

Behavioral Cloning

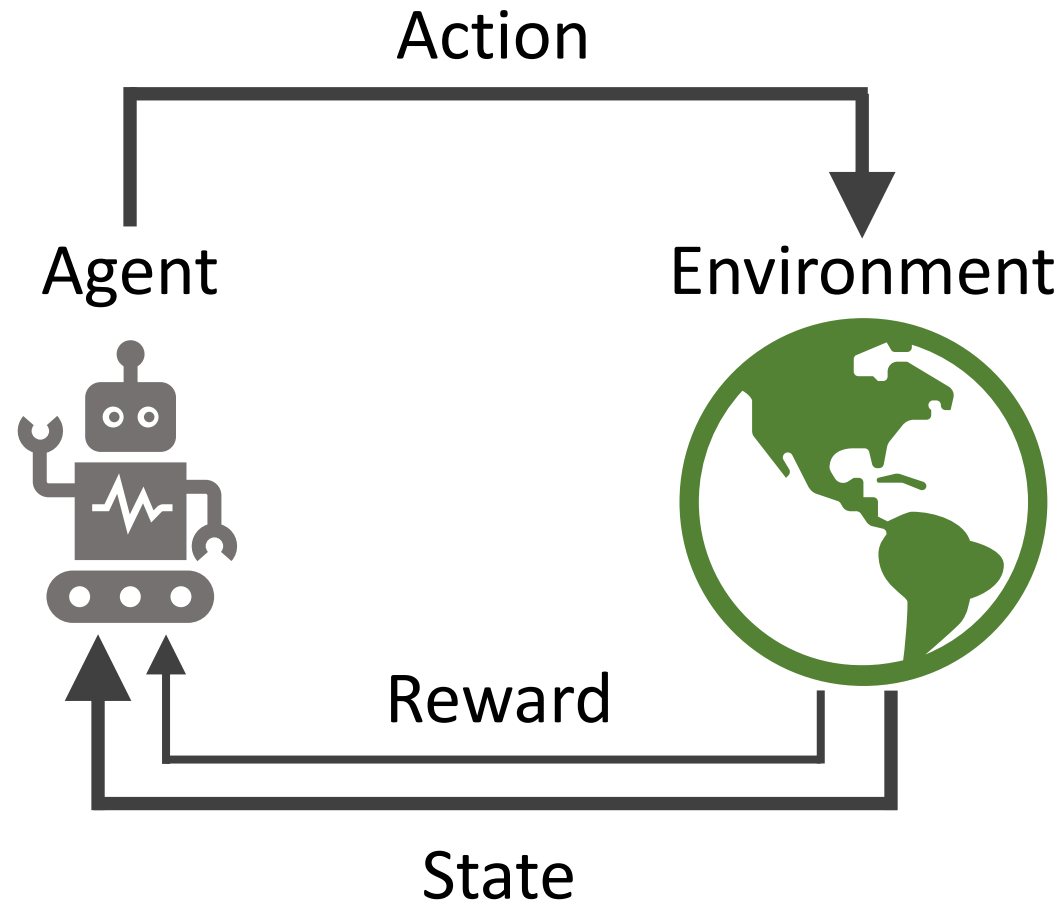
CMPT 729 G100

Jason Peng

Overview

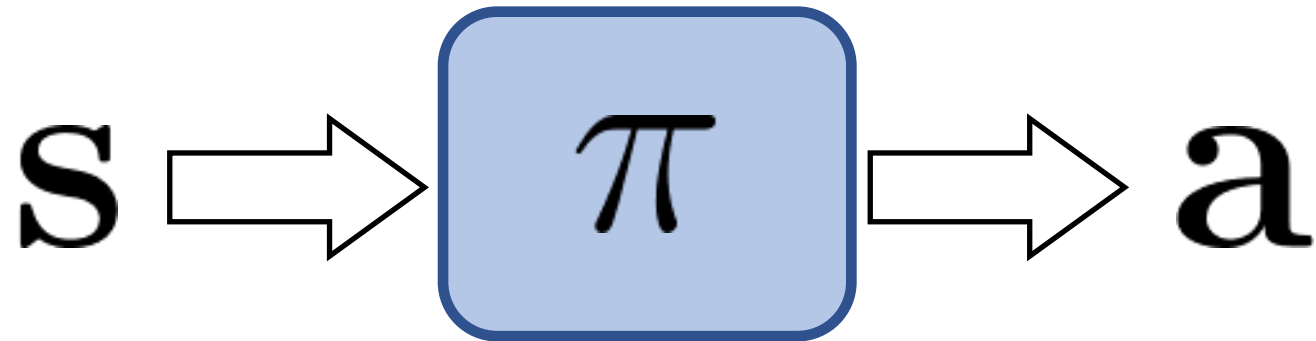
- Behavioral Cloning
- Drift
- Theoretical Analysis
- DAgger
- Applications

Agent-Environment Interface



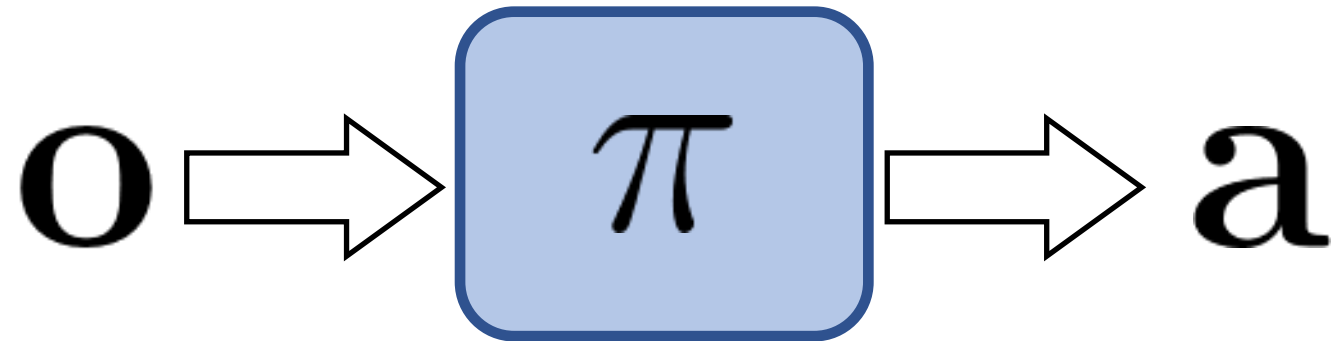
Policy

$$\pi(\mathbf{a}|\mathbf{s})$$



Policy

$$\pi(\mathbf{a}|\mathbf{o})$$

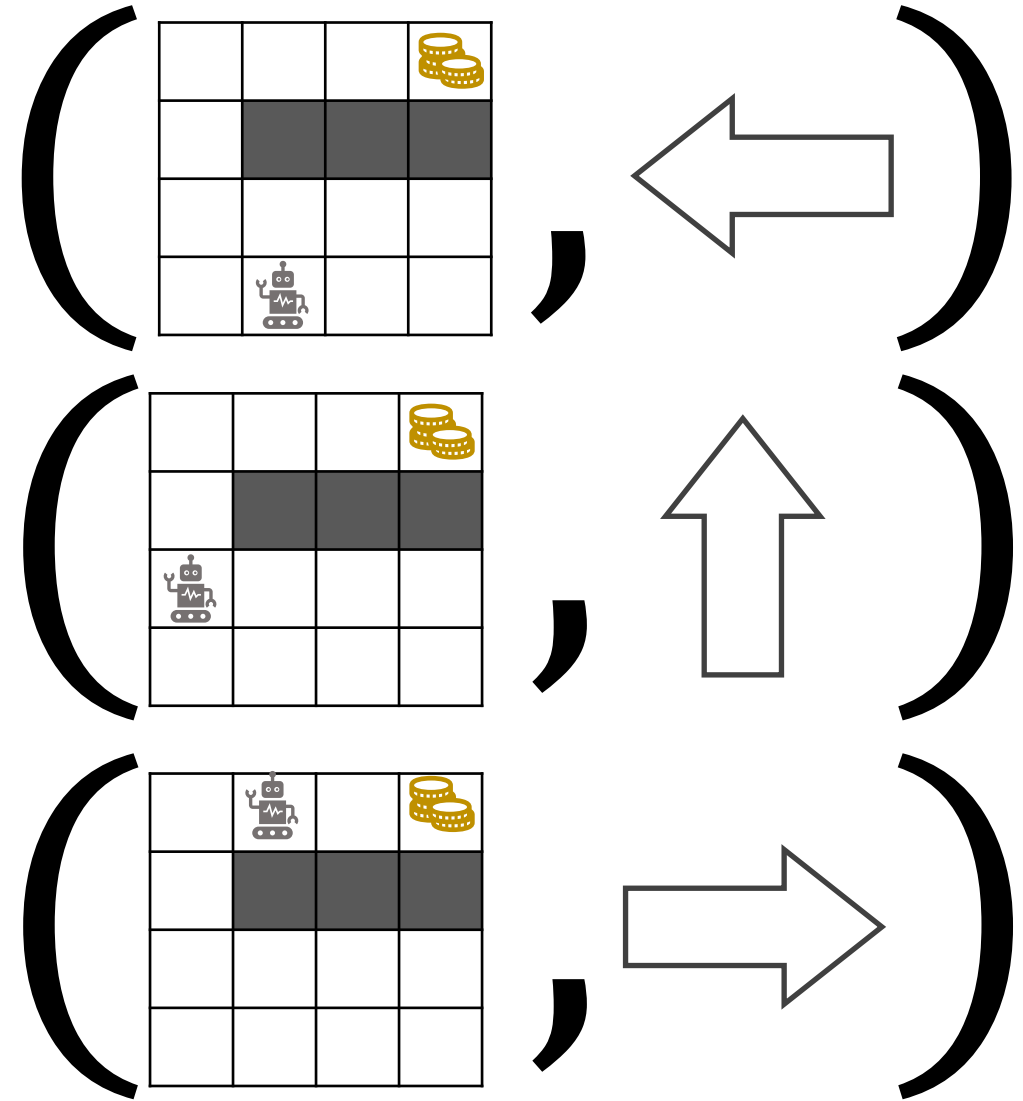


Supervised Learning

$\{(\mathbf{o}_0, \mathbf{a}_0), (\mathbf{o}_1, \mathbf{a}_1), \dots\}$



Dataset



Supervised Learning

$$\{(\mathbf{o}_0, \mathbf{a}_0), (\mathbf{o}_1, \mathbf{a}_1), \dots\}$$



Dataset



Nvidia Automotive Simulation
[NVIDIA]

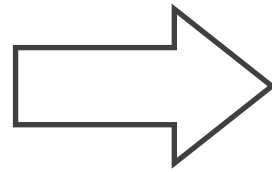


Supervised Learning

$$\{(\mathbf{o}_0, \mathbf{a}_0), (\mathbf{o}_1, \mathbf{a}_1), \dots\}$$



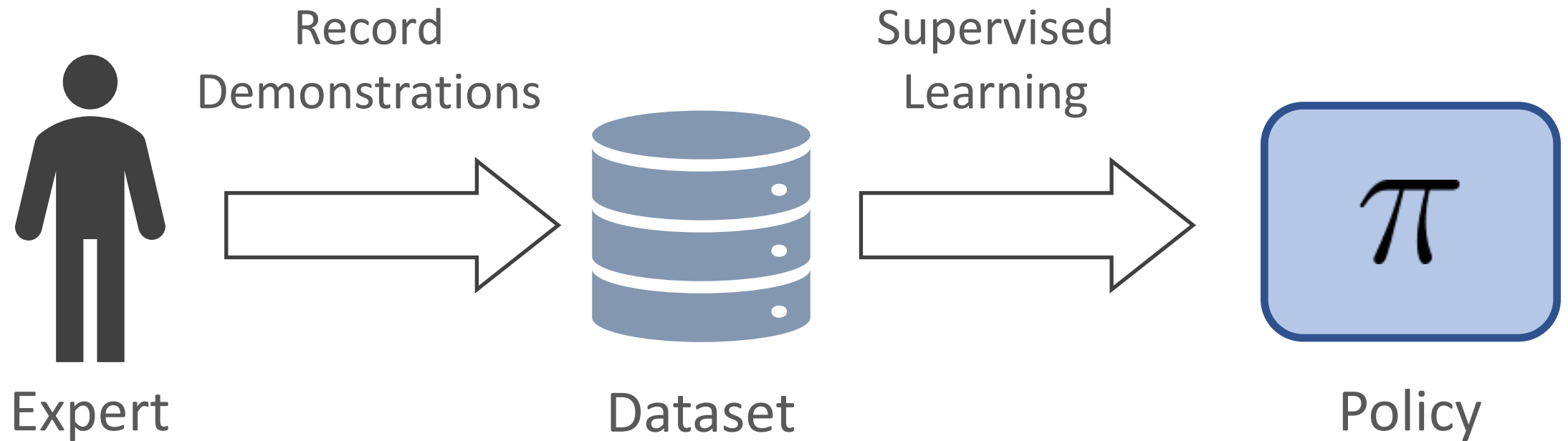
Dataset



$$\min_{\pi} \mathbb{E}_{(\mathbf{o}, \mathbf{a}) \sim \mathcal{D}} [-\log \pi(\mathbf{a} | \mathbf{o})]$$

