Introduction to Imitation Learning & the Behavior Cloning Algorithm

Annoucements

1. We had a typo in 2.2 of the homework, fixed and updated pdf/latex are posted on ED

2. Releasing the next reading quiz on DPO

3. No class this Wednesday and no office hour this Thursday — traveling to DC for DoD meetings

Recap

Infinite horizon Discounted MDPs

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

Average state distribution: $d^{\pi} = (1-\gamma)\sum_{h=0}^{\infty}\gamma^{h}\mathbb{P}_{h}^{\pi}$

Recap

Infinite horizon Discounted MDPs

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

What if r is unknown

Average state distribution: $d^{\pi} = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h \mathbb{P}_h^{\pi}$

Recap

Infinite horizon Discounted MDPs

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

What if r is unknown

We have covered how to learn a reward from binary preference data…

Recap

Infinite horizon Discounted MDPs

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

What if r is unknown

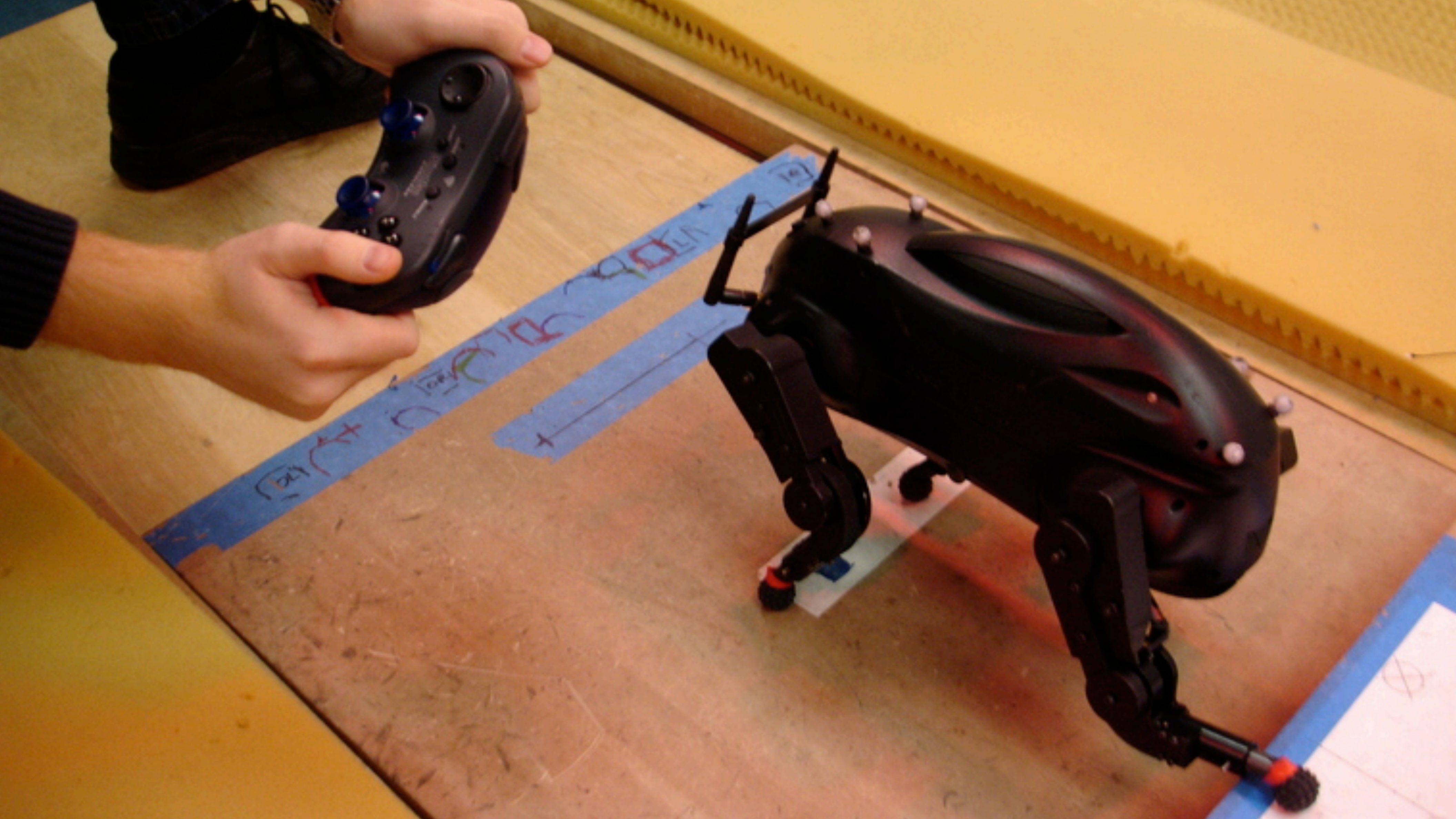
We have covered how to learn a reward from binary preference data…

Today: how to learn directly from expert demonstations

Outline for today:

