\sqrt{T}} \right)
$$

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

$$
\min_{\pi \in \Pi} \sum_{t=0}^{T-1} \mathcal{L}_t(\pi) \leq \sum_{t=0}^{T-1} \mathcal{L}_t(\pi^*) = 0
$$

$$
\mathcal{L}_t(\pi^*) = \mathbb{E}_{s \sim d_\pi^*} \left[ c(\pi^*, s, \pi^*(s)) \right] = 0
$$
DAgger Analysis: A reduction to no-regret online learning

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

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

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

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

Which implies that:

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

\[
\ell_t(\pi^t) \leq \epsilon_{\text{reg}}
\]
DAgger Analysis: A reduction to no-regret online learning

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

$$\mathcal{L}_t (\pi^t) \leq \varepsilon_{\text{reg}}$$

Recall the definition of $\mathcal{L}_i (\pi^t)$

$$\mathcal{L}_t (\pi^t) = \mathbb{E}_{s \sim d_{\mu}^{\pi^t}} \left[ c \left( \pi^t, s, \pi^* (s) \right) \right] \leq \varepsilon_{\text{reg}}$$
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:

\[
\mathcal{L}_t (\pi^t) \leq \epsilon_{\text{reg}}
\]

Recall the definition of \( \mathcal{L}_t (\pi^t) \)

\[
\mathcal{L}_t (\pi^t) = \mathbb{E}_{s \sim d_{\mu}^{\pi^t}} \left[ c \left( \pi^t, s, \pi^*(s) \right) \right] \leq \epsilon_{\text{reg}}
\]

\( \pi^t \) matches to \( \pi^* \) 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:

$$\mathcal{L}_t (\pi^t) \leq \epsilon_{\text{reg}}$$

Recall the definition of $\mathcal{L}_t (\pi^t)$

$$\mathcal{L}_t (\pi^t) = \mathbb{E}_{s \sim d^t_\mu} \left[ c \left( \hat{\pi}, s, \pi^*(s) \right) \right] \leq \epsilon_{\text{reg}}$$

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

Recall BC, we had:

$$\mathbb{E}_{s \sim d^{\pi^*}} \left[ c(\hat{\pi}, s, \pi^*(s)) \right] \leq \epsilon, \text{ i.e., we matched to } \pi^* \text{ under } \pi^*\text{'s distribution}$$
Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration $t$, such that:

$$V^{\pi^*} - V^{\pi^t} \leq \frac{\max_{s,a} A^{\pi^*}(s, a)}{1 - \gamma} \cdot \epsilon_{\text{reg}}$$

This bound indicates that:
Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration $t$, such that:

$$V_{\pi^*} - V_{\pi'} \leq \frac{\max_{s,a} \left| A_{\pi^*}(s, a) \right|}{1 - \gamma} \cdot \epsilon_{\text{reg}}$$

This bound indicates that:

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

$$\max | A_{\pi^*}(s, a) | \leq c \in \mathbb{R}^+$$

Small number

$$| (\cdot) - (\cdot) | \text{ is small}$$
Finally, turn things into the performance bound using PDL:

**Theorem:** There exists a iteration $t$, such that:

$$V^\pi - V^{\pi_t} \leq \frac{\max_{s,a} |A^\pi(s, a)|}{1 - \gamma} \cdot \varepsilon_{reg} \leq \frac{c}{1 - \gamma} \varepsilon_{reg}$$

This bound indicates that:

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

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

i.e., at $s$, taking $a$ then following $\pi^*$ is almost as good as following $\pi^*$ directly.
Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration $t$, such that:

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

This bound indicates that:

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

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

i.e., at $s$, taking $a$ then following $\pi^*$ is almost as good as following $\pi^*$ directly
Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration $t$, such that:

$$V_{\pi^*} - V_{\pi^t} \leq \max_{s,a} \left| A_{\pi^*}(s, a) \right| \frac{1 - \gamma}{1 - \gamma} \cdot \epsilon_{\text{reg}}$$