Behavioral Cloning

Behavioral Cloning



Behavioral Cloning

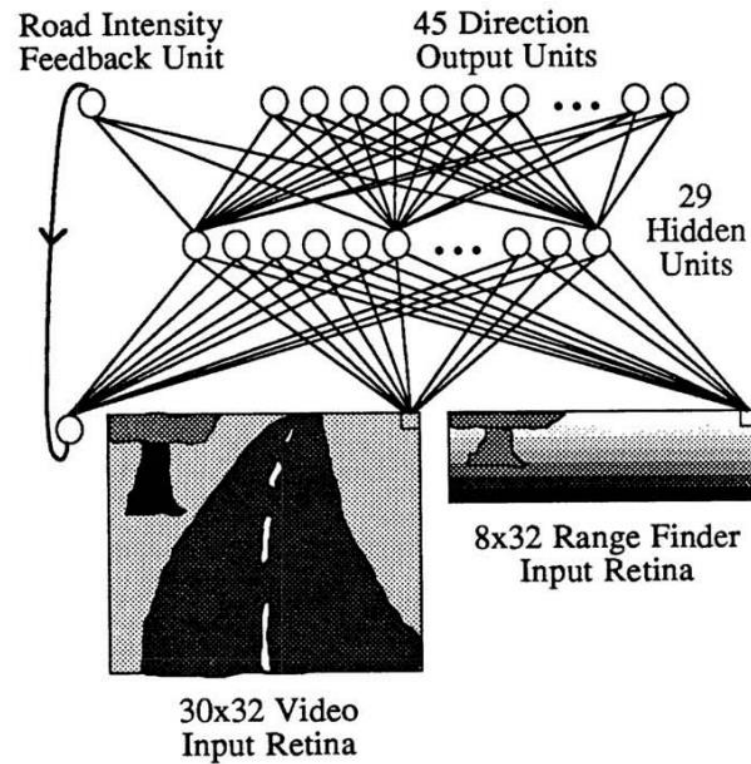
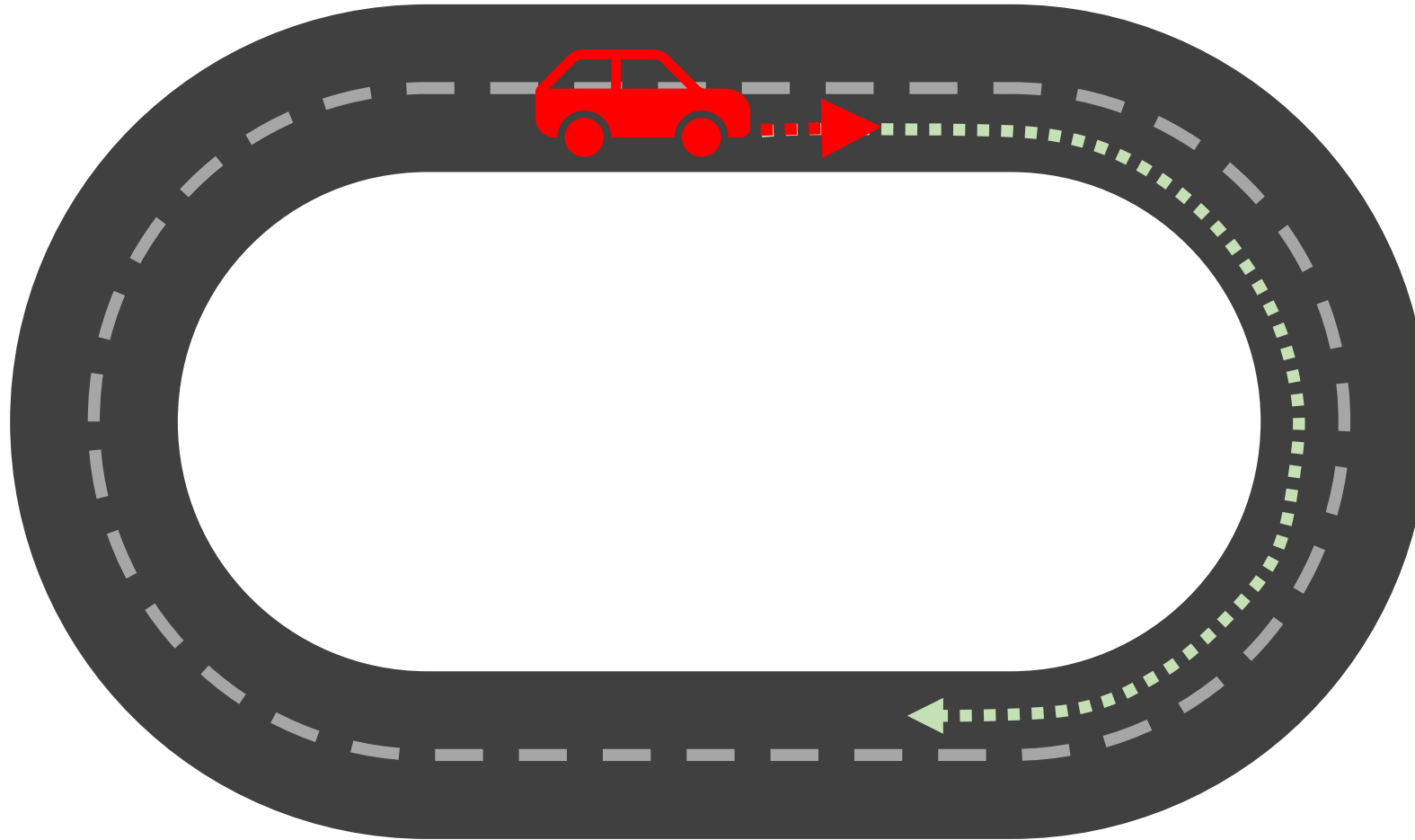


Figure 1: ALVINN Architecture

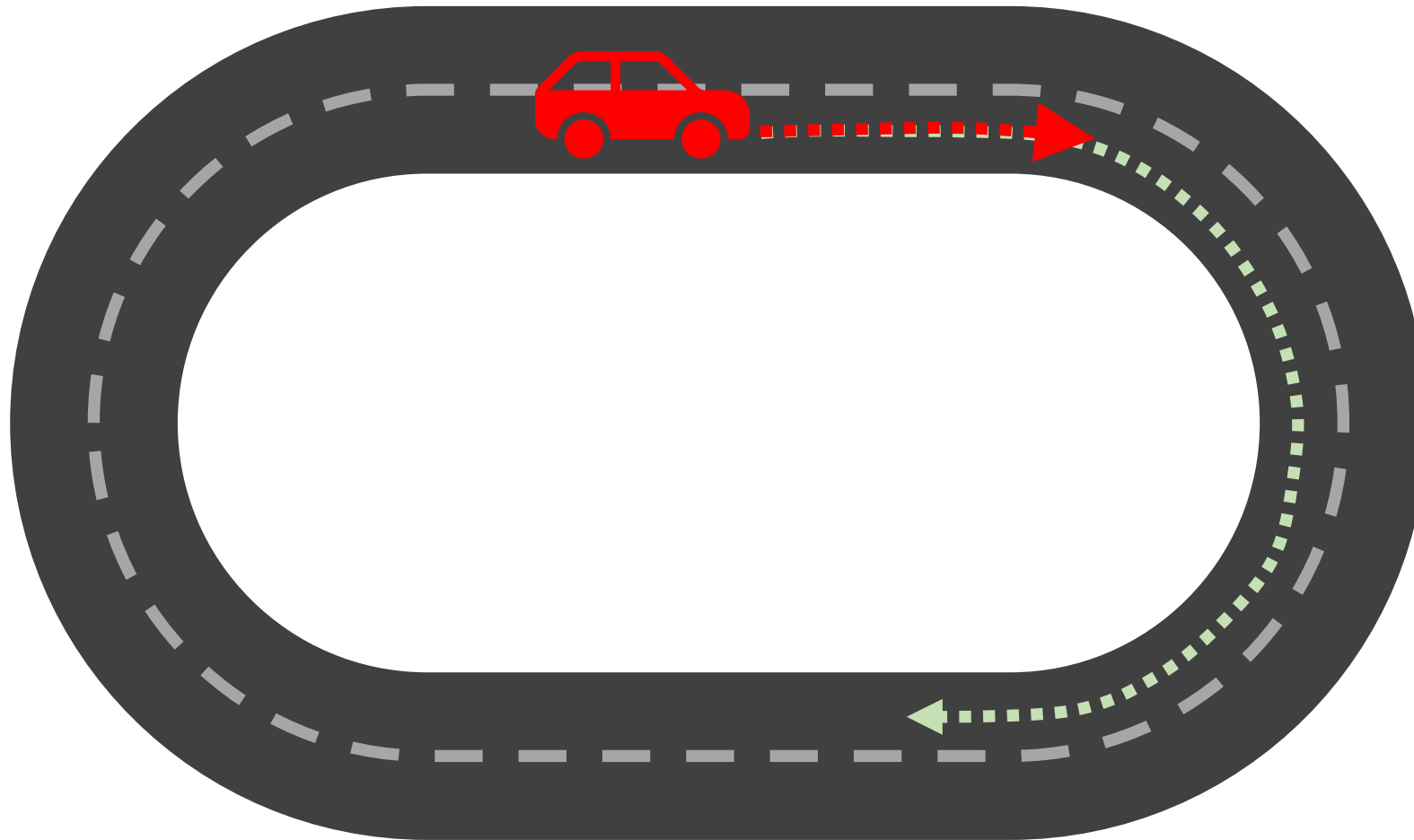


ALVINN: An Autonomous Land Vehicle in a Neural Network
[Pomerleau 1989]

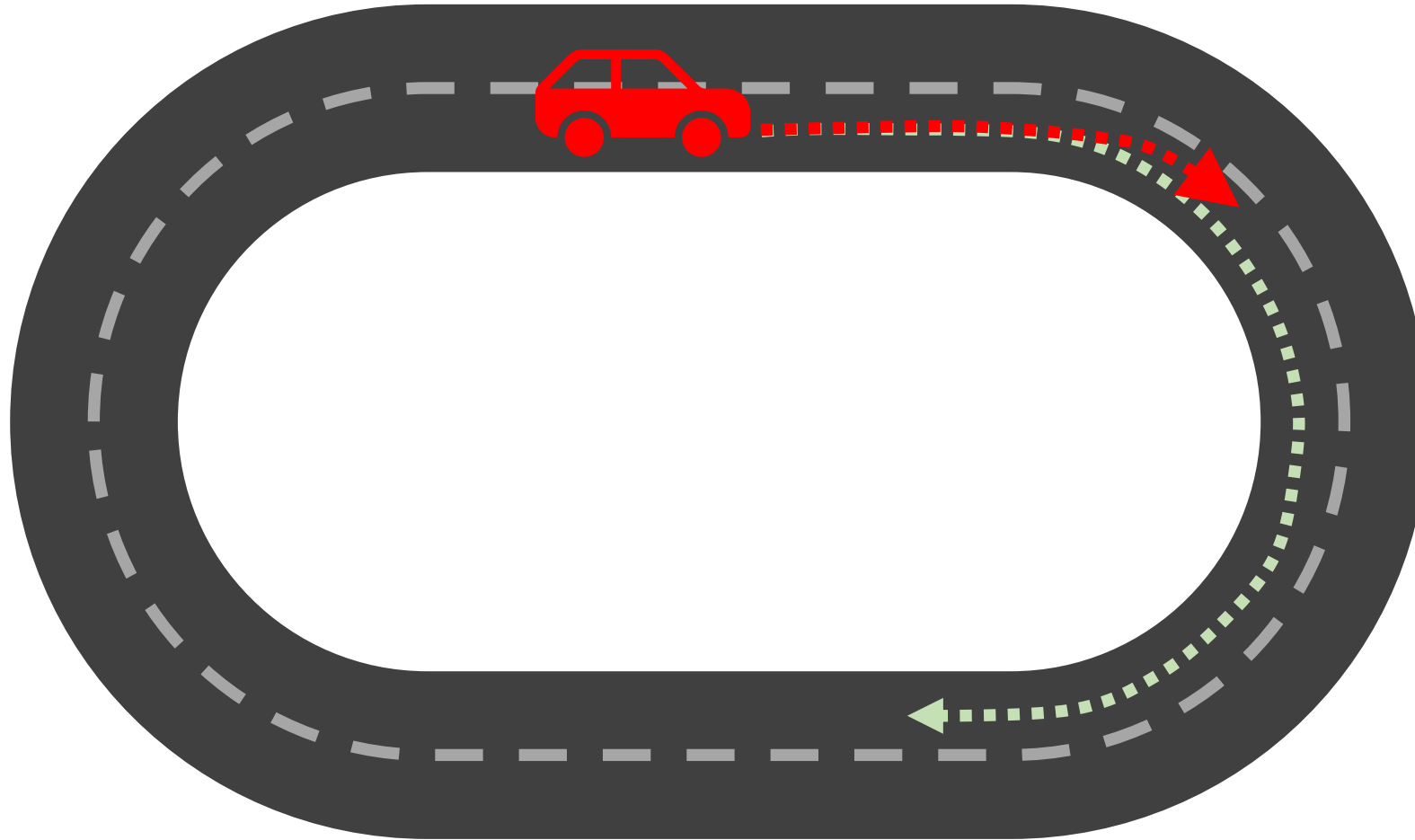
Does it work?