1. Offline Imitation Learning: Behavior Cloning

2. Performance difference lemma and its application to proving BC's bound



An Autonomous Land Vehicle In A Neural Network *[Pomerleau, NIPS ‘88]*

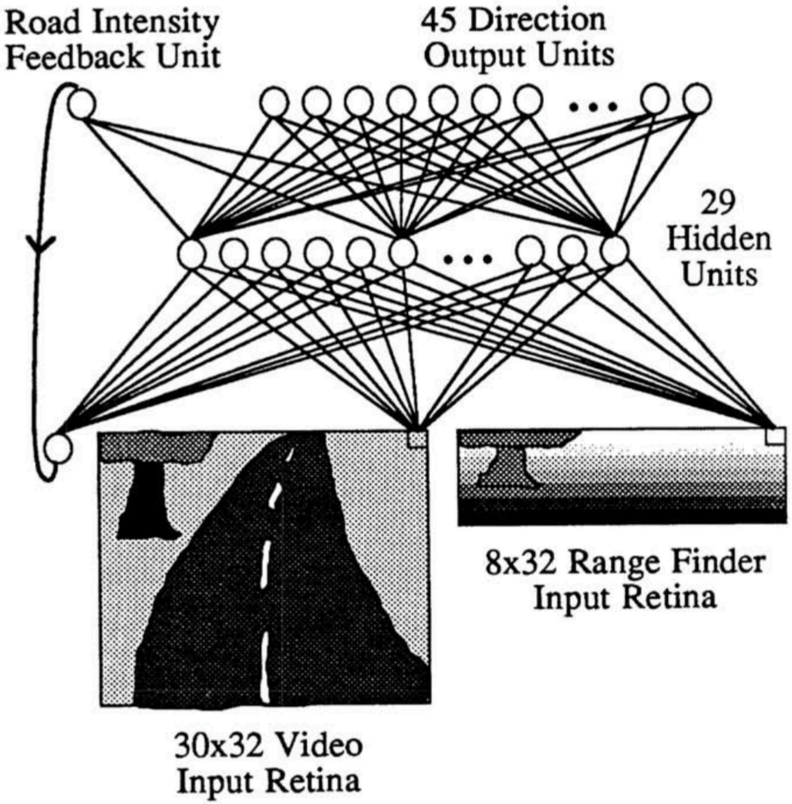
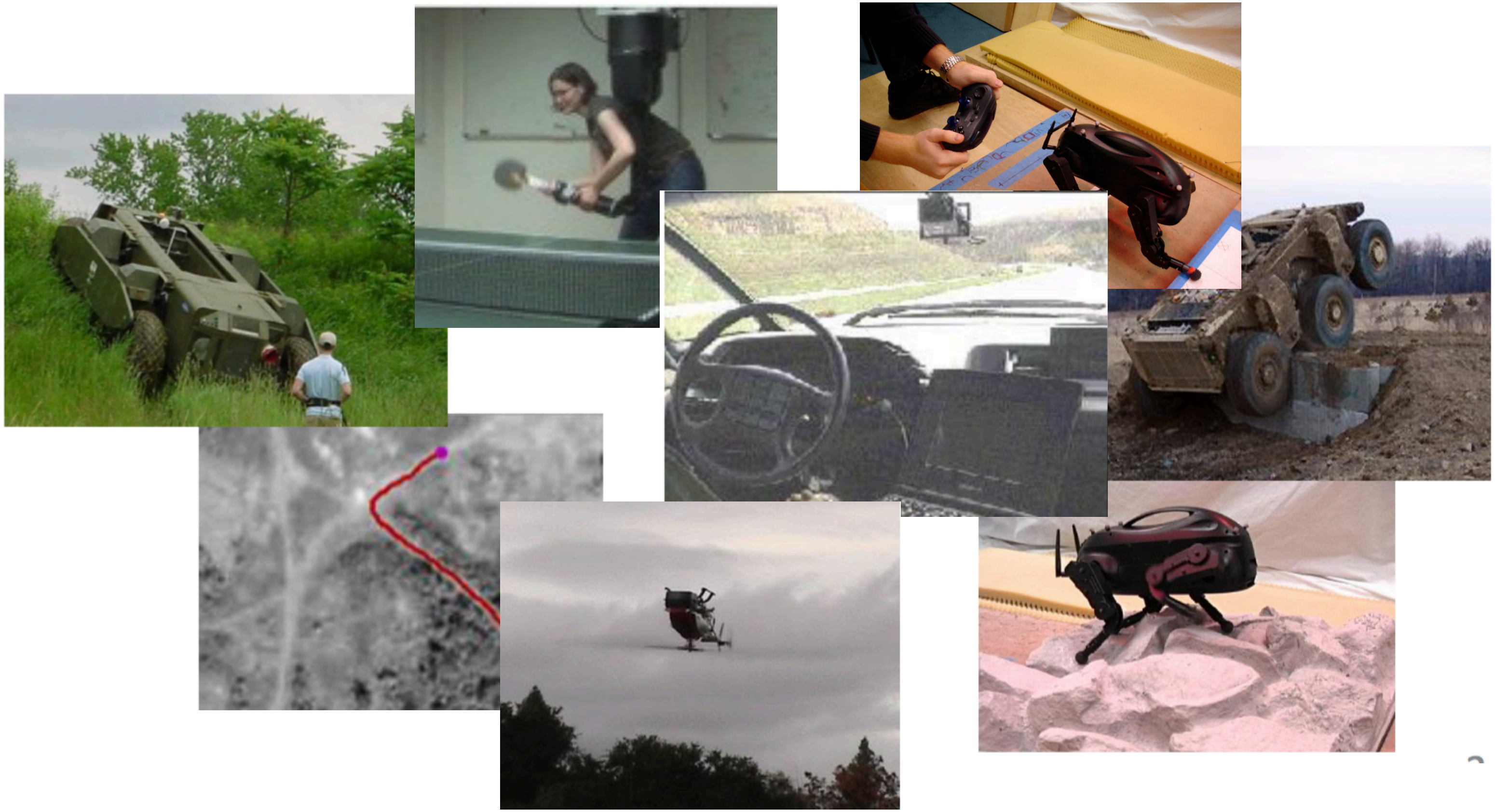


