Introduction to Imitation Learning & the Behavior Cloning Algorithm

Recap

Infinite horizon Discounted MDPs

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

State visitation: $d_\mu^\pi(s) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s; s_0)$

Recap

Infinite horizon Discounted MDPs

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

State visitation: $d_\mu^\pi(s) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s; s_0)$

Performance Difference Lemma:

What's the perf difference between π & π'?

Recap

Infinite horizon Discounted MDPs

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

State visitation: $d_\mu^\pi(s) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s; s_0)$

Performance Difference Lemma:

What's the perf difference between π & π'?

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

Recap

Infinite horizon Discounted MDPs

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

State visitation: $d_\mu^\pi(s) = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h \mathbb{P}_h^\pi(s; s_0)$

Performance Difference Lemma:

What's the perf difference between π & π'?

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

The adv against π' averaged over the state distribution of π

Recap

Infinite horizon Discounted MDPs

$$\mathcal{M} = \{S, A, \gamma, r, P, \mu\}$$ What if r is unknown

State visitation: $d^\pi_\mu(s) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}^\pi_h(s; s_0)$

Performance Difference Lemma:

What's the perf difference between π & π'?

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

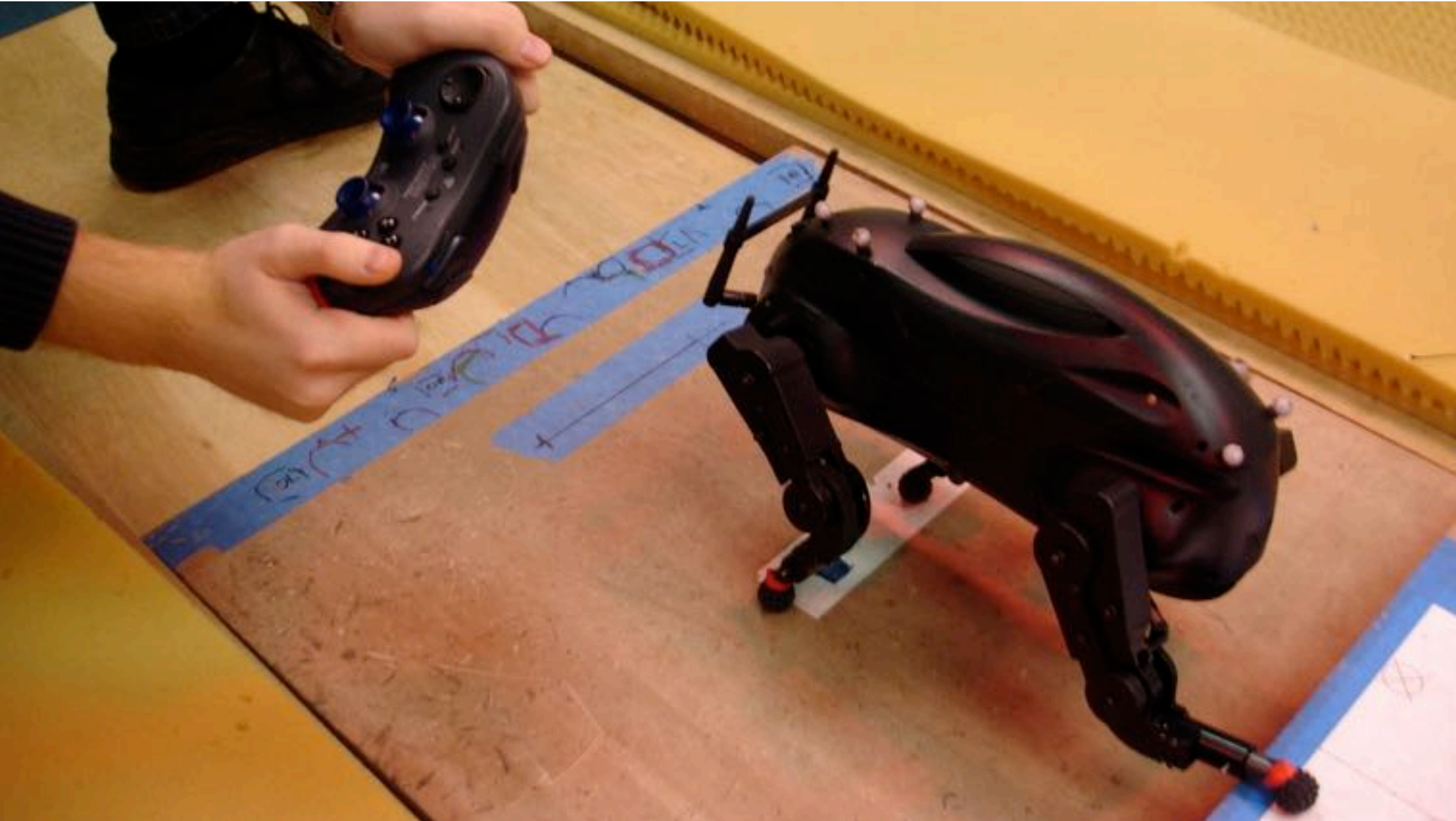
The adv against π' averaged over the state distribution of π