This bound indicates that:

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

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

i.e., at $s$, taking $a$ then following $\pi^*$ is almost as good as following $\pi^*$ directly.
Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration $t$, such that:

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

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

This bound indicates that:

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

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

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

\[
\text{from Dagger: } \mathbb{E}_t(\pi^t) \leq \epsilon_{\text{reg}} \\
\Rightarrow \mathbb{E}_{s \sim d^\pi} 1\{\pi^t(s) \neq \pi^*(s)\} \leq \epsilon_{\text{reg}}
\]
Finally, turn things into the performance bound using PDL:

Theorem: There exists a iteration $t$, such that:

$$V_{\pi^*} - V_{\pi^t} \leq \frac{\max_{s,a} |A_{\pi^*}(s,a)|}{1 - \gamma} \cdot \epsilon_{\text{reg}}$$

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

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

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

$$V_{\pi^*} - V_{\pi^t} \leq \frac{\max_{s,a} |A_{\pi^*}(s,a)|}{1 - \gamma} \cdot \epsilon_{\text{reg}}$$

This bound indicates that:

We avoid quadratic error if expert $\pi^*$ can quickly recover from a mistake, i.e., at $s$, taking $a$ then following $\pi^*$ is almost as good as following $\pi^*$ directly.
Summary of DAgger
Summary of DAgger

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

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

$\mu$-Regret

Argumed

CfTRL
Summary of DAgger

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

\[ \mathbb{E}_{s \sim d_{\mu}^{\hat{\pi}}} \left[ 1 \{ \hat{\pi}(s) \neq \pi^*(s) \} \right] \leq \epsilon_{\text{reg}} = O\left(\frac{1}{\sqrt{T}}\right) \]

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

\[ V^{\pi^*} - V^{\pi^i} \leq O \left( \frac{1}{1 - \gamma} \cdot \epsilon_{\text{reg}} \right) \]

\[ \mathbb{E}_{S \sim d_{\mu}^{\hat{\pi}}} \left[ 1 \{ \hat{\pi}(s, \pi^*(s)) \} \right] = 0 \]

\[ \mathbb{E}_{S \sim d_{\mu}^{\pi^i}} \left[ A^\hat{\pi}(s, \pi^*(s)) \right] \]
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 \( \mathcal{D} = (s_i^*, a_i^* )_{i=1}^M \sim \mathcal{d}^{\pi^*} \)

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 \[ \mathcal{D} = (s_i^*, a_i^*)_i^{M} \sim d^{\pi^*} \]

No expert interaction, no real world interaction

2. Interactive IL setting:

We have access to \( \pi^* \) 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 $\mathcal{D} = (s_i^*, a_i^*)_{i=1}^M \sim d^{\pi^*}$ (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 $\mathcal{M} = \{S, A, H, c, P, \mu, \pi^*\}$
Setting

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

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

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

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

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

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

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

We have a dataset $\mathcal{D} = (s_{i*}, a_{i*})_{i=1}^M \sim d^{\pi^*}$

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

Key Assumption on cost:

\[ c(s, a) = \langle \theta^*, \phi(s, a) \rangle, \text{linear wrt feature } \phi(s, a) \]

**Fig. 4.** Classifier feature response maps. Top left is the original image.
Running Example: Define feature map

Key Assumption on cost:
\[ c(s, a) = \langle \theta^*, \phi(s, a) \rangle, \text{linear wrt feature } \phi(s, a) \]

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:
\[ c(s, a) = \langle \theta^*, \phi(s, a) \rangle, \text{ linear wrt feature } \phi(s, a) \]

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

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

Fig. 4. Classifier feature response maps. Top left is the original image.
Running Example: Define feature map

Key Assumption on cost:
\[ c(s, a) = \langle \theta^*, \phi(s, a) \rangle, \text{linear wrt feature } \phi(s, a) \]

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

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

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

Maybe colliding with cars or buildings has **high** cost, but walking on sidewalk 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