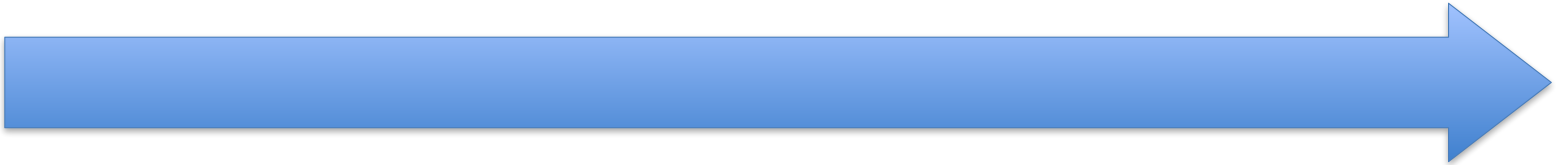
Figure 1: ALVINN Architecture

Imitation Learning

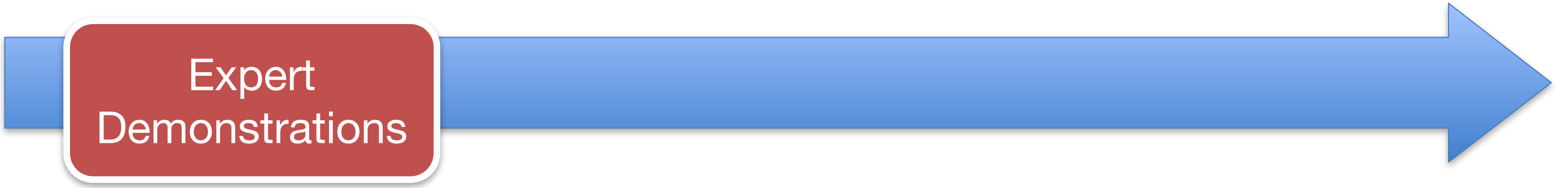


Imitation Learning

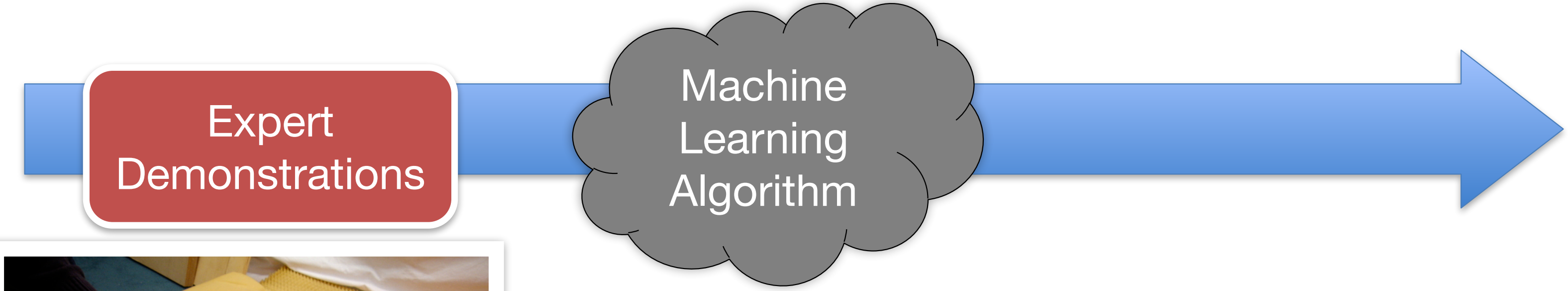
Imitation Learning

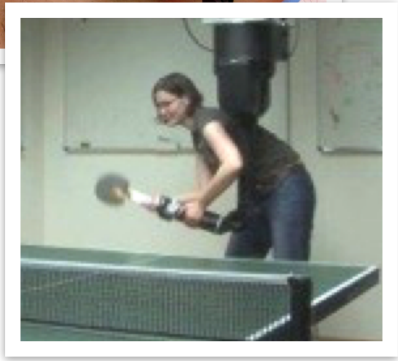


Imitation Learning



Imitation Learning

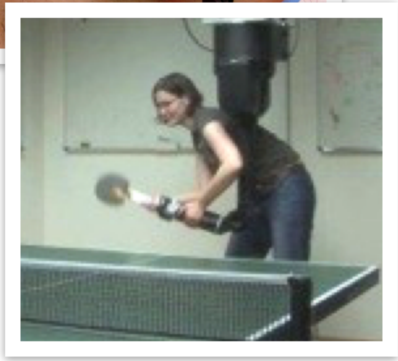




- SVM
- Gaussian Process
- Kernel Estimator
- Deep Networks
- Random Forests
- LWR
- …

Imitation Learning

Expert Demonstrations → Machine Learning Algorithm → Policy π



- SVM
- Gaussian Process
- Kernel Estimator
- Deep Networks
- Random Forests
- LWR
- …

Maps *states* to <u>actions</u>

Learning to Drive by Imitation

[Pomerleau89, Saxena05, Ross11a]

Input:

Camera Image

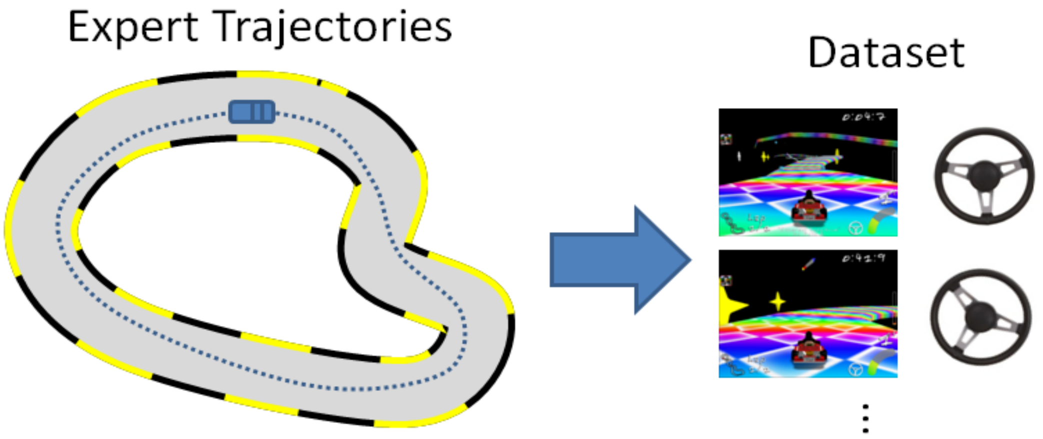
Policy

Output:

Steering Angle
in [-1, 1]