Outline for today:

1. Introduction of Imitation Learning

2. Offline Imitation Learning: Behavior Cloning

3. The distribution shift issue in BC



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

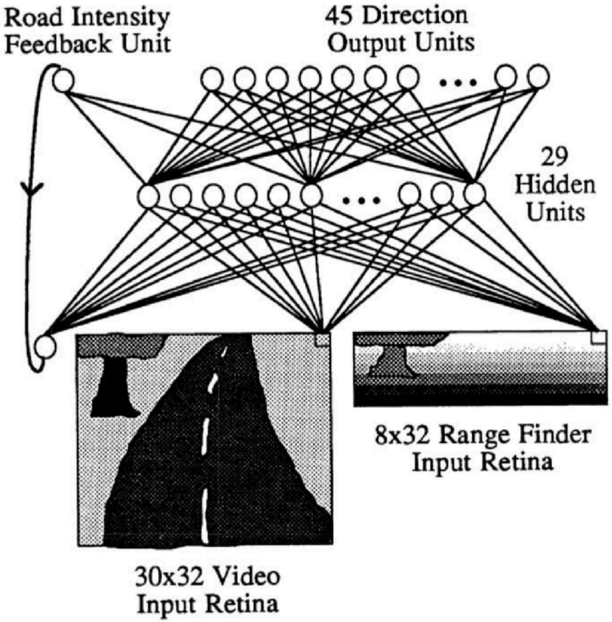
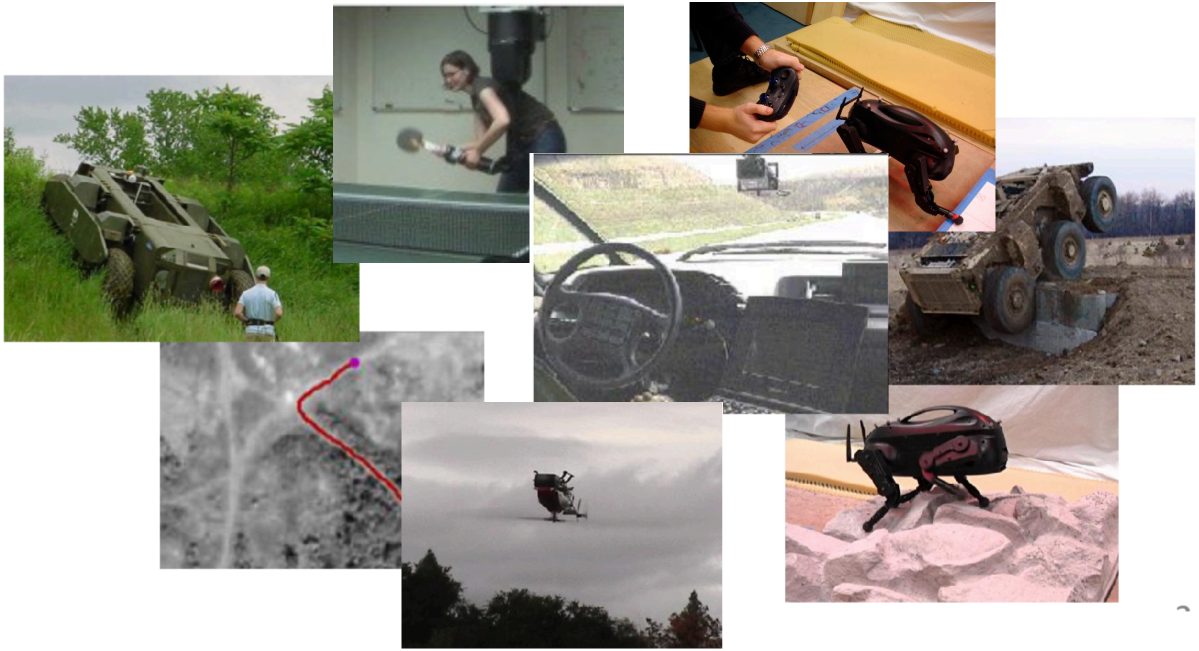