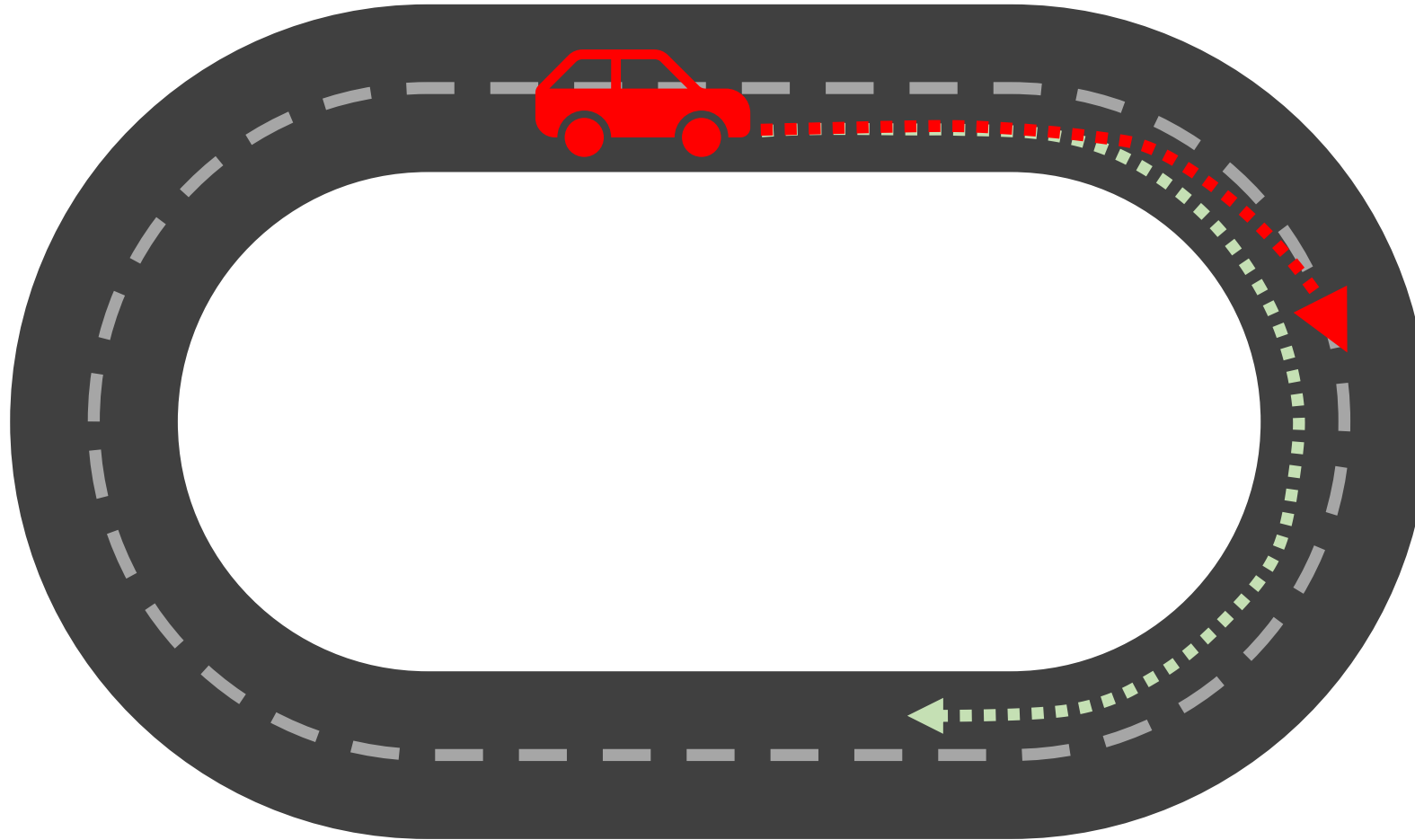
Does it work?



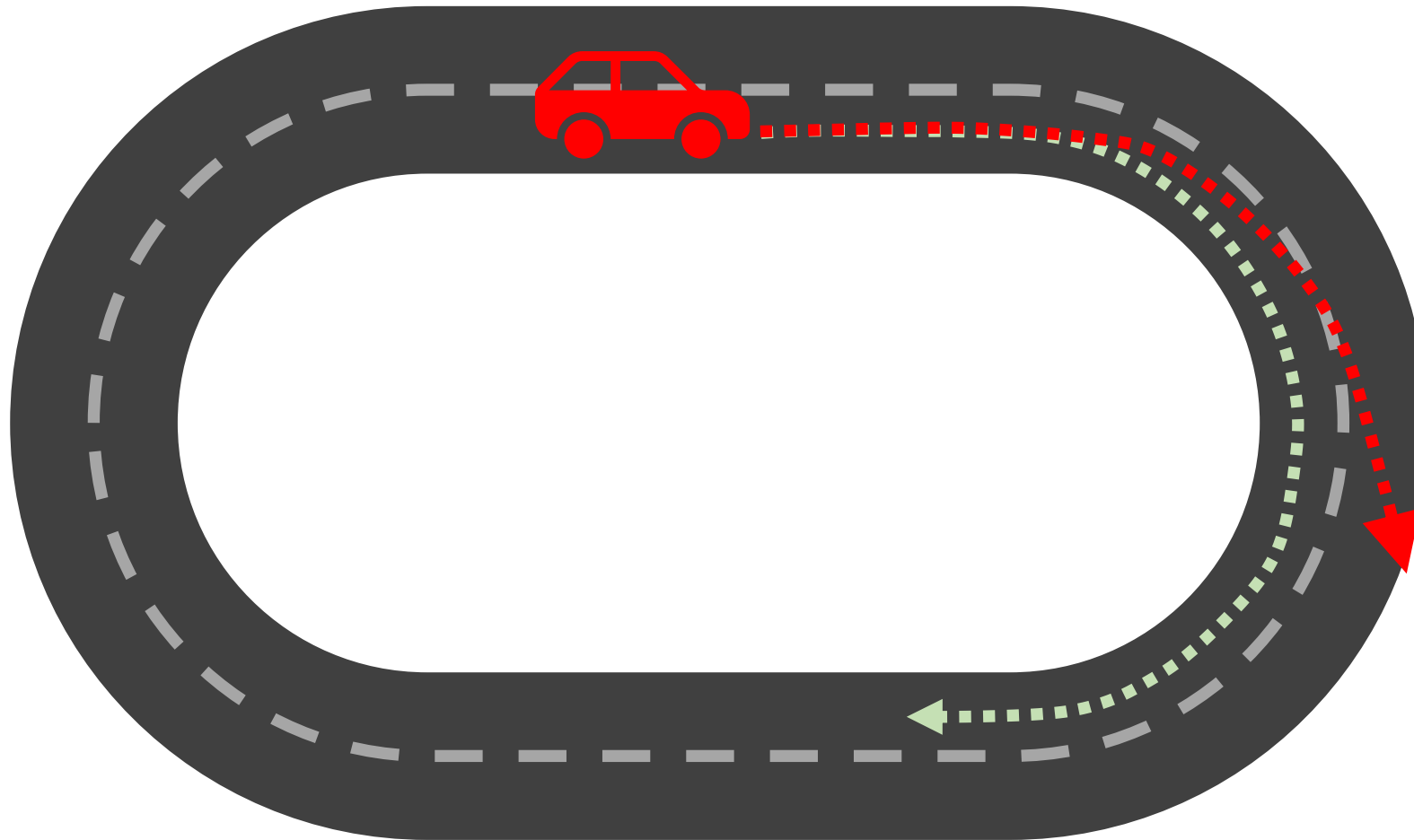
Does it work?



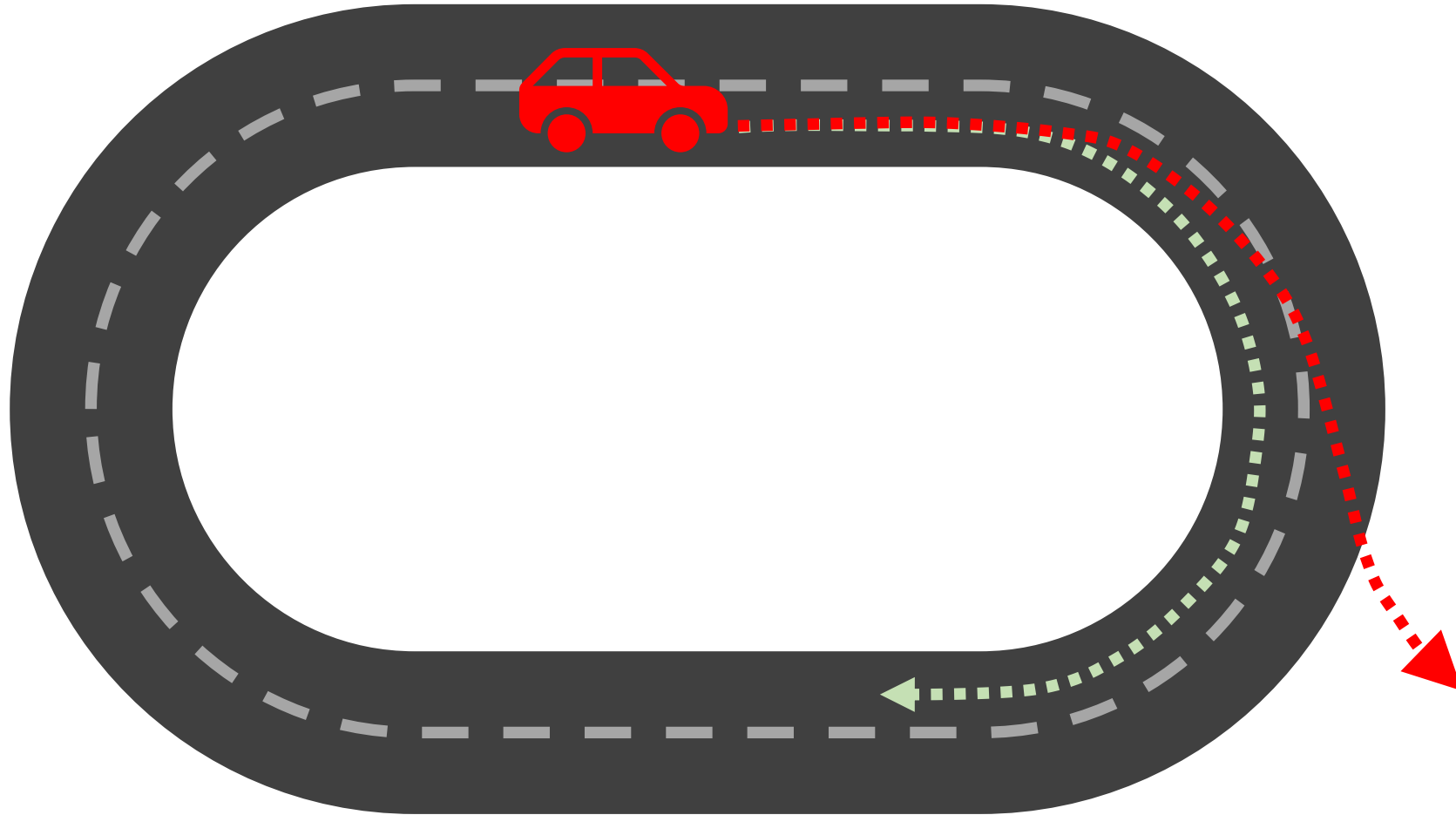
Does it work?



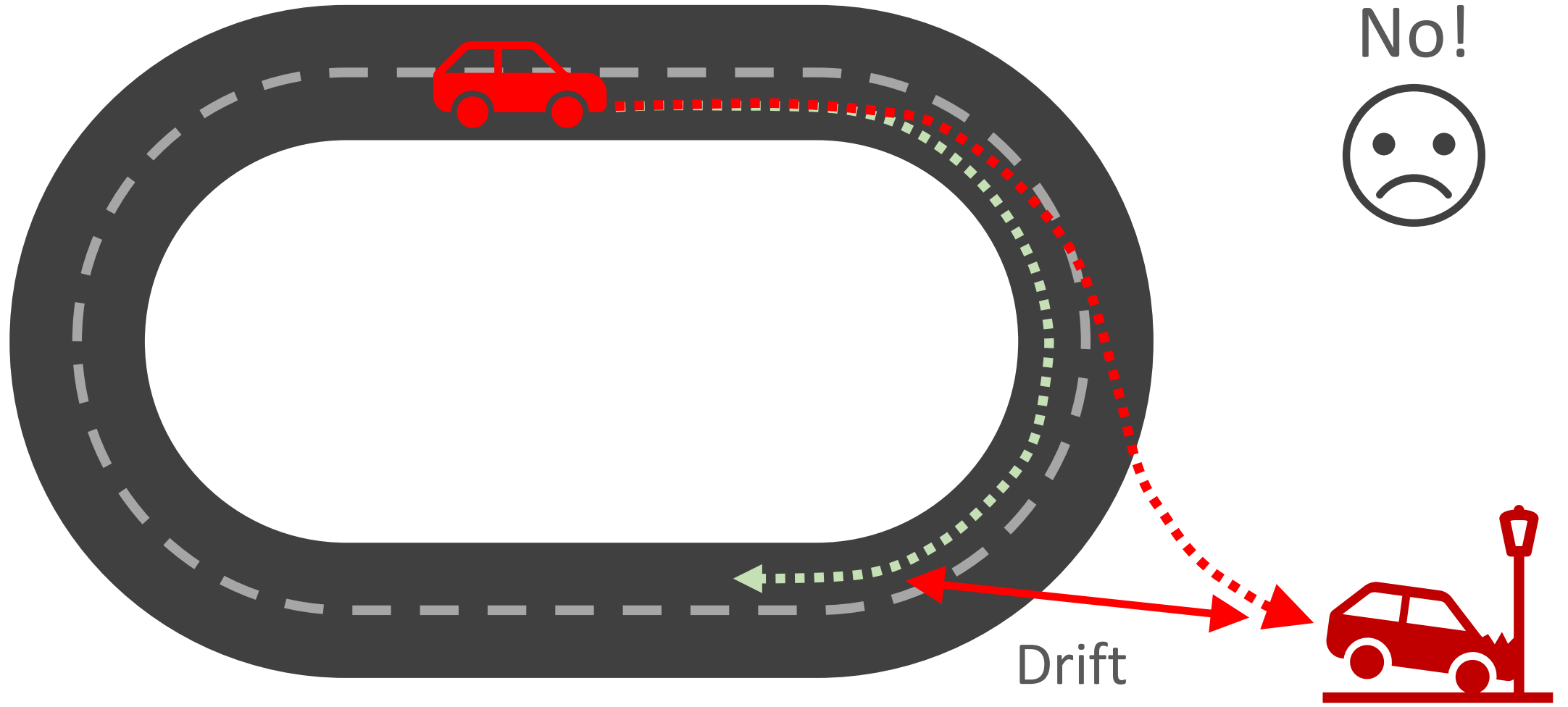
Does it work?