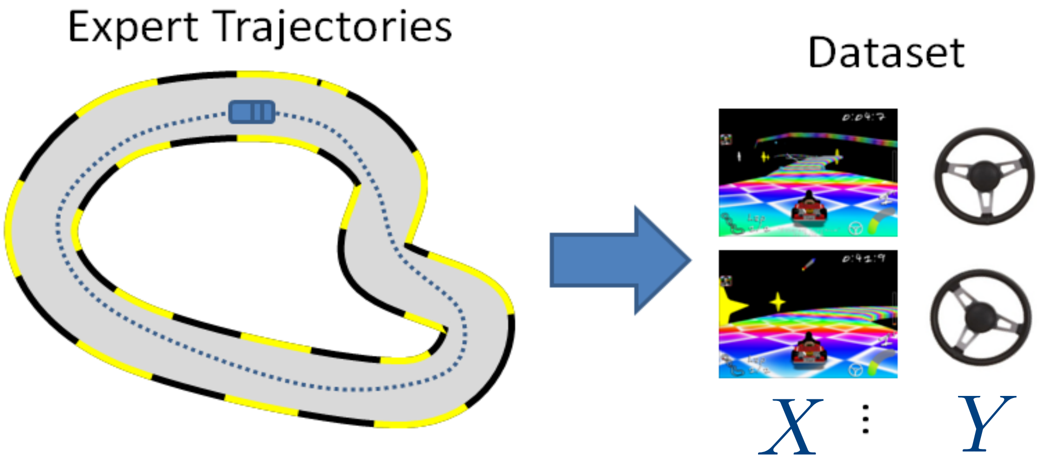
[Widrow64,Pomerleau89]

Supervised Learning Approach: Behavior Cloning

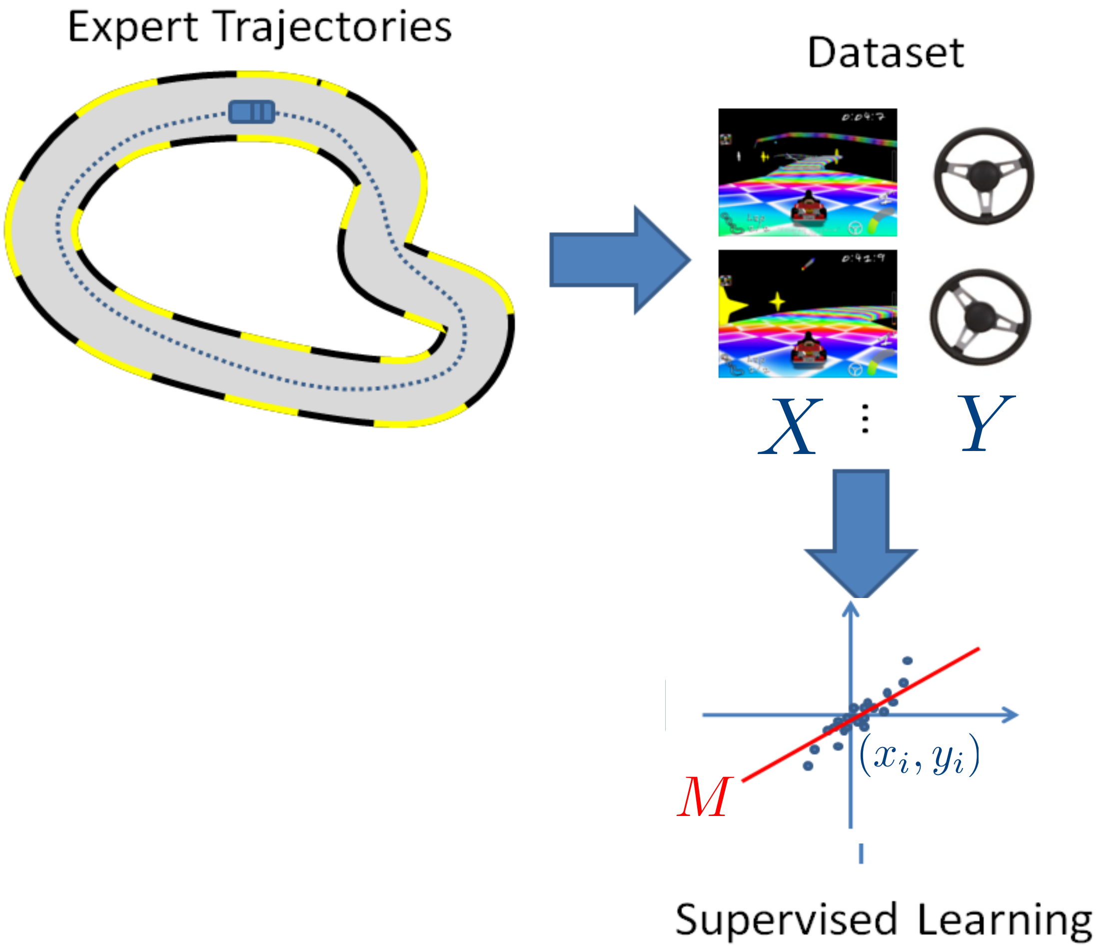


[Widrow64,Pomerleau89]