Figure 1: ALVINN Architecture

Imitation Learning

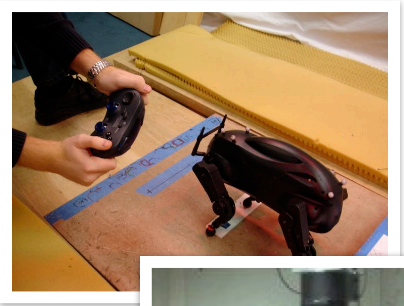


Imitation Learning

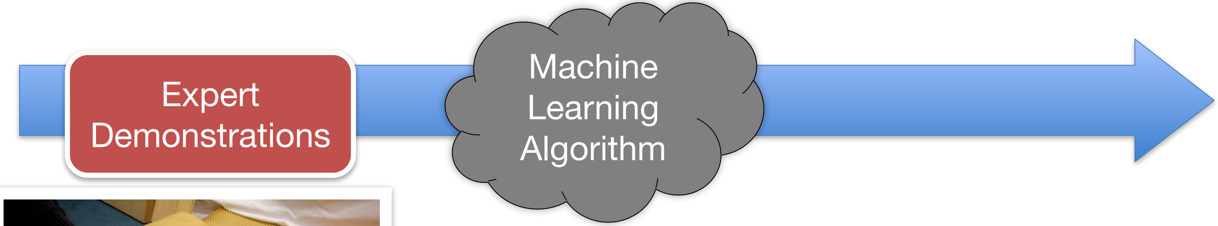
Imitation Learning

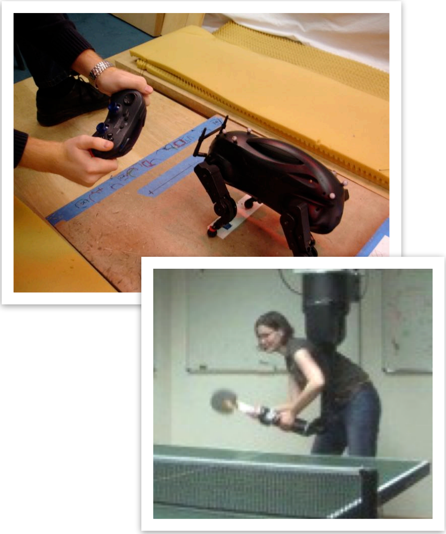
Imitation Learning

Expert Demonstrations



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

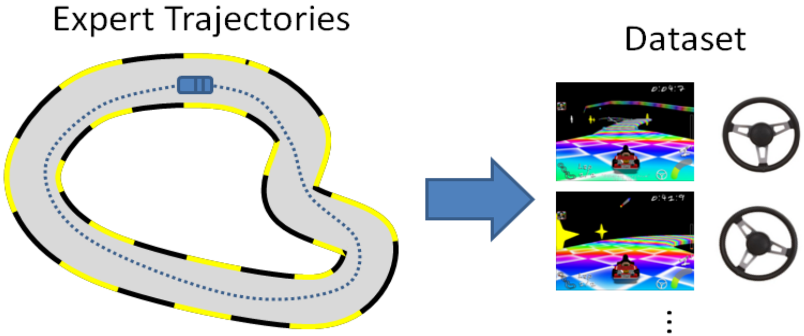
Policy

Output:

Steering Angle
in [-1, 1]

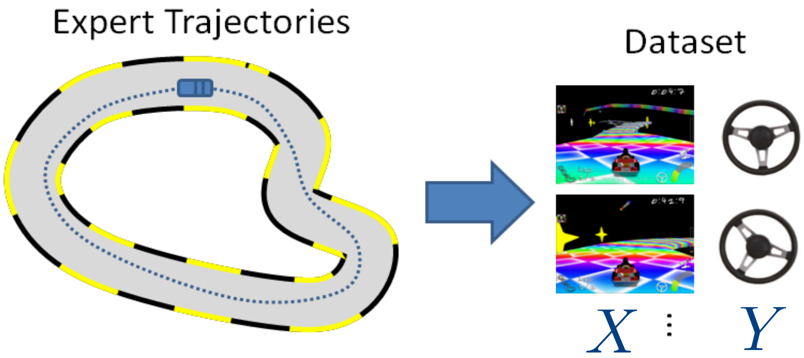
[Widrow64,Pomerleau89]

Supervised Learning Approach: Behavior Cloning



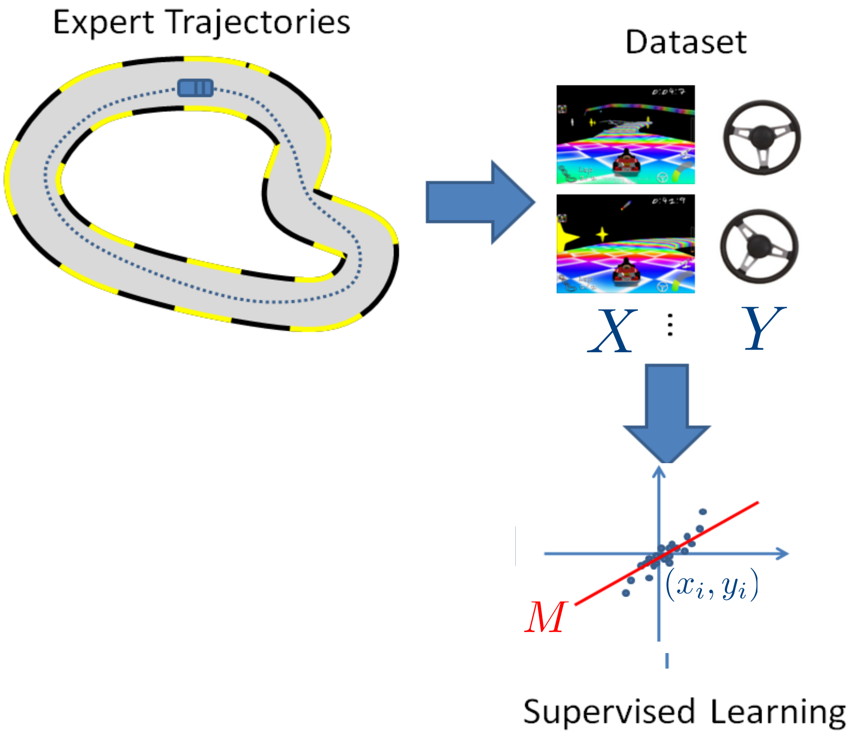
[Widrow64,Pomerleau89]

