Behavioral Cloning