Supervised Learning Approach: Behavior Cloning

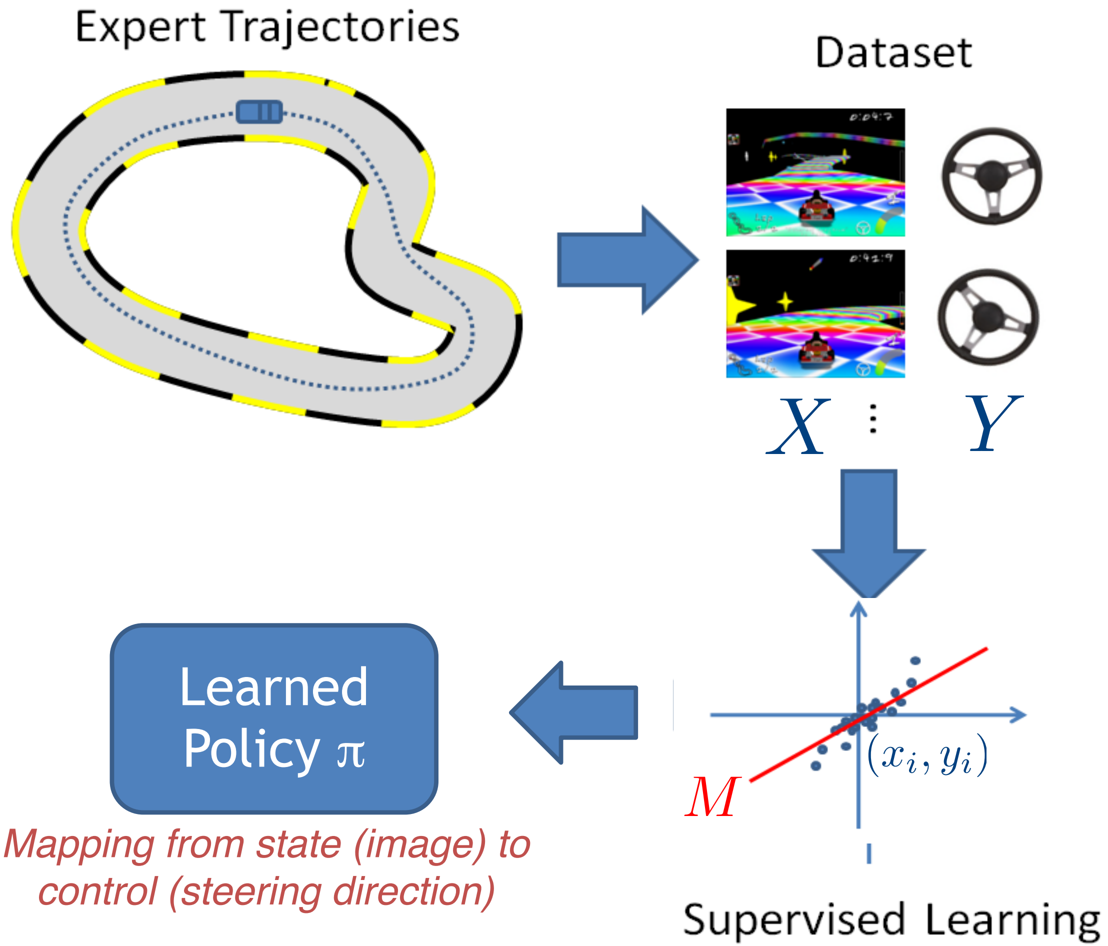


Supervised Learning Approach: Behavior Cloning



[Widrow64,Pomerleau89]

Supervised Learning Approach: Behavior Cloning




Ready!


Ready!


Ready!

LLMs are trained via BC in their pre-training phase

Take a sentence from the web:

State

Action

Reinforcement learning (RL) is an interdisciplinary area of machine learning and optimal control concerned with how an intelligent agent should take actions in a dynamic environment in order to maximize a reward signal.

LLMs are trained via BC in their pre-training phase

Take a sentence from the web:

State Action

Reinforcement learning (RL) is an interdisciplinary area of machine learning and optimal control concerned with how an intelligent agent should take actions in a dynamic environment in order to maximize a reward signal.

LLMs are trained via BC in their pre-training phase

Take a sentence from the web:

State — Action

Reinforcement learning (RL) is an interdisciplinary area of **machine** *learning and optimal control concerned with how an intelligent agent should take actions in a dynamic environment in order to maximize a reward signal.*

Forcing LLM to predict the next "action" conditioned on past…

Let's formalize the offline IL Setting and the Behavior Cloning algorithm

Discounted infinite horizon MDP $\mathscr{M} = \{S, A, \gamma, r, P, \rho, \pi^\star\}$

Let’s formalize the offline IL Setting and the Behavior Cloning algorithm

Discounted infinite horizon MDP $\mathcal{M} = \{S, A, \gamma, r, P, \rho, \pi^\star\}$

Ground truth reward $r(s, a) \in [0,1]$ is unknown;

For simplicity, let’s assume expert is a (nearly) optimal policy $\pi^\star$

Let's formalize the offline IL Setting and the Behavior Cloning algorithm

Discounted infinite horizon MDP $\mathscr{M} = \{S, A, \gamma, r, P, \rho, \pi^\star\}$

Ground truth reward $r(s, a) \in [0,1]$ is unknown;

For simplicity, let's assume expert is a (nearly) optimal policy $\pi^\star$

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

Let’s formalize the offline IL Setting and the Behavior Cloning algorithm

Discounted infinite horizon MDP $\mathscr{M} = \{S, A, \gamma, r, P, \rho, \pi^\star\}$

Ground truth reward $r(s, a) \in [0,1]$ is unknown;

For simplicity, let’s assume expert is a (nearly) optimal policy $\pi^\star$

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

Goal: learn a policy from $\mathscr{D}$ that is as good as the expert $\pi^\star$

Let's formalize the Behavior Cloning algorithm

BC Algorithm input: a restricted policy class $\Pi = \{\pi : S \mapsto \Delta(A)\}$

BC is a Reduction to Supervised Learning:

Let's formalize the Behavior Cloning algorithm

BC Algorithm input: a restricted policy class $\Pi = \{\pi : S \mapsto \Delta(A)\}$

BC is a Reduction to Supervised Learning:

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{M} \ell\left(\pi, s^\star, a^\star\right)$$

Let's formalize the Behavior Cloning algorithm

BC Algorithm input: a restricted policy class $\Pi = \{\pi : S \mapsto \Delta(A)\}$

BC is a Reduction to Supervised Learning:

$$\hat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{M} \ell\left(\pi, s^\star, a^\star\right)$$

Many choices of loss functions:

Let's formalize the Behavior Cloning algorithm

BC Algorithm input: a restricted policy class $\Pi = \{\pi : S \mapsto \Delta(A)\}$

BC is a Reduction to Supervised Learning:

$$\hat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{M} \ell\left(\pi, s^\star, a^\star\right)$$

Many choices of loss functions:

1. Negative log-likelihood (NLL): $\ell(\pi, s^\star, a^\star) = -\ln \pi(a^\star \mid s^\star)$

Let’s formalize the Behavior Cloning algorithm

BC Algorithm input: a restricted policy class $\Pi = \{\pi : S \mapsto \Delta(A)\}$

BC is a Reduction to Supervised Learning:

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{M} \ell\left(\pi, s^\star, a^\star\right)$$