Supervised Learning Approach: Behavior Cloning



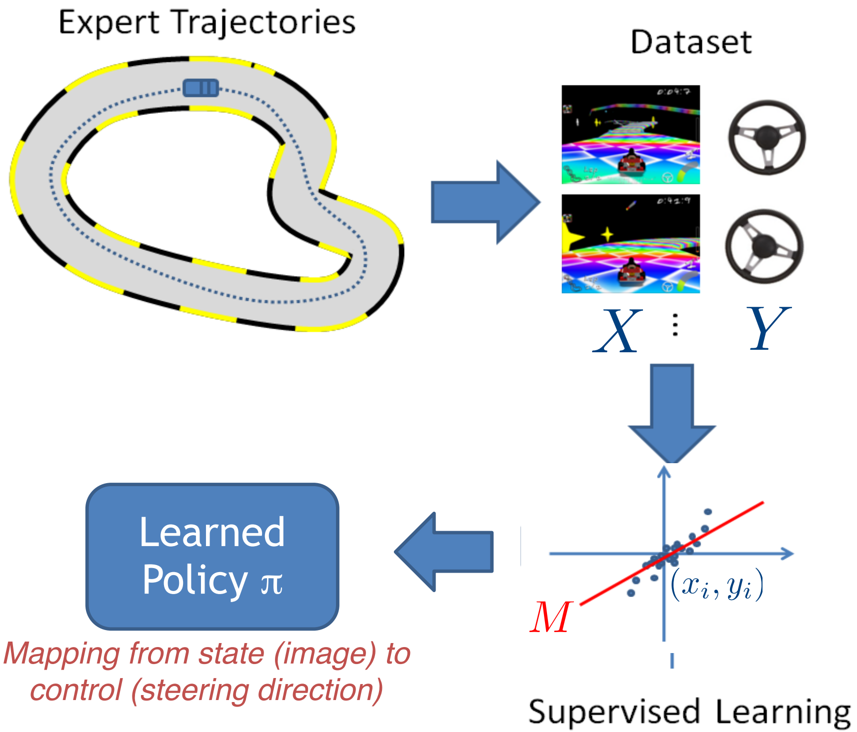
[Widrow64,Pomerleau89]

Supervised Learning Approach: Behavior Cloning



[Widrow64,Pomerleau89]

Supervised Learning Approach: Behavior Cloning




Ready!


Ready!


Ready!

Outline

1. Introduction of Imitation Learning

2. Offline Imitation Learning: Behavior Cloning

3. The distribution shift issue in BC

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:

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

loss function

$\mathcal{D} = \{s^*, a^*\}$

$s^* \sim d^{\pi^*}_{\mu}$

$a^* \sim \pi^*(s^*)$

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): $\ell(\pi, s, a^\star) = -\ln \pi(a^\star \mid s^\star)$

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:

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

Many choices of loss functions:

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

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

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

Analysis

Assumption: we are going to assume Supervised Learning succeeded

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

Analysis

Assumption: we are going to assume Supervised Learning succeeded

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

Training Distribution

$\approx \sqrt{\frac{1}{M}}$ ← # of exper samples

We care $V^{\hat{\pi}}$ versus V^{π^*}

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

Analysis

Assumption: we are going to assume Supervised Learning succeeded

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

Note that here training and testing mismatch at this stage!

Analysis

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

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

← SL-error

Analysis

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

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

$\pi^\star$, $\widehat{\pi}$ ← deterministic

↙ PDL

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

Δ

Analysis

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

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

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

$$= \mathbb{E}_{s \sim d^{\pi^\star}} A^{\hat{\pi}}(s, \pi^\star(s)) - \mathbb{E}_{s \sim d^{\pi^\star}} A^{\hat{\pi}}(s, \hat{\pi}(s)) \geq 0$$

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$

$$0 - \frac{1}{1-\gamma} \leq A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \leq \frac{1}{1-\gamma}$$

$0 \leq \cdot \leq \frac{1}{1-\gamma}$, ≥ 0

$$A^{\pi}(s,a) \in \left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right], \forall s,a$$

Analysis

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

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

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

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

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

$$= \begin{cases} 0 & \text{if } \widehat{\pi}(s) = \pi^\star(s) \\ \le \frac{2}{1-\gamma} & \text{if } \widehat{\pi}(s) \ne \pi^\star(s). \end{cases} \quad \left(\because A^{\widehat{\pi}}(s,a) \in \left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right]\right)$$

Analysis

MDP = { S, A, γ, P, r, μ } (r: unknown)