Does it work?



Does it work?



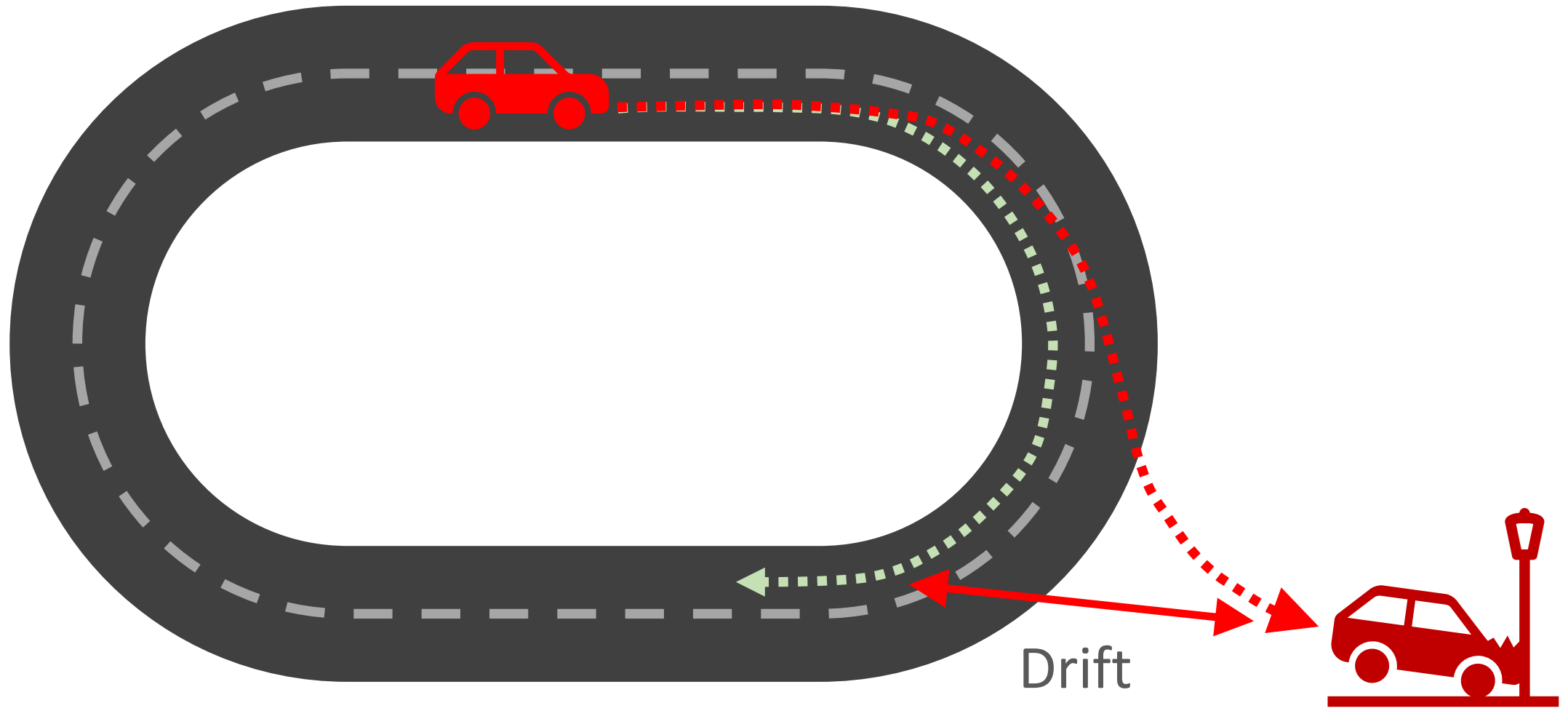
Drift

- Expert is too good
- Lack of corrective feedback
- Policy inaccuracies
- Errors compound over time

Drift

- Expert is too good
- Lack of corrective feedback
- Policy inaccuracies
- Errors compound over time

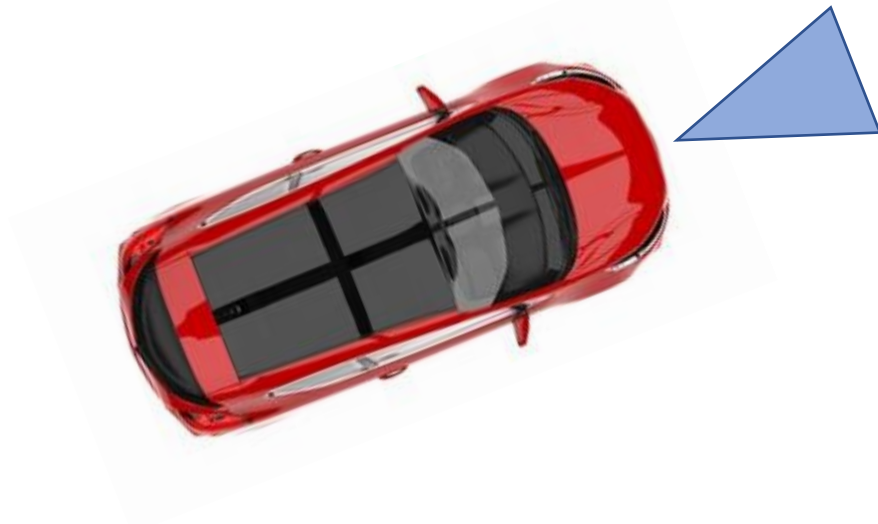
Feedback



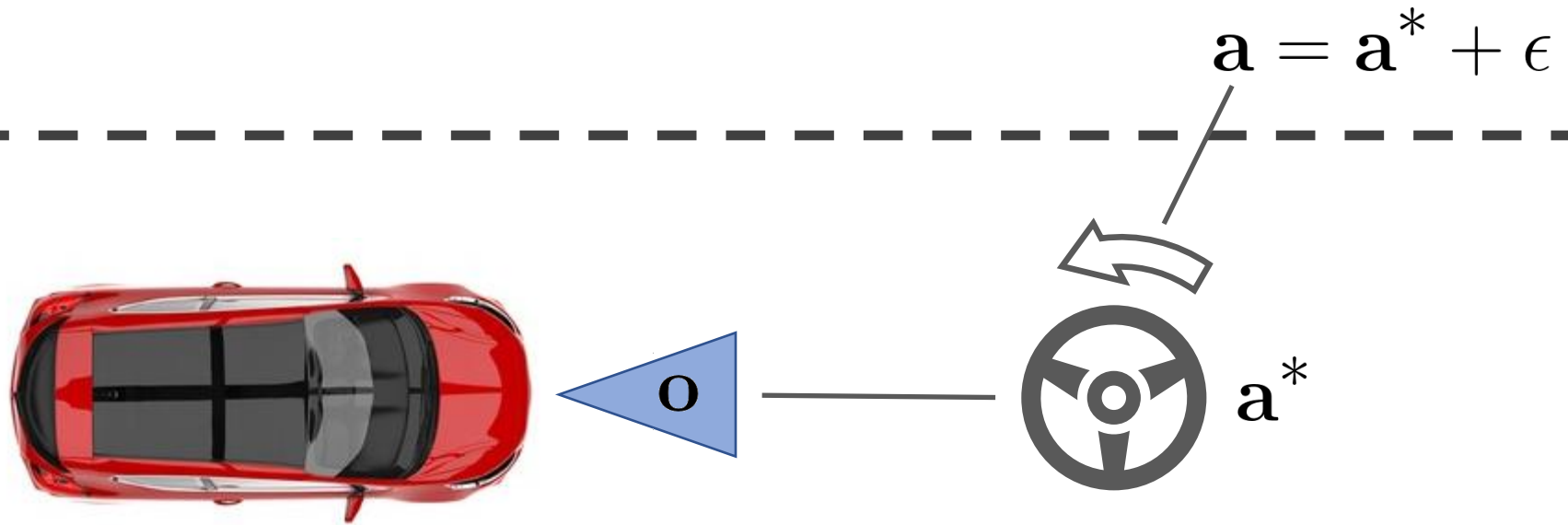
Feedback



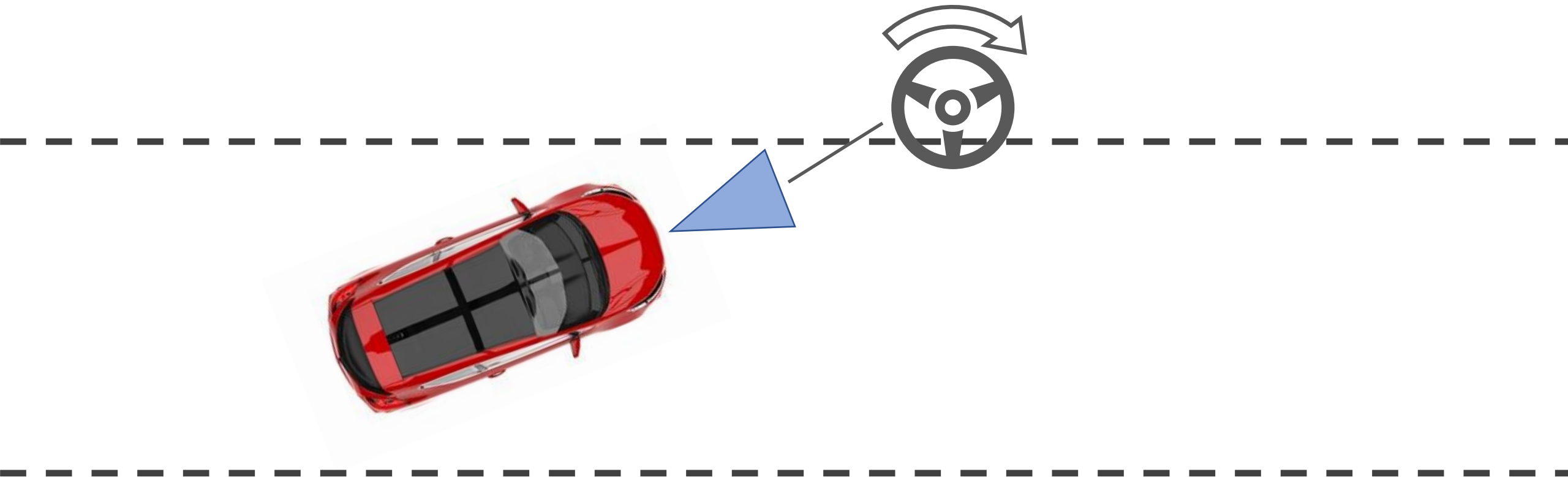
Feedback



Noise Injection



Noise Injection



DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection

ALGORITHM 2: BC with Noise Injection

- 1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset
 - 2: **for** timestep t **do**
 - 3: $\mathbf{o}_t \leftarrow$ record observation
 - 4: $\mathbf{a}_t^* \leftarrow$ query expert for an action
 - 5: $\epsilon_t \leftarrow$ sample noise
 - 6: $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
 - 7: Apply \mathbf{a}_t to environment
 - 8: Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset \mathcal{D}
 - 9: **end for**
 - 10: $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
 - 11: return π^{BC}
-

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection

ALGORITHM 2: BC with Noise Injection

```
1:  $\mathcal{D} \leftarrow \emptyset$  initialize dataset

2: for timestep  $t$  do
3:    $\mathbf{o}_t \leftarrow$  record observation
4:    $\mathbf{a}_t^* \leftarrow$  query expert for an action

5:    $\epsilon_t \leftarrow$  sample noise
6:    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$ 
7:   Apply  $\mathbf{a}_t$  to environment

8:   Store  $(\mathbf{o}_t, \mathbf{a}_t^*)$  in dataset  $\mathcal{D}$ 
9: end for

10:  $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$ 
11: return  $\pi^{\text{BC}}$ 
```

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection

ALGORITHM 2: BC with Noise Injection

- 1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset
 - 2: **for** timestep t **do**
 - 3: $\mathbf{o}_t \leftarrow$ record observation
 - 4: $\mathbf{a}_t^* \leftarrow$ query expert for an action
 - 5: $\epsilon_t \leftarrow$ sample noise
 - 6: $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
 - 7: Apply \mathbf{a}_t to environment
 - 8: Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset \mathcal{D}
 - 9: **end for**
 - 10: $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
 - 11: return π^{BC}
-

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection

ALGORITHM 2: BC with Noise Injection

- 1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset
 - 2: **for** timestep t **do**
 - 3: $\mathbf{o}_t \leftarrow$ record observation
 - 4: $\mathbf{a}_t^* \leftarrow$ query expert for an action
 - 5: $\epsilon_t \leftarrow$ sample noise
 - 6: $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
 - 7: Apply \mathbf{a}_t to environment
 - 8: Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset \mathcal{D}
 - 9: **end for**
 - 10: $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
 - 11: return π^{BC}
-

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection

ALGORITHM 2: BC with Noise Injection

- 1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset
 - 2: **for** timestep t **do**
 - 3: $\mathbf{o}_t \leftarrow$ record observation
 - 4: $\mathbf{a}_t^* \leftarrow$ query expert for an action
 - 5: $\epsilon_t \leftarrow$ sample noise
 - 6: $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
 - 7: Apply \mathbf{a}_t to environment
 - 8: Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset \mathcal{D}
 - 9: **end for**
 - 10: $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
 - 11: return π^{BC}
-