Many choices of loss functions:

1. Negative log-likelihood (NLL): $\ell(\pi, s^\star, a^\star) = -\ln \pi(a^\star \mid s^\star)$

2. square loss (i.e., regression for continuous action): $\ell(\pi, s^\star, a^\star) = \|\pi(s^\star) - a^\star\|_2^2$

$$\widehat{\pi} = \arg\min_{\pi\in\Pi} \sum_{i=1}^{M} \ell\left(\pi, s^\star, a^\star\right)$$

Analysis

Assumption: we are going to assume Supervised Learning succeeded

$$\hat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{M} \ell\left(\pi, s^\star, a^\star\right)$$

Analysis

Assumption: we are going to assume Supervised Learning succeeded

$$\mathbb{E}_{s \sim d^{\pi^\star}} \mathbf{1}\left[\hat{\pi}(s) \neq \pi^\star(s)\right] \leq \epsilon \in \mathbb{R}^+$$

$$\hat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{M} \ell\left(\pi, s^\star, a^\star\right)$$

Analysis

Assumption: we are going to assume Supervised Learning succeeded

$$\mathbb{E}_{s \sim d^{\pi^\star}} \mathbf{1}\left[\hat{\pi}(s) \neq \pi^\star(s)\right] \leq \epsilon \in \mathbb{R}^+$$

Does that imply that $\hat{\pi}$ is a good policy? What's the performance difference between $\hat{\pi}$ and $\pi^\star$?

Outline for today:

1. Offline Imitation Learning: Behavior Cloning

2. Performance difference lemma and its application to proving BC's bound

Performance Difference Lemma

Given two policies $\pi : S \mapsto \Delta(A)$, $\pi' : S \mapsto \Delta(A)$, recall $V^{\pi}(s_0) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,|\, \pi\right]$

$V^{\pi} - V^{\pi'}$

Performance Difference Lemma

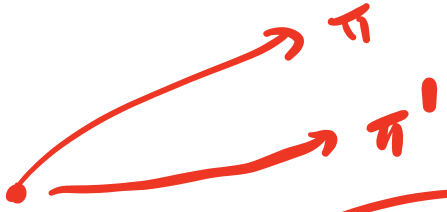
Given two policies $\pi : S \mapsto \Delta(A)$, $\pi' : S \mapsto \Delta(A)$, recall $V^{\pi}(s_0) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,|\, \pi\right]$

$s_0 \leftarrow$ Intial state

Performance Difference Lemma (PDL):

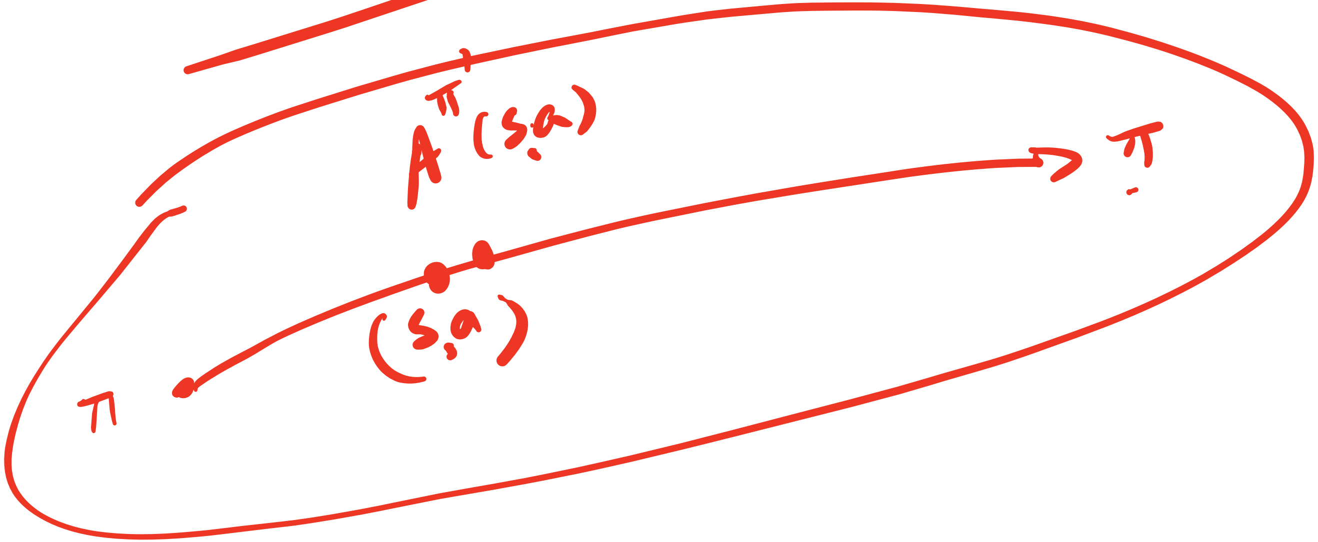
$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s \sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a \sim \pi(\cdot|s)} Q^{\pi'}(s,a) - V^{\pi'}(s)\right]$$

$$:= \frac{1}{1-\gamma}\mathbb{E}_{s \sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a \sim \pi(\cdot|s)} A^{\pi'}(s,a)\right]$$