But (s, a) ~ π*

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

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

$\mathbb{E}\left[\sum_h \gamma^h r(s_h, a_h)\right]$

PDL

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

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

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

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

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

RL → SL → RL

Analysis

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

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

The quadratic amplification is annoying

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

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

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

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

Outline

1. Introduction of Imitation Learning

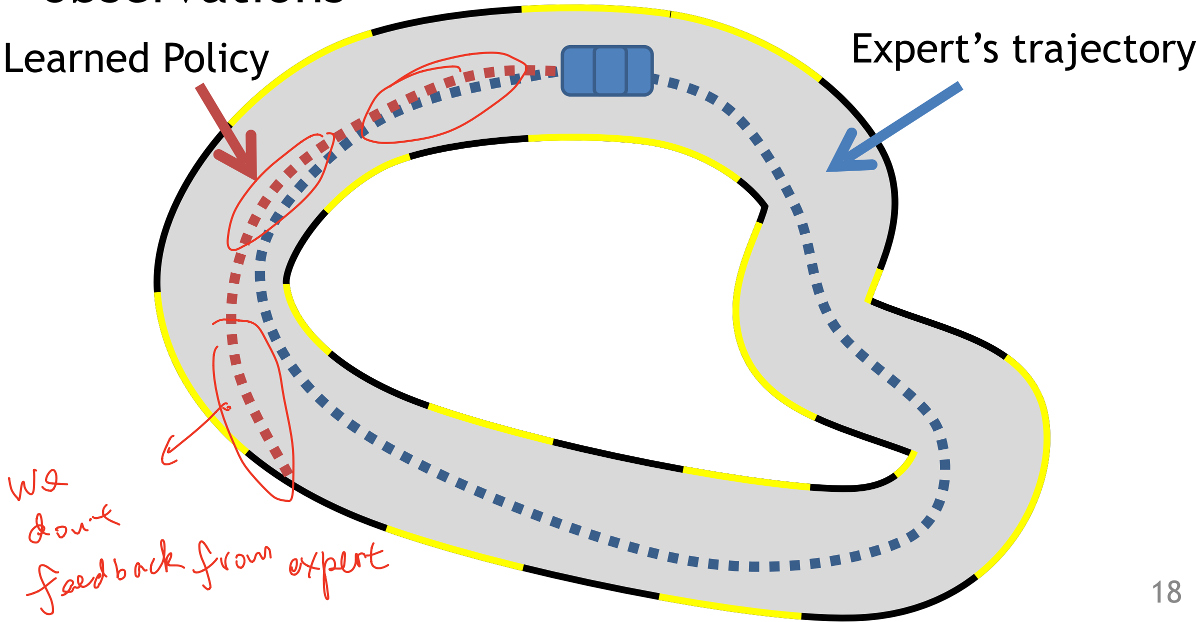
2. Offline Imitation Learning: Behavior Cloning

3. The distribution shift issue in BC

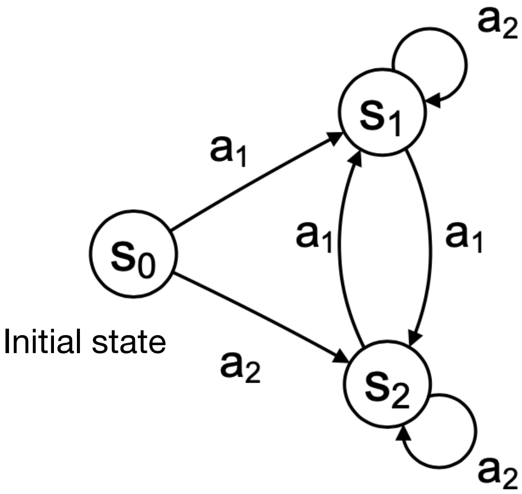
What could go wrong?

[Pomerleau89,Daume09]

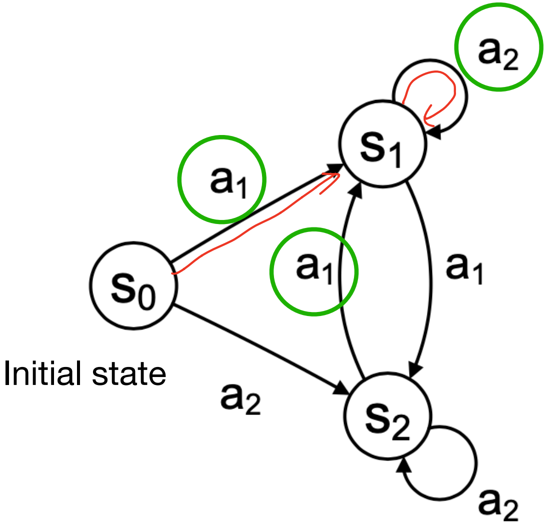
- Predictions affect future inputs/ observations