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection

ALGORITHM 2: BC with Noise Injection

- 1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset
 - 2: **for** timestep t **do**
 - 3: $\mathbf{o}_t \leftarrow$ record observation
 - 4: $\mathbf{a}_t^* \leftarrow$ query expert for an action
 - 5: $\epsilon_t \leftarrow$ sample noise
 - 6: $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
 - 7: Apply \mathbf{a}_t to environment
 - 8: Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset \mathcal{D}
 - 9: **end for**
 - 10: $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
 - 11: return π^{BC}
-

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection

ALGORITHM 2: BC with Noise Injection

- 1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset
 - 2: **for** timestep t **do**
 - 3: $\mathbf{o}_t \leftarrow$ record observation
 - 4: $\mathbf{a}_t^* \leftarrow$ query expert for an action
 - 5: $\epsilon_t \leftarrow$ sample noise
 - 6: $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
 - 7: Apply \mathbf{a}_t to environment
 - 8: Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset \mathcal{D}
 - 9: **end for**
 - 10: $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
 - 11: return π^{BC}
-

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection

ALGORITHM 2: BC with Noise Injection

```
1:  $\mathcal{D} \leftarrow \emptyset$  initialize dataset

2: for timestep  $t$  do
3:    $\mathbf{o}_t \leftarrow$  record observation
4:    $\mathbf{a}_t^* \leftarrow$  query expert for an action

5:    $\epsilon_t \leftarrow$  sample noise
6:    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$ 
7:   Apply  $\mathbf{a}_t$  to environment

8:   Store  $(\mathbf{o}_t, \mathbf{a}_t^*)$  in dataset  $\mathcal{D}$ 
9: end for

10:  $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$ 
11: return  $\pi^{\text{BC}}$ 
```

Noise Injection

ALGORITHM 2: BC with Noise Injection

```
1:  $\mathcal{D} \leftarrow \emptyset$  initialize dataset

2: for timestep  $t$  do
3:    $\mathbf{o}_t \leftarrow$  record observation
4:    $\mathbf{a}_t^* \leftarrow$  query expert for an action

5:    $\epsilon_t \leftarrow$  sample noise
6:    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$ 
7:   Apply  $\mathbf{a}_t$  to environment

8:   Store  $(\mathbf{o}_t, \mathbf{a}_t^*)$  in dataset  $\mathcal{D}$ 
9: end for

10:  $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$ 
11: return  $\pi^{\text{BC}}$ 
```

Noise Injection

ALGORITHM 2: BC with Noise Injection

```
1:  $\mathcal{D} \leftarrow \emptyset$  initialize dataset

2: for timestep  $t$  do
3:    $\mathbf{o}_t \leftarrow$  record observation
4:    $\mathbf{a}_t^* \leftarrow$  query expert for an action

5:    $\epsilon_t \leftarrow$  sample noise
6:    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$ 
7:   Apply  $\mathbf{a}_t$  to environment

8:   Store  $(\mathbf{o}_t, \mathbf{a}_t^*)$  in dataset  $\mathcal{D}$ 
9: end for

10:  $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$ 
11: return  $\pi^{\text{BC}}$ 
```

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection

ALGORITHM 2: BC with Noise Injection

- 1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset
 - 2: **for** timestep t **do**
 - 3: $\mathbf{o}_t \leftarrow$ record observation
 - 4: $\mathbf{a}_t^* \leftarrow$ query expert for an action
 - 5: $\epsilon_t \leftarrow$ sample noise
 - 6: $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
 - 7: Apply \mathbf{a}_t to environment
 - 8: Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset \mathcal{D}
 - 9: **end for**
 - 10: $\pi^{\text{BC}} = \arg \min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
 - 11: return π^{BC}
-

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

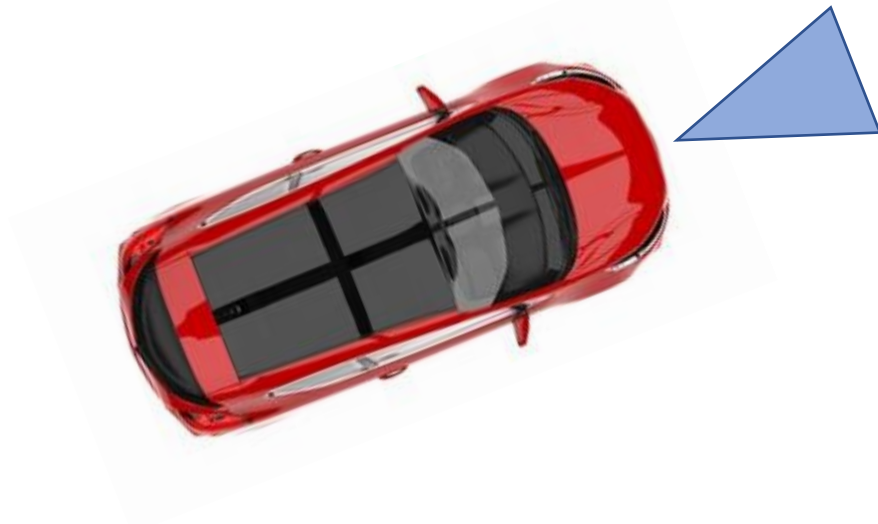
Noise Injection

- ✓ Simple method to get corrective feedback
- ✓ Can work well in practice
- ✗ Dangerous for expert!
- ✗ Difficult to pick effective perturbations