PDL Explanation

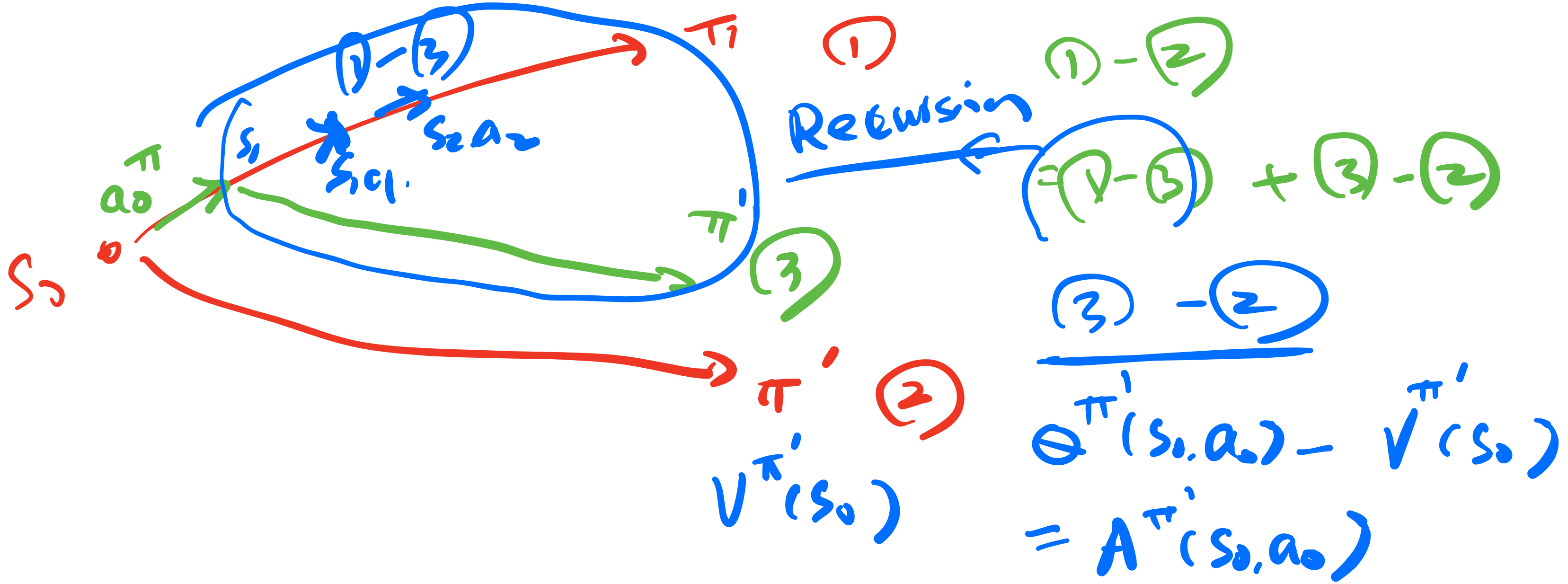
$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]$$



PDL Proof

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}_{s_0}} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} A^{\pi'}(s,a) \right]$$

Proof of Sketch (see reading material for detailed steps)



PDL Proof

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}_{s_0}} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} A^{\pi'}(s, a) \right]$$

Proof of Sketch (see reading material for detailed steps)

$$V^{\pi}(s_0) - V^{\pi'}(s_0)$$

$$= V^{\pi}(s_0) - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi'}(s') \right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi'}(s') \right] - V^{\pi'}(s_0)$$

PDL Proof

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]$$

Proof of Sketch (see reading material for detailed steps)

$$V^{\pi}(s_0) - V^{\pi'}(s_0)$$

$$= V^{\pi}(s_0) - \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)}V^{\pi'}(s')\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)}V^{\pi'}(s')\right] - V^{\pi'}(s_0)$$

$$= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)}V^{\pi'}(s')\right] - V^{\pi'}(s_0)$$

PDL Proof

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s \sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a \sim \pi(\cdot|s)} A^{\pi'}(s,a)\right]$$

Proof of Sketch (see reading material for detailed steps)

$$V^{\pi}(s_0) - V^{\pi'}(s_0)$$

$$= V^{\pi}(s_0) - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0)$$

$$= \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \mathbb{E}_{s_1 \sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0)$$

$$= \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \mathbb{E}_{s_1 \sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[Q^{\pi'}(s_0,a_0) - V^{\pi'}(s_0)\right]$$

PDL Proof

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]$$

Proof of Sketch (see reading material for detailed steps)

$$V^{\pi}(s_0) - V^{\pi'}(s_0)$$

$$= V^{\pi}(s_0) - \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)}V^{\pi'}(s')\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)}V^{\pi'}(s')\right] - V^{\pi'}(s_0)$$

$$= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)}V^{\pi'}(s')\right] - V^{\pi'}(s_0)$$

$$= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[Q^{\pi'}(s_0,a_0) - V^{\pi'}(s_0)\right]$$

$$= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[A^{\pi'}(s_0,a_0)\right]$$

An Application of PDL in Policy Iteration

Recall that $\pi^{t+1}(s) = \arg\max_a A^{\pi^t}(s,a)$

Show monotonic improvement using PDL:

$\arg\max_a Q^{\pi^t}(s,a)$

$= \arg\max_a A^{\pi^t}(s,a)$

$V^{\pi^{t+1}} - V^{\pi^t} \geq 0$

$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi^{t+1}}} \mathbb{E}_{a \sim \pi^{t+1}} A^{\pi^t}(s,a)$

π^{t+1} is deterministic

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

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

$= Q^{\pi^t}(s, \pi^t(s)) - V^{\pi^t}(s) = 0$

An Application of PDL in Policy Iteration

Recall that $\pi^{t+1}(s) = \arg\max_a A^{\pi^t}(s, a)$

Show monotonic improvement using PDL:

$$V^{\pi^{t+1}}(s_0) - V^{\pi^t}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)}A^{\pi^t}(s, a)$$

An Application of PDL in Policy Iteration

Recall that $\pi^{t+1}(s) = \arg\max_a A^{\pi^t}(s, a)$

Show monotonic improvement using PDL:

$$V^{\pi^{t+1}}(s_0) - V^{\pi^t}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi^{t+1}}} \mathbb{E}_{a \sim \pi^{t+1}(\cdot|s)} A^{\pi^t}(s, a)$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi^{t+1}}} A^{\pi^t}(s, \pi^{t+1}(s))$$

An Application of PDL in Policy Iteration

Recall that $\pi^{t+1}(s) = \arg\max_a A^{\pi^t}(s,a)$

Show monotonic improvement using PDL:

$$V^{\pi^{t+1}}(s_0) - V^{\pi^t}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)}A^{\pi^t}(s,a)$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi^{t+1}}}A^{\pi^t}(s,\pi^{t+1}(s))$$

$$\geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi^{t+1}}}A^{\pi^t}(s,\pi^t(s))$$