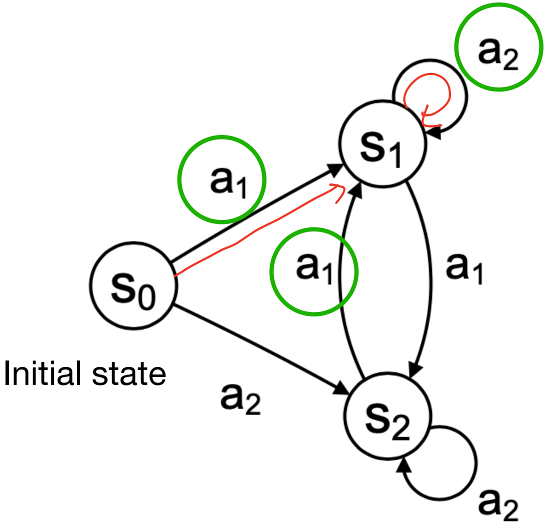
Distribution Shift: Example



Distribution Shift: Example



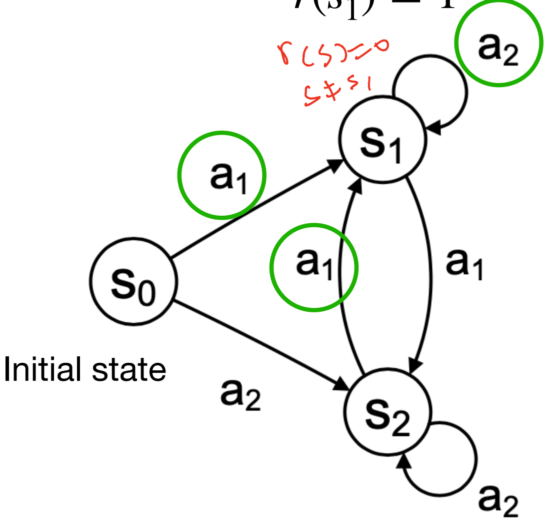
Distribution Shift: Example



$d_{s_0}^{\pi^\star}(s_0) = 1 - \gamma,\ d_{s_0}^{\pi^\star}(s_1) = \gamma,\ d_{s_0}^{\pi^\star}(s_2) = 0$

Distribution Shift: Example

$r(s_1) = 1$



$d_{s_0}^{\pi^\star}(s_0) = 1 - \gamma,\ d_{s_0}^{\pi^\star}(s_1) = \gamma,\ d_{s_0}^{\pi^\star}(s_2) = 0$

$d_{s_0}^{\pi}(s) = (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^{\pi}(s; s_0)$

s_0:

$\mathbb{P}_0^{\pi^*}(s_0) = 1$

$P_1^{\pi^*}(s_0) = 0$

$P_2^{\pi^*}(s_0) = 0$

$\vdots$

$d_{s_0}^{\pi^*}(s_0) = (1-\gamma) \sum [1 + 0 + 0 \ldots]$

$= 1 - \gamma$

s_1:

$\mathbb{P}_0^{\pi^*}(s_1) = 0$

$\mathbb{P}_1^{\pi^*}(s_1) = 1$

$\mathbb{P}_2^{\pi^*}(s_1) = 1$

$\vdots$

$d_{s_0}^{\pi^*}(s_1)$

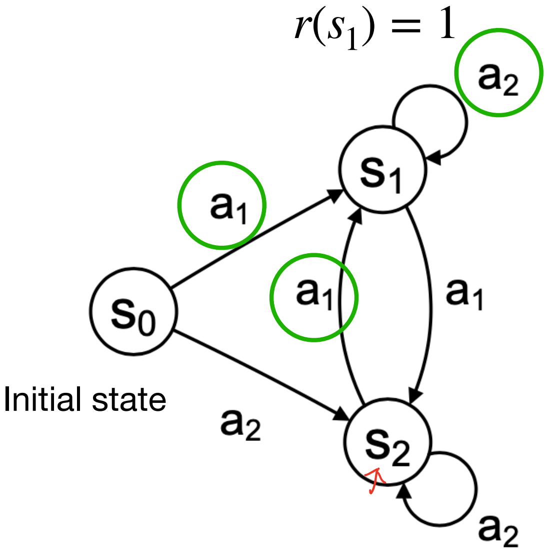
$= (1-\gamma)[0 + \gamma \cdot 1 + \gamma^2 1 + \ldots]$

$= \gamma$

s_2:

$d_{s_0}^{\pi^*}(s_2) = 0$

Distribution Shift: Example



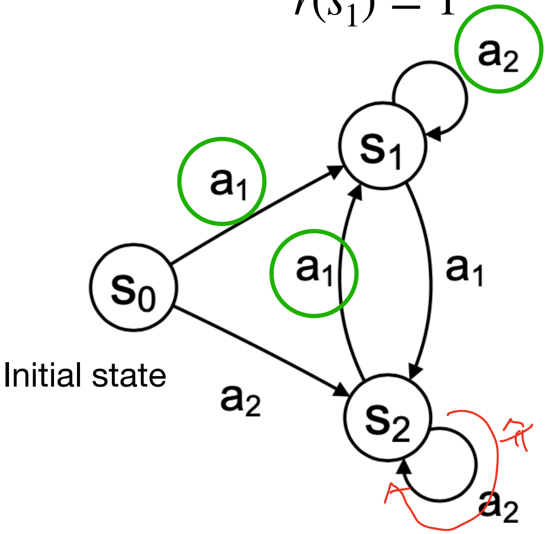
$$d_{s_0}^{\pi^\star}(s_0) = 1 - \gamma,\ d_{s_0}^{\pi^\star}(s_1) = \gamma,\ d_{s_0}^{\pi^\star}(s_2) = 0$$

$$V_{s_0}^{\pi^\star} = \frac{\gamma}{1-\gamma} = [0 + \gamma \cdot 1 + \gamma^2 \cdot 1 + \dots]$$