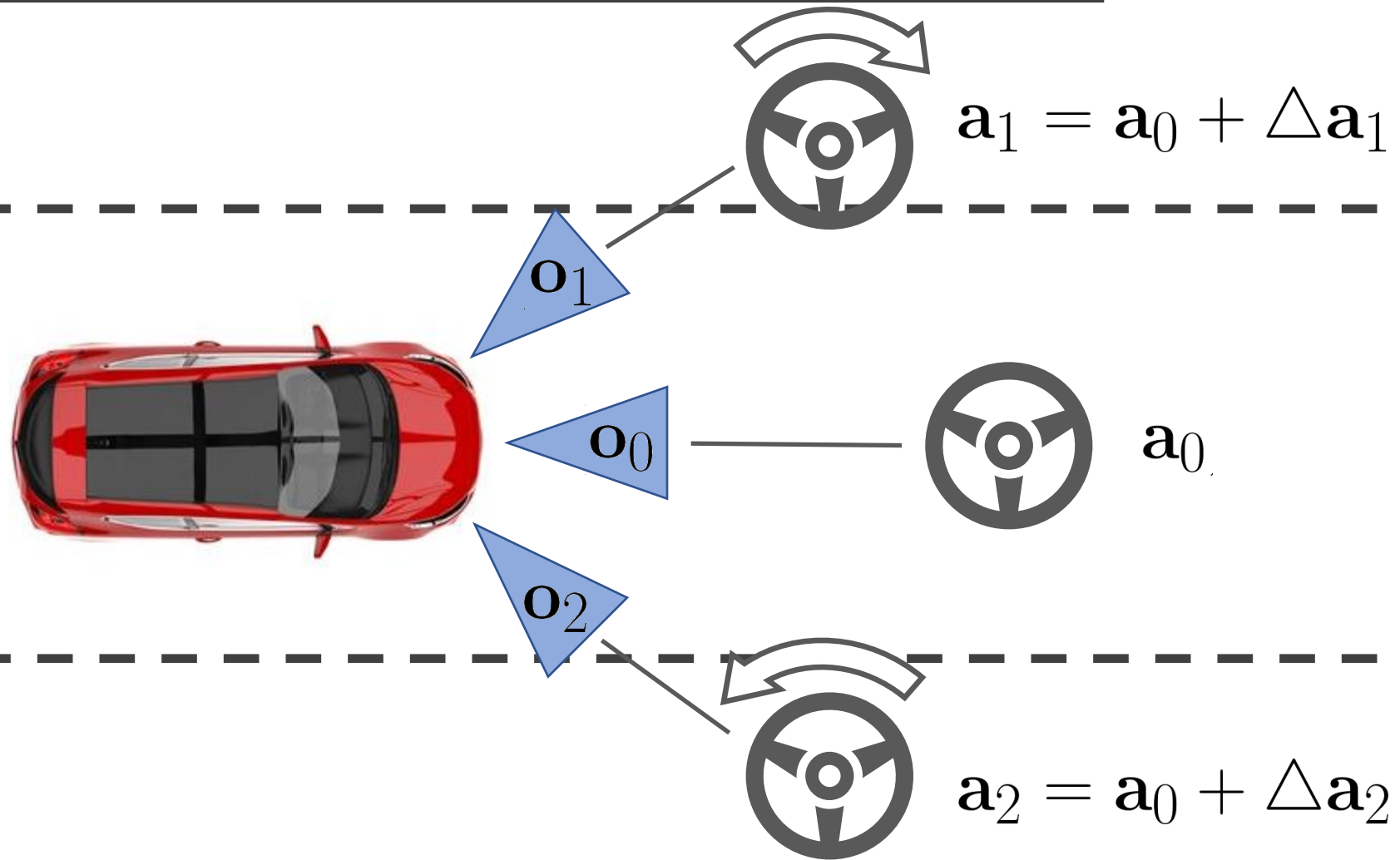
Data Augmentation



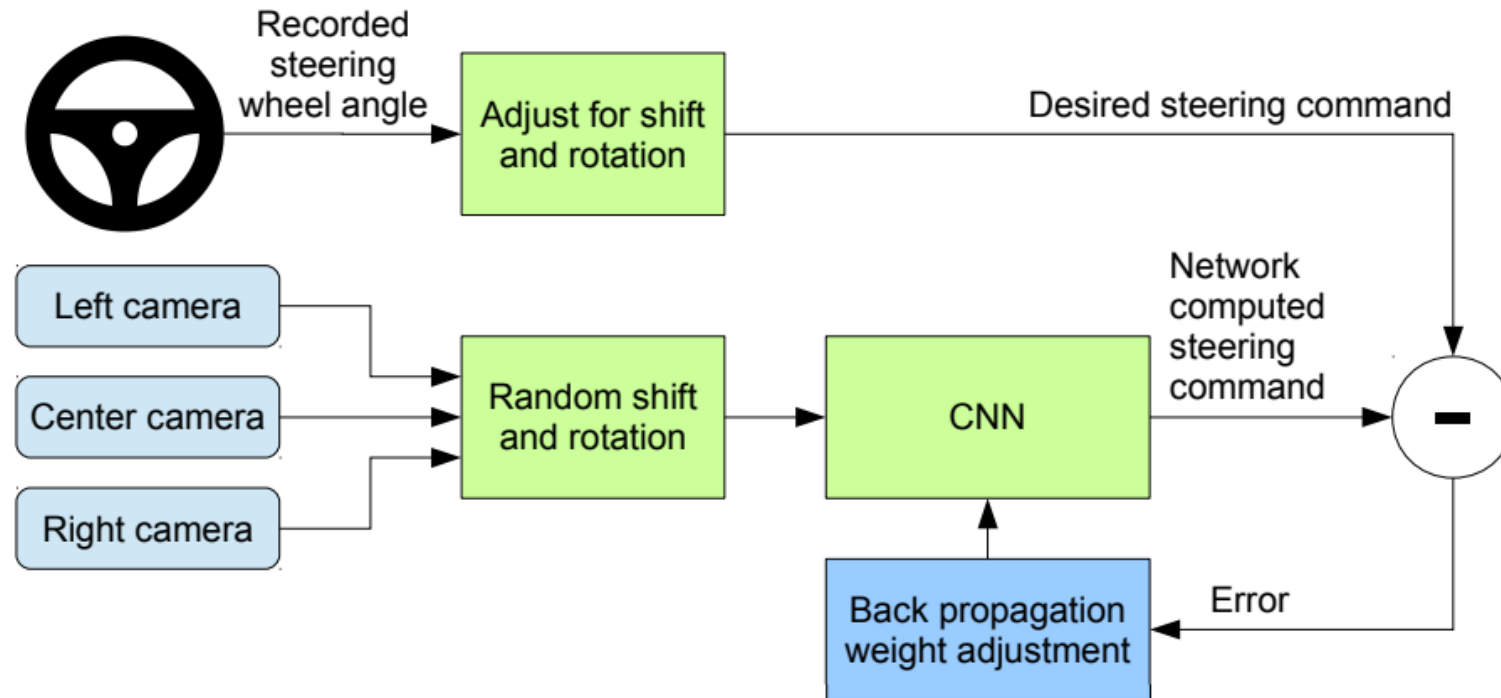
Data Augmentation



Data Augmentation



Data Augmentation



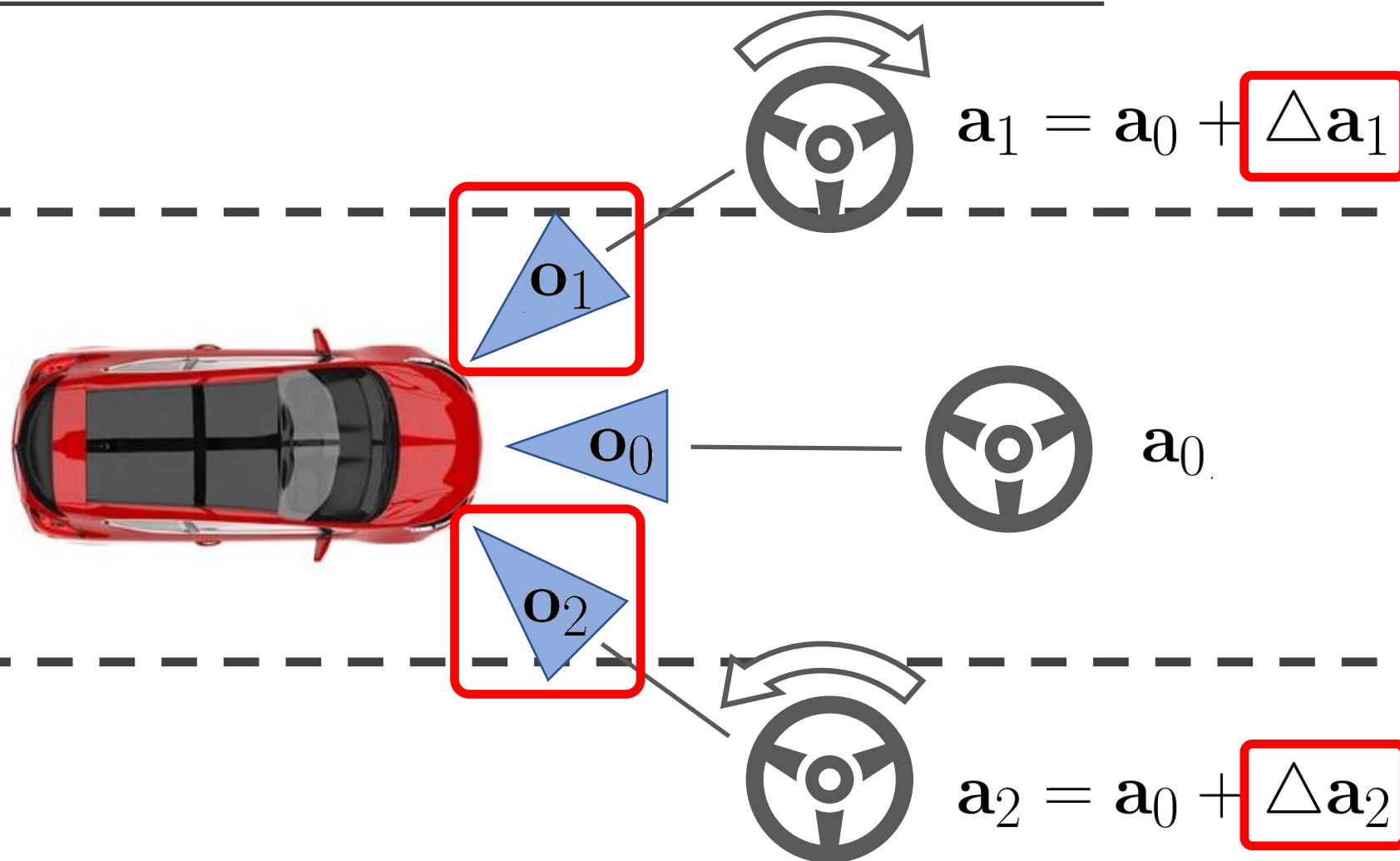
End to End Learning for Self-Driving Cars
[Bojarski et al. 2016]

Data Augmentation



End to End Learning for Self-Driving Cars
[Bojarski et al. 2016]

Data Augmentation



Drift

- Expert is too good
- Lack of corrective feedback
- Policy inaccuracies
- Errors compound over time

Theoretical Analysis

Analyze the number of mistakes π makes over time

Theorem 1. The number of mistakes grow $O(\epsilon T^2)$

Theoretical Analysis

Given dataset sampled from $p_{\text{data}}(\mathbf{s}, \mathbf{a})$

$$\min_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim p_{\text{data}}(\mathbf{s}, \mathbf{a})} [-\log \pi(\mathbf{a} | \mathbf{s})]$$

Such that

$$\pi(\mathbf{a} \neq \pi^*(\mathbf{s}) | \mathbf{s}) \leq \epsilon \text{ for all } \mathbf{s} \sim p_{\text{data}}(\mathbf{s})$$

i.e. the probability of π making a mistake is bounded.

$$\text{Cost: } c(\mathbf{s}, \mathbf{a}) = \begin{cases} 0 & \text{if } \mathbf{a} = \pi^*(\mathbf{s}) \\ 1 & \text{otherwise} \end{cases}$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$\underline{p_{\pi}^t(\mathbf{s})} = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

probability of being in \mathbf{s} after following π for t timesteps

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = \underbrace{(1 - \epsilon)^t}_{\text{no mistakes in } t \text{ timesteps}} p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

no mistakes in t timesteps

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t \underbrace{p_{\text{data}}^t(\mathbf{s})}_{\text{no mistakes in } t \text{ timesteps}} + \underbrace{(1 - (1 - \epsilon)^t)}_{\text{at least 1 mistakes in } t \text{ timesteps}} \underbrace{p_{\text{mistake}}^t(\mathbf{s})}_{\text{at least 1 mistakes in } t \text{ timesteps}}$$

no mistakes in t timesteps

at least 1 mistakes in t timesteps

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]$$

expected cost

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{\underline{p_{\pi}^t(\mathbf{s})}} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [\underline{c(\mathbf{s}, \mathbf{a})}]$$

$$c(\mathbf{s}, \mathbf{a}) = \begin{cases} 0 & \text{if } \mathbf{a} = \pi^*(\mathbf{s}) \\ 1 & \text{otherwise} \end{cases}$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \underbrace{\mathbb{E}_{p_{\pi}^t(\mathbf{s})}} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] = \sum_t \sum_{\underbrace{\mathbf{s}}} p_{\pi}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - \underbrace{p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s})}_{=0} \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \end{aligned}$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} \underline{p_{\text{data}}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]} + \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \end{aligned}$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} p_{\text{data}}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] + \sum_t \sum_{\mathbf{s}} \left(\underline{p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s})} \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \end{aligned}$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} p_{\text{data}}^t(\mathbf{s}) \underbrace{\mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]}_{\leq \epsilon} + \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \end{aligned}$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \underbrace{\sum_t \sum_{\mathbf{s}} p_{\text{data}}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]}_{\leq \epsilon} + \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \end{aligned}$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \underbrace{\sum_t \sum_{\mathbf{s}} p_{\text{data}}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]}_{\leq \epsilon T} + \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \end{aligned}$$

Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\pi}^t(\mathbf{s}) = (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &= \sum_t \sum_{\mathbf{s}} p_{\text{data}}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] + \sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &\leq \epsilon T + \underbrace{\sum_t \sum_{\mathbf{s}} \left(p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]}_{?} \end{aligned}$$

Theoretical Analysis

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$
$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - \underline{(1-\epsilon)^t p_{\text{data}}^t(\mathbf{s})} = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

Theoretical Analysis

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})$$

Theoretical Analysis

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - \underbrace{(1-\epsilon)^t p_{\text{data}}^t(\mathbf{s})} - \underbrace{\left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})} = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})$$

Theoretical Analysis

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \underbrace{\left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})} - \underbrace{\left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})}$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s})$$

Theoretical Analysis

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) &= \left(1 - (1-\epsilon)^t\right) \sum_{\mathbf{s}} p_{\text{mistake}}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \\ &\leq \left(1 - (1-\epsilon)^t\right) \underbrace{\sum_{\mathbf{s}} \left| p_{\text{mistake}}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right|}_{\text{total variation distance} \leq 2} \end{aligned}$$

Theoretical Analysis

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) = \left(1 - (1-\epsilon)^t\right) \sum_{\mathbf{s}} p_{\text{mistake}}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s})$$

$$\leq \left(1 - (1-\epsilon)^t\right) \sum_{\mathbf{s}} \left| p_{\text{mistake}}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right|$$

$$\leq 2 \left(1 - (1-\epsilon)^t\right)$$

$$\leq 2\epsilon t$$

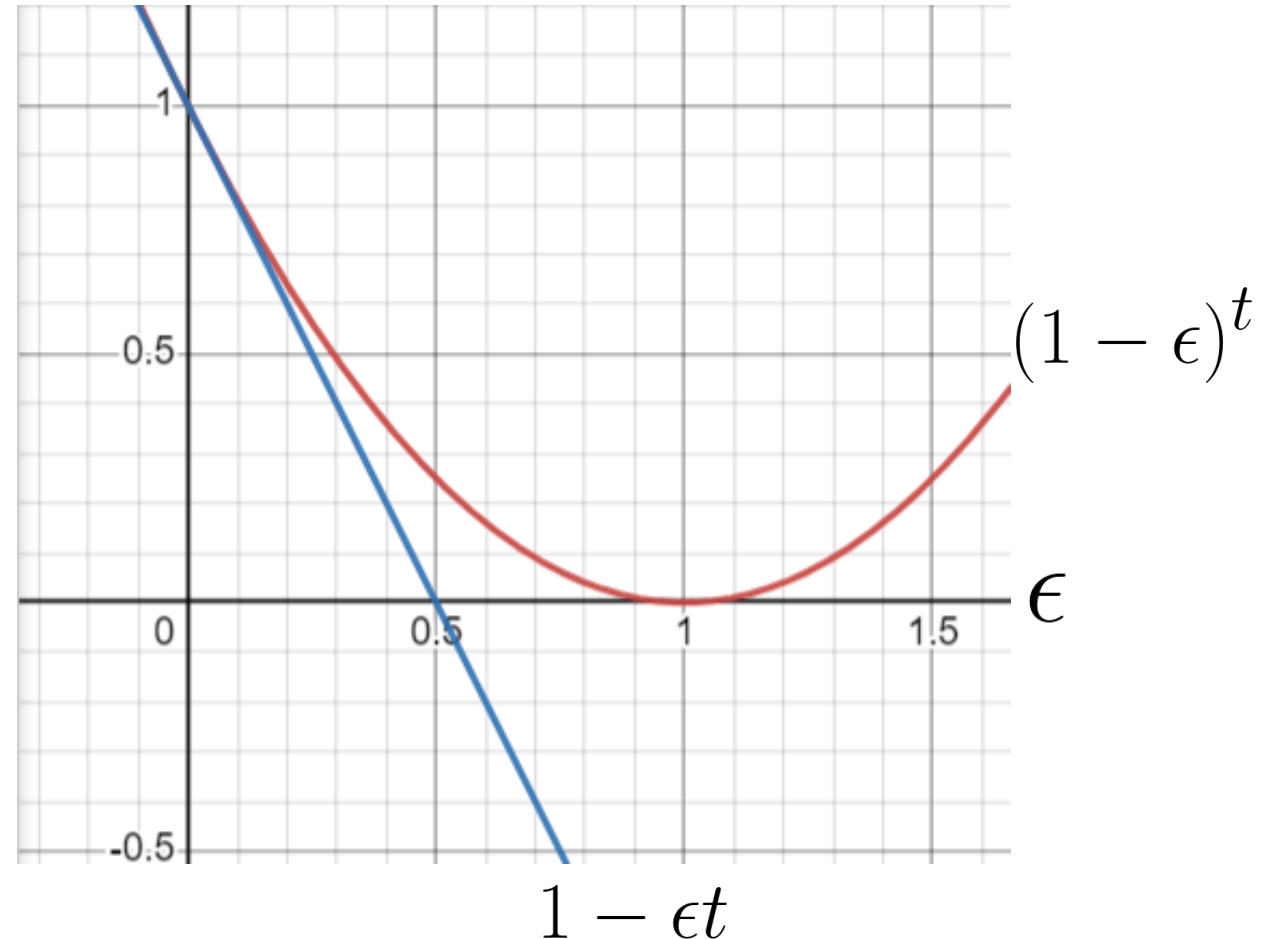
Note: $(1-\epsilon)^t \geq 1 - \epsilon t$ for $\epsilon \in [0, 1]$

Theoretical Analysis

$$\begin{aligned}\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) &\leq 2 \left(1 - (1 - \epsilon)^t\right) \\ &\leq 2(1 - (1 - \epsilon t)) \\ &\leq 2\epsilon t\end{aligned}$$

$$\sum_{\mathbf{s}} p_{\pi}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \leq 2\epsilon t$$

Note: $(1 - \epsilon)^t \geq 1 - \epsilon t$ for $\epsilon \in [0, 1]$



Theoretical Analysis

$$\begin{aligned} \sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &\leq \epsilon T + \sum_t \underbrace{\sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right)}_{\leq 2\epsilon t} \underbrace{\mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]}_{\leq 1} \end{aligned}$$

Theoretical Analysis

$$\begin{aligned}\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &\leq \epsilon T + \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &\leq \epsilon T + \sum_t 2\epsilon t\end{aligned}$$

Theoretical Analysis

$$\begin{aligned} \sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &\leq \epsilon T + \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &\leq \epsilon T + \sum_t 2\epsilon t \\ &\leq \epsilon T + 2\epsilon T^2 \in \boxed{O(\epsilon T^2)} \end{aligned}$$

Worst Case

$$\sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \leq \epsilon T + 2\epsilon T^2$$



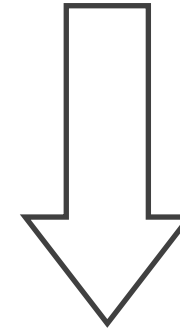
Worst Case

$$\sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \leq \epsilon T + 2\epsilon T^2$$