An Application of PDL in Policy Iteration

Recall that $\pi^{t+1}(s) = \arg\max_a A^{\pi^t}(s,a)$

Show monotonic improvement using PDL:

$$V^{\pi^{t+1}}(s_0) - V^{\pi^t}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)}A^{\pi^t}(s,a)$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi^{t+1}}}A^{\pi^t}(s,\pi^{t+1}(s))$$

$$\geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi^{t+1}}}A^{\pi^t}(s,\pi^{t}(s)) \quad = 0$$

Analysis of BC

Theorem [BC Performance] With probability at least $1-\delta$, BC returns a policy $\widehat{\pi}$:

$$V^{\pi^\star} - V^{\widehat{\pi}} \leq \frac{2}{(1-\gamma)^2}\epsilon$$

Analysis of BC

Theorem [BC Performance] With probability at least $1-\delta$, BC returns a policy $\widehat{\pi}$:

$$V^{\pi^\star} - V^{\widehat{\pi}} \le \frac{2}{(1-\gamma)^2}\epsilon$$

PDL

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

0

Analysis of BC

Theorem [BC Performance] With probability at least $1-\delta$, BC returns a policy $\hat{\pi}$:

$$V^{\pi^\star} - V^{\hat{\pi}} \le \frac{2}{(1-\gamma)^2}\epsilon$$

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

$$= \mathbb{E}_{s\sim d^{\pi^\star}} A^{\hat{\pi}}(s, \pi^\star(s)) - \mathbb{E}_{s\sim d^{\pi^\star}} A^{\hat{\pi}}(s, \hat{\pi}(s))$$

$$\underbrace{\qquad}_{a} \qquad \underbrace{\qquad}_{b} = 0$$

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

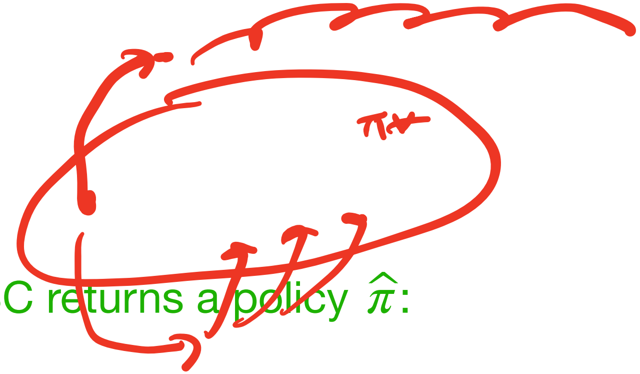
① $\pi^*(s) = \hat{\pi}(s)$

$a - b = 0$

② $\pi^*(s) \neq \hat{\pi}(s)$

$$\left| A^{\hat{\pi}}(s, \pi^*(s)) - A^{\hat{\pi}}(s, \hat{\pi}(s)) \right| \le \frac{2}{1-\gamma} \quad \left(\because A^{\pi}(s,a) \in \left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right]\right)$$

Analysis of BC



Theorem [BC Performance] With probability at least 1 − δ, BC returns a policy $\widehat{\pi}$:

$$V^{\pi^\star} - V^{\widehat{\pi}} \le \frac{2}{(1-\gamma)^2}\epsilon$$

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

$$= \mathbb{E}_{s\sim d^{\pi^\star}} A^{\widehat{\pi}}(s, \pi^\star(s)) - \mathbb{E}_{s\sim d^{\pi^\star}} A^{\widehat{\pi}}(s, \widehat{\pi}(s))$$

$$\le \mathbb{E}_{s\sim d^{\pi^\star}} \frac{2}{1-\gamma} \mathbf{1}\left\{\widehat{\pi}(s) \neq \pi^\star(s)\right\}$$

$$= \frac{2}{1-\gamma} \underbrace{\mathbb{E}_{s\sim d^{\pi^\star}} \mathbf{1}\left\{\widehat{\pi}(s) \neq \pi^\star(s)\right\}}_{\epsilon}$$

$$\left| A^{\pi}(s,a) - A^{\pi}(s,a') \right|$$

Analysis of BC

Theorem [BC Performance] With probability at least $1-\delta$, BC returns a policy $\widehat{\pi}$:

$$V^{\pi^\star} - V^{\widehat{\pi}} \leq \frac{2}{(1-\gamma)^2}\epsilon$$

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

$$= \mathbb{E}_{s\sim d^{\pi^\star}} A^{\widehat{\pi}}(s, \pi^\star(s)) - \mathbb{E}_{s\sim d^{\pi^\star}} A^{\widehat{\pi}}(s, \widehat{\pi}(s))$$

$$\leq \mathbb{E}_{s\sim d^{\pi^\star}} \frac{2}{1-\gamma}\mathbf{1}\left\{\widehat{\pi}(s) \neq \pi^\star(s)\right\}$$

$$\leq \frac{2}{1-\gamma}\epsilon$$

Analysis of BC

Theorem [BC Performance] With probability at least $1-\delta$, BC returns a policy $\hat{\pi}$:

$$V^{\pi^\star} - V^{\hat{\pi}} \leq \frac{2}{(1-\gamma)^2}\epsilon$$

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

$$= \mathbb{E}_{s\sim d^{\pi^\star}} A^{\hat{\pi}}(s, \pi^\star(s)) - \mathbb{E}_{s\sim d^{\pi^\star}} A^{\hat{\pi}}(s, \hat{\pi}(s))$$

$$\leq \mathbb{E}_{s\sim d^{\pi^\star}} \frac{2}{1-\gamma} \mathbf{1}\left\{\hat{\pi}(s) \neq \pi^\star(s)\right\}$$

$$\leq \frac{2}{1-\gamma}\epsilon$$

The quadratic amplification is annoying;
Related to pre-trained LLM hallucination;
Will see how to fix it next lecture

Summary

$\{s, a\} \sim \pi^*$

BC: simple algorithm that directly learns from human demonstrations; used in robotics and NLP

PDL: how to capture the performance difference between two policies
(as important as Simulation lemma)