Distribution Shift: Example

$\epsilon \to$ small number

$r(s_1) = 1$



Assume SL returned such policy $\hat{\pi}$

$$\hat{\pi}(s_0) = \begin{cases} a_1 & \text{w/ prob } 1 - \epsilon/(1-\gamma) \\ a_2 & \text{w/ prob } \epsilon/(1-\gamma) \end{cases}, \quad \hat{\pi}(s_1) = a_2, \hat{\pi}(s_2) = a_2$$

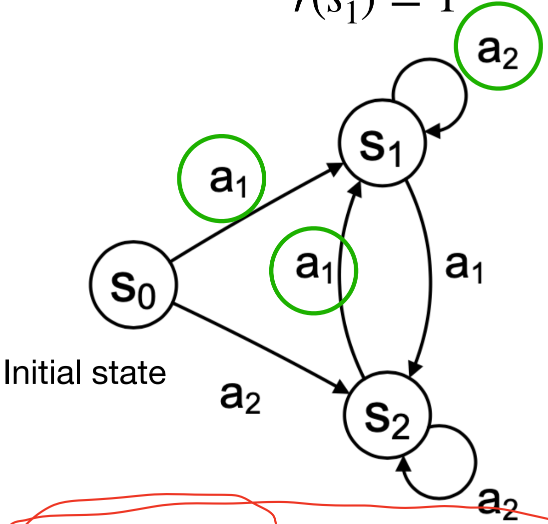
small

$$d_{s_0}^{\pi^\star}(s_0) = 1 - \gamma, d_{s_0}^{\pi^\star}(s_1) = \gamma, d_{s_0}^{\pi^\star}(s_2) = 0$$

$$V_{s_0}^{\pi^\star} = \frac{\gamma}{1-\gamma}$$

Distribution Shift: Example

$r(s_1) = 1$



$$d_{s_0}^{\pi^\star}(s_0) = 1-\gamma,\ d_{s_0}^{\pi^\star}(s_1) = \gamma,\ d_{s_0}^{\pi^\star}(s_2) = 0$$

$$V_{s_0}^{\pi^\star} = \frac{\gamma}{1-\gamma}$$

Assume SL returned such policy $\widehat{\pi}$

$$\widehat{\pi}(s_0) = \begin{cases} a_1 & \text{w/ prob } 1 - \epsilon/(1-\gamma) \\ a_2 & \text{w/ prob } \epsilon/(1-\gamma) \end{cases},\quad \widehat{\pi}(s_1) = a_2,\ \widehat{\pi}(s_2) = a_2$$

We will have good supervised learning error:

$$\mathbb{E}_{s \sim d_{s_0}^{\pi^\star}} \mathbb{E}_{a \sim \widehat{\pi}(\cdot|s)} \mathbf{1}\left(a \neq \pi^\star(s)\right) = \epsilon$$

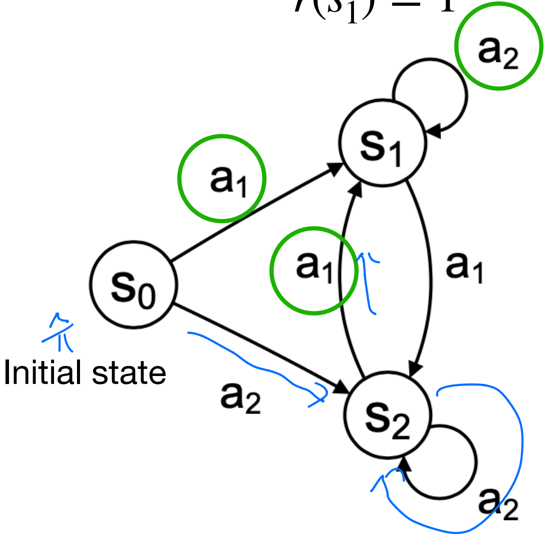
at s_0: $(1-\gamma) \cdot \frac{\epsilon}{1-\gamma} + \gamma \cdot 0 + 0 \cdot 1$

$d_{s_0}^{\pi^\star}(s_0)$ $d_{s_0}^{\pi^\star}(s_1)$ $d_{s_0}^{\pi^\star}(s_2)$

Distribution Shift: Example

$V^{\hat{\pi}}_{s_0} = \left(1 - \frac{\epsilon}{1-\gamma}\right)\frac{\gamma}{1-\gamma} + \frac{\epsilon}{1-\gamma} \cdot 0$