Distribution Shift

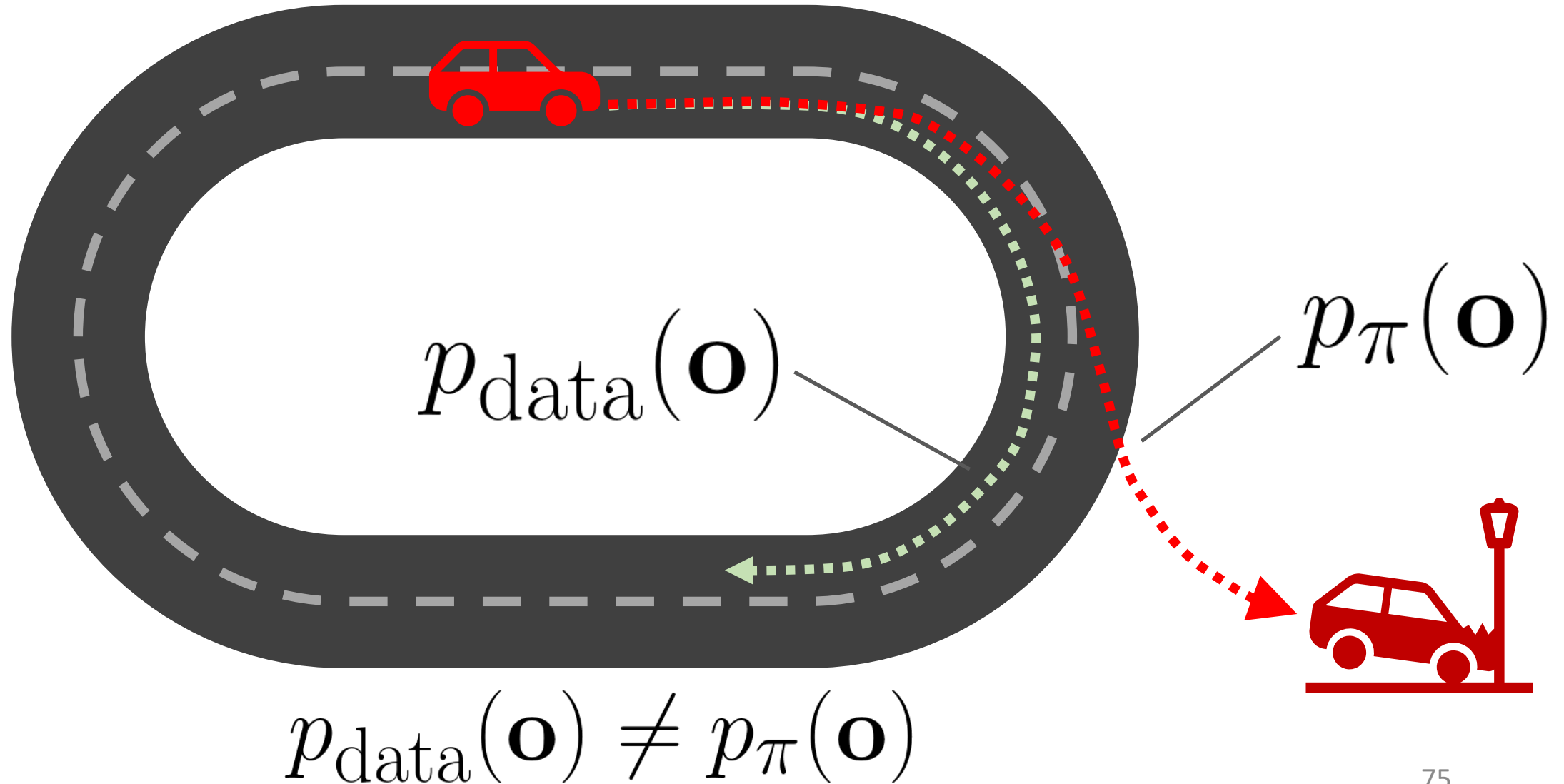
$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \leq \epsilon T + 2\epsilon T^2$$



$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \leq \epsilon T + \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]$$


$$p_\pi^t(\mathbf{s}) \neq p_{\text{data}}^t(\mathbf{s})$$

Distribution Shift



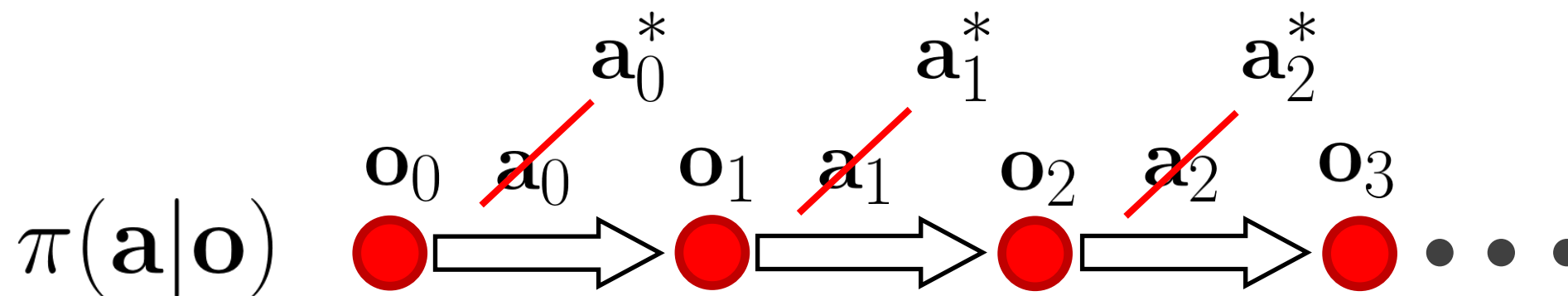
Dataset Aggregation

Can we make $p_{\text{data}}(\mathbf{o}) = p_{\pi}(\mathbf{o})$?

Key idea:

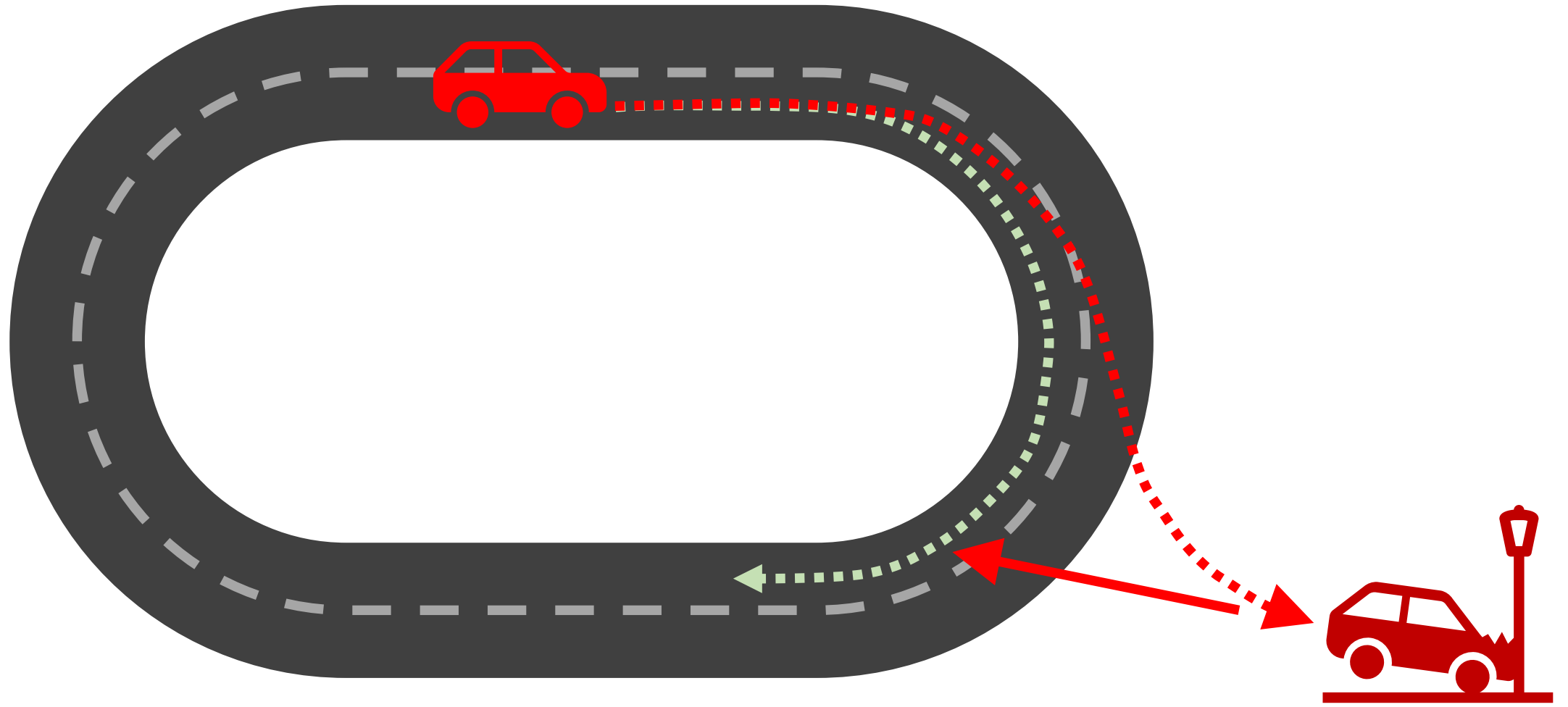
- Collect observations from $p_{\pi}(\mathbf{o})$ instead of $p_{\text{data}}(\mathbf{o})$
- Label actions with expert
- DAgger: Dataset Aggregation [Ross et al. 2011]

DAgger



Train with $(\mathbf{o}_i, \mathbf{a}_i^*)$

Dagger



DAgger

ALGORITHM: DAgger

- 1: **for** iteration $i = 0, \dots, k - 1$ **do**
 - 2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, \dots\}$
 - 3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, \dots\}$
 - 4: Label \mathcal{D}_π with actions \mathbf{a}_i from expert
 - 5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
 - 6: **end for**
-

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

DAgger

ALGORITHM: DAgger

- 1: **for** iteration $i = 0, \dots, k - 1$ **do**
 - 2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, \dots\}$
 - 3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, \dots\}$
 - 4: Label \mathcal{D}_π with actions \mathbf{a}_i from expert
 - 5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
 - 6: **end for**
-

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

DAgger

ALGORITHM: DAgger

- 1: **for** iteration $i = 0, \dots, k - 1$ **do**
 - 2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, \dots\}$
 - 3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, \dots\}$
 - 4: Label \mathcal{D}_π with actions \mathbf{a}_i from expert
 - 5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
 - 6: **end for**
-

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

DAgger

ALGORITHM: DAgger

- 1: **for** iteration $i = 0, \dots, k - 1$ **do**
 - 2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, \dots\}$
 - 3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, \dots\}$
 - 4: Label \mathcal{D}_π with actions \mathbf{a}_i from expert
 - 5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
 - 6: **end for**
-

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

DAgger

ALGORITHM: DAgger

- 1: **for** iteration $i = 0, \dots, k - 1$ **do**
 - 2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, \dots\}$
 - 3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, \dots\}$
 - 4: Label \mathcal{D}_π with actions \mathbf{a}_i from expert
 - 5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
 - 6: **end for**
-

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

DAgger

ALGORITHM: DAgger

- 1: **for** iteration $i = 0, \dots, k - 1$ **do**
 - 2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, \dots\}$
 - 3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, \dots\}$
 - 4: Label \mathcal{D}_π with actions \mathbf{a}_i from expert
 - 5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
 - 6: **end for**
-

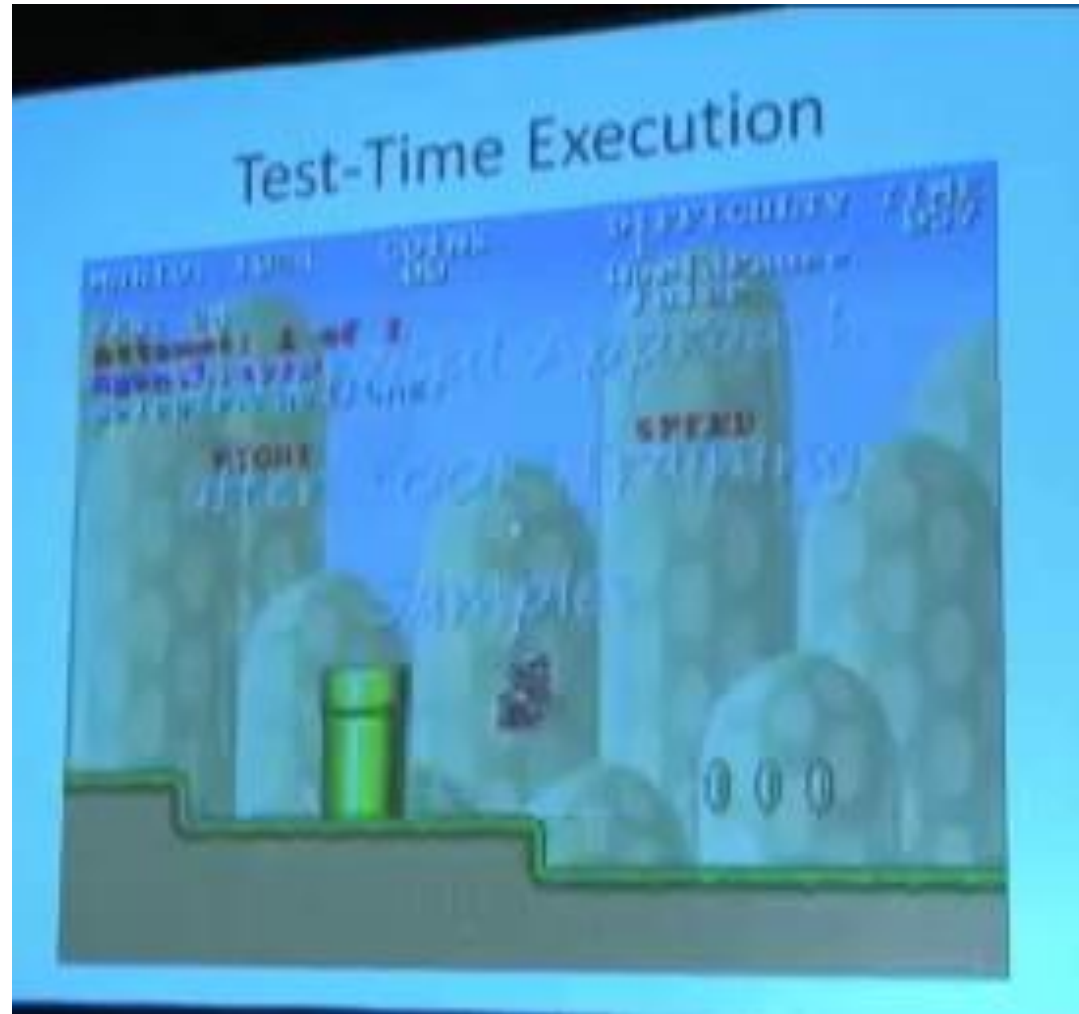
A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