$r(s_1) = 1$



$d^{\pi^\star}_{s_0}(s_0) = 1 - \gamma,\ d^{\pi^\star}_{s_0}(s_1) = \gamma,\ d^{\pi^\star}_{s_0}(s_2) = 0$

$$V^{\pi^\star}_{s_0} = \frac{\gamma}{1-\gamma}$$

Assume SL returned such policy $\hat{\pi}$

$$\hat{\pi}(s_0) = \begin{cases} a_1 & \text{w/ prob } 1 - \epsilon/(1-\gamma) \\ a_2 & \text{w/ prob } \epsilon/(1-\gamma) \end{cases}, \quad \hat{\pi}(s_1) = a_2, \hat{\pi}(s_2) = a_2$$

We will have good supervised learning error:

$$\mathbb{E}_{s \sim d^{\pi^\star}_{s_0}} \mathbb{E}_{a \sim \hat{\pi}(\cdot|s)} \mathbf{1}\left(a \neq \pi^\star(s)\right) = \epsilon$$

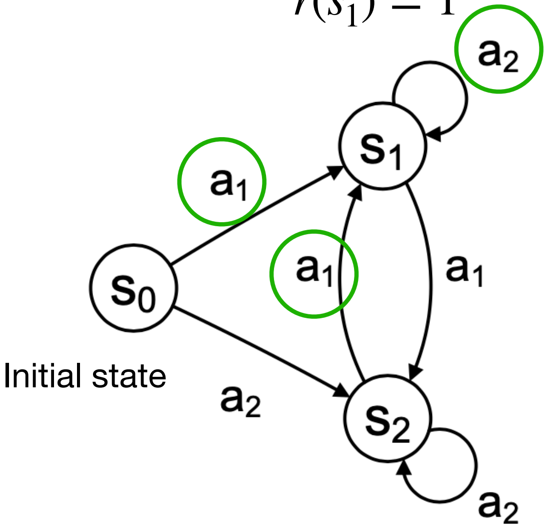
But we have quadratic error in performance:

$$V^{\hat{\pi}}_{s_0} = \frac{\gamma}{1-\gamma} - \frac{\epsilon\gamma}{(1-\gamma)^2} = V^{\pi^\star}_{s_0} - \frac{\epsilon\gamma}{(1-\gamma)^2}$$

$\frac{\gamma}{1-\gamma}$ Gap

Distribution Shift: Example

$r(s_1) = 1$



$$d_{s_0}^{\pi^\star}(s_0) = 1-\gamma,\ d_{s_0}^{\pi^\star}(s_1) = \gamma,\ d_{s_0}^{\pi^\star}(s_2) = 0$$

$$V_{s_0}^{\pi^\star} = \frac{\gamma}{1-\gamma}$$

Assume SL returned such policy $\widehat{\pi}$

$$\widehat{\pi}(s_0) = \begin{cases} a_1 & \text{w/ prob } 1 - \epsilon/(1-\gamma) \\ a_2 & \text{w/ prob } \epsilon/(1-\gamma) \end{cases}, \quad \widehat{\pi}(s_1) = a_2,\ \widehat{\pi}(s_2) = a_2$$

We will have good supervised learning error:

$$\mathbb{E}_{s\sim d_{s_0}^{\pi^\star}} \mathbb{E}_{a\sim\widehat{\pi}(\cdot|s)} \mathbf{1}\left(a \neq \pi^\star(s)\right) = \epsilon$$

But we have quadratic error in performance:

$$V_{s_0}^{\widehat{\pi}} = \frac{\gamma}{1-\gamma} - \frac{\epsilon\gamma}{(1-\gamma)^2} = V_{s_0}^{\pi^\star} - \frac{\epsilon\gamma}{(1-\gamma)^2}$$

Issue: once we make a mistake at s_0, we end up in s_2 which is not in the training data!

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

"If the network is not presented with sufficient variability in its training exemplars to cover the conditions it is likely to encounter…[it] will perform poorly"

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

"If the network is not presented with sufficient variability in its training exemplars to cover the conditions it is likely to encounter…[it] will perform poorly"

Summary for Today:

1. The most common imitation learning algorithm: BC

A reduction to supervised Learning, e.g., training classifier from $s^\star \sim d_\mu^{\pi^\star}, a^\star = \pi^\star(s^\star)$

Summary for Today:

1. The most common imitation learning algorithm: BC

A reduction to supervised Learning, e.g., training classifier from $s^\star \sim d_\mu^{\pi^\star}, a^\star = \pi^\star(s^\star)$

2. Distribution shift:

When execute the learned policy,
we may deviate from the expert trajectories,
causing compounding error

Summary for Today:

1. The most common imitation learning algorithm: BC

A reduction to supervised Learning, e.g., training classifier from $s^\star \sim d_\mu^{\pi^\star}, a^\star = \pi^\star(s^\star)$

2. Distribution shift:

When execute the learned policy,
we may deviate from the expert trajectories,
causing compounding error

3. **Again this demonstrates why RL/IL is harder than SL:**
we need to test our model on new data generated by our model