DAgger



A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

DAgger



A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

Dagger Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\text{data}}(\mathbf{s}) = p_{\pi}(\mathbf{s})!$$

$$\begin{aligned} p_{\pi}^t(\mathbf{s}) &= (1 - \epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}^t(\mathbf{s}) \\ &= p_{\text{data}}^t(\mathbf{s}) \end{aligned}$$

DAgger Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\text{data}}(\mathbf{s}) = p_{\pi}(\mathbf{s})!$$

$$p_{\pi}^t(\mathbf{s}) = p_{\text{data}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] = \sum_t \mathbb{E}_{p_{\text{data}}^t(\mathbf{s})} \underbrace{\mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})]}_{\leq \epsilon}$$

DAgger Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_{\text{data}}(\mathbf{s}) = p_{\pi}(\mathbf{s})!$$

$$p_{\pi}^t(\mathbf{s}) = p_{\text{data}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_t \mathbb{E}_{p_{\pi}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] &= \sum_t \mathbb{E}_{p_{\text{data}}^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s}, \mathbf{a})] \\ &\leq \sum_t \epsilon \\ &\leq \epsilon T \in O(\epsilon T) \end{aligned}$$

DAgger

ALGORITHM: DAgger

- 1: **for** iteration $i = 0, \dots, k - 1$ **do**
 - 2: train $\pi(\mathbf{a}|\mathbf{o})$ from expert dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, \dots\}$
 - 3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, \dots\}$
 - 4: Label \mathcal{D}_π with actions \mathbf{a}_i from expert
 - 5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
 - 6: **end for**
-

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

DAgger

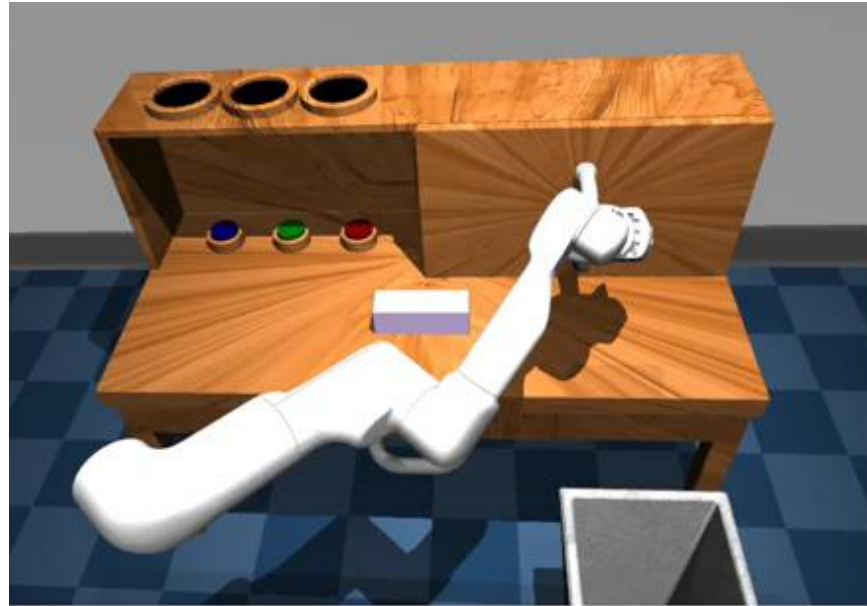
ALGORITHM: DAgger

- 1: **for** iteration $i = 0, \dots, k - 1$ **do**
 - 2: train $\pi(\mathbf{a}|\mathbf{o})$ from expert dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, \dots\}$
 - 3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, \dots\}$
 - 4: Label \mathcal{D}_π with actions \mathbf{a}_i from expert
 - 5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
 - 6: **end for**
-

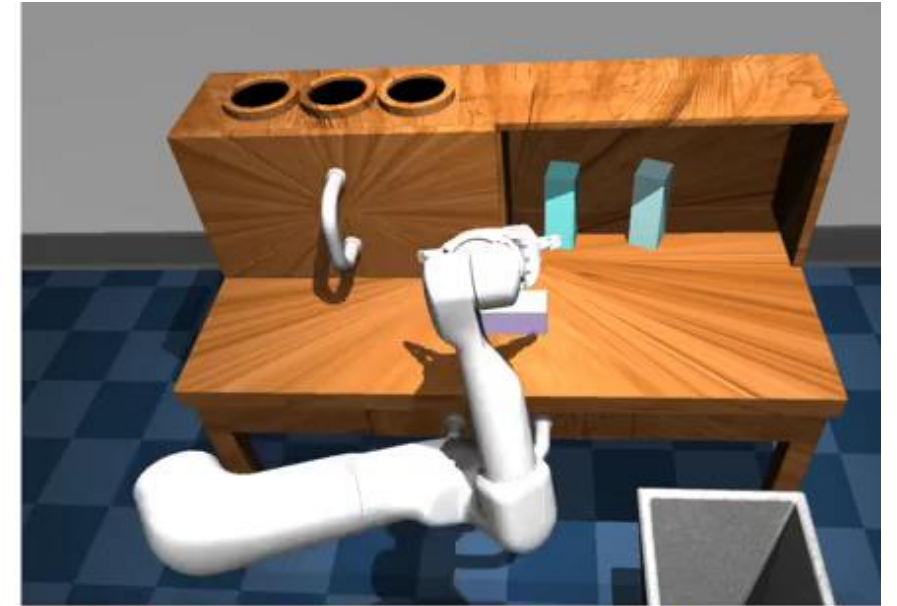
A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

Applications

Applications



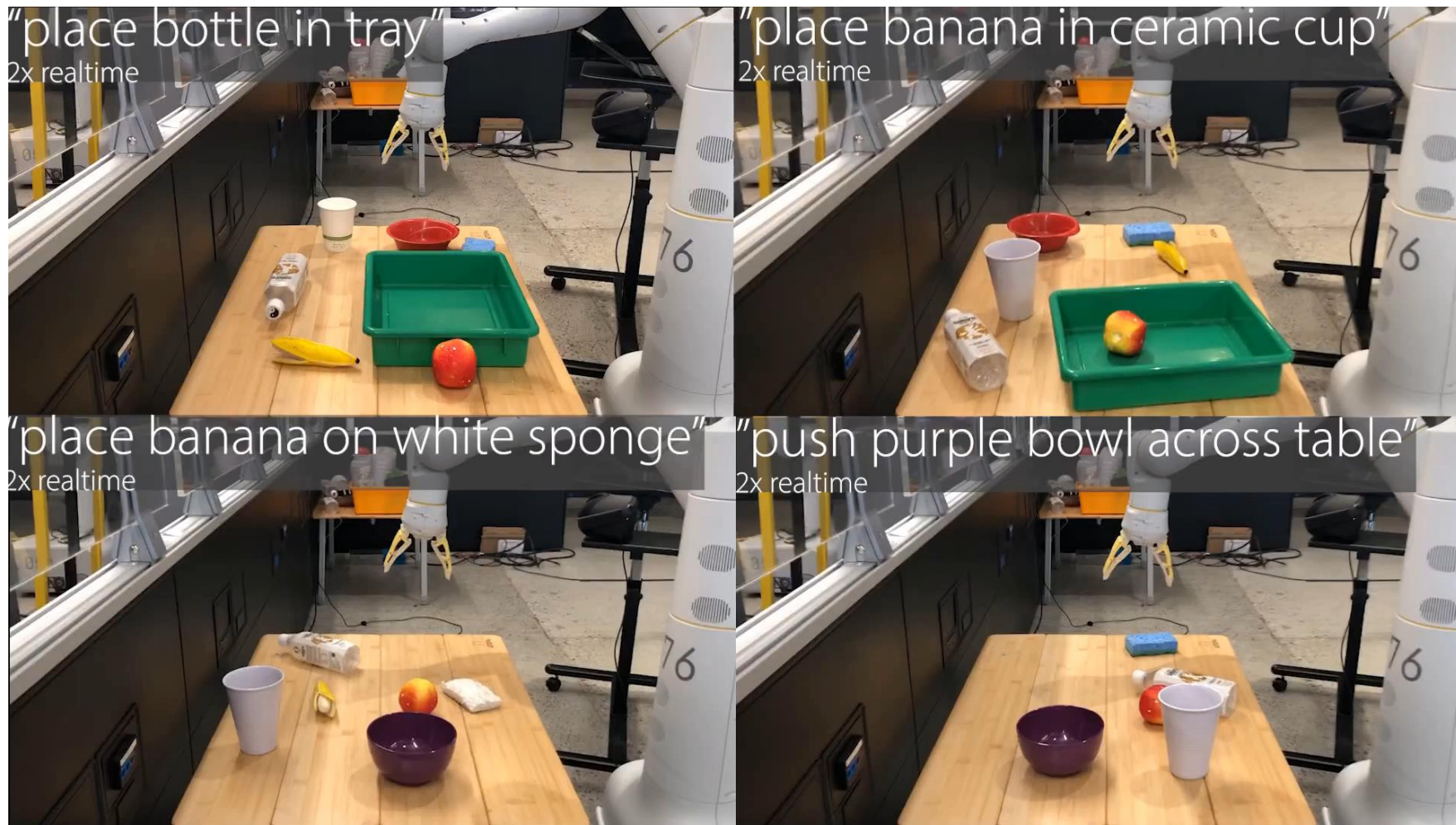
Goal



Single Play-LMP policy

Learning Latent Plans from Play
[Lynch et al. 2019]

Applications



BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning
[Jang et al. 2021]

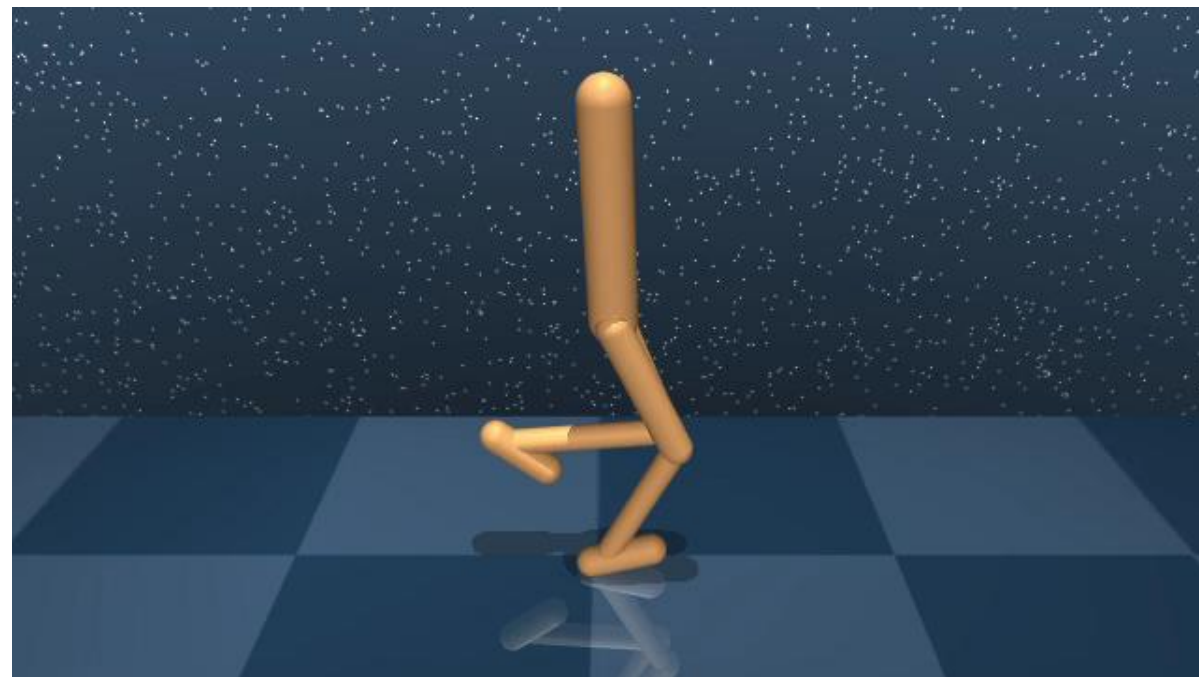
Summary

- Behavioral Cloning
- Drift
- Theoretical Analysis
- DAgger
- Applications

Assignment 1: Behavioral Cloning



Cheetah



Walker

Assignment 1: Behavioral Cloning

The screenshot shows the GitHub interface for the repository `xbpeng / rl_assignments`. The repository is public and has 3 stars, 2 watchers, and 3 forks. The main branch is `main`, with 1 branch and 0 tags. The repository contains a commit by Jason Peng titled "fixing potential laoding issueg" (note the typo) 19 hours ago, with 4 commits. The file list includes folders `a1`, `data`, `envs`, `learning`, `tools`, and `util`, and files `.gitignore`, `LICENSE`, and `README.md`. The right sidebar shows the "About" section with no description, the "Releases" section with no releases published, and the "Packages" section.

xbpeng / rl_assignments Public

Pin Unwatch 2 Fork 3 Star 3

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags Go to file Add file <> Code

Jason Peng fixing potential laoding issueg 55b171e 19 hours ago 4 commits

folder	a1	a1	2 days ago
folder	data	a1	2 days ago
folder	envs	a1	2 days ago
folder	learning	fixing potential laoding issueg	19 hours ago
folder	tools	a1	2 days ago
folder	util	a1	2 days ago
file	.gitignore	a1	2 days ago
file	LICENSE	a1	2 days ago
file	README.md	readme	2 days ago

About
No description, website, or topics provided.
Readme
BSD-3-Clause license
3 stars
2 watching
3 forks

Releases
No releases published
[Create a new release](#)

Packages

github.com/xbpeng/rl_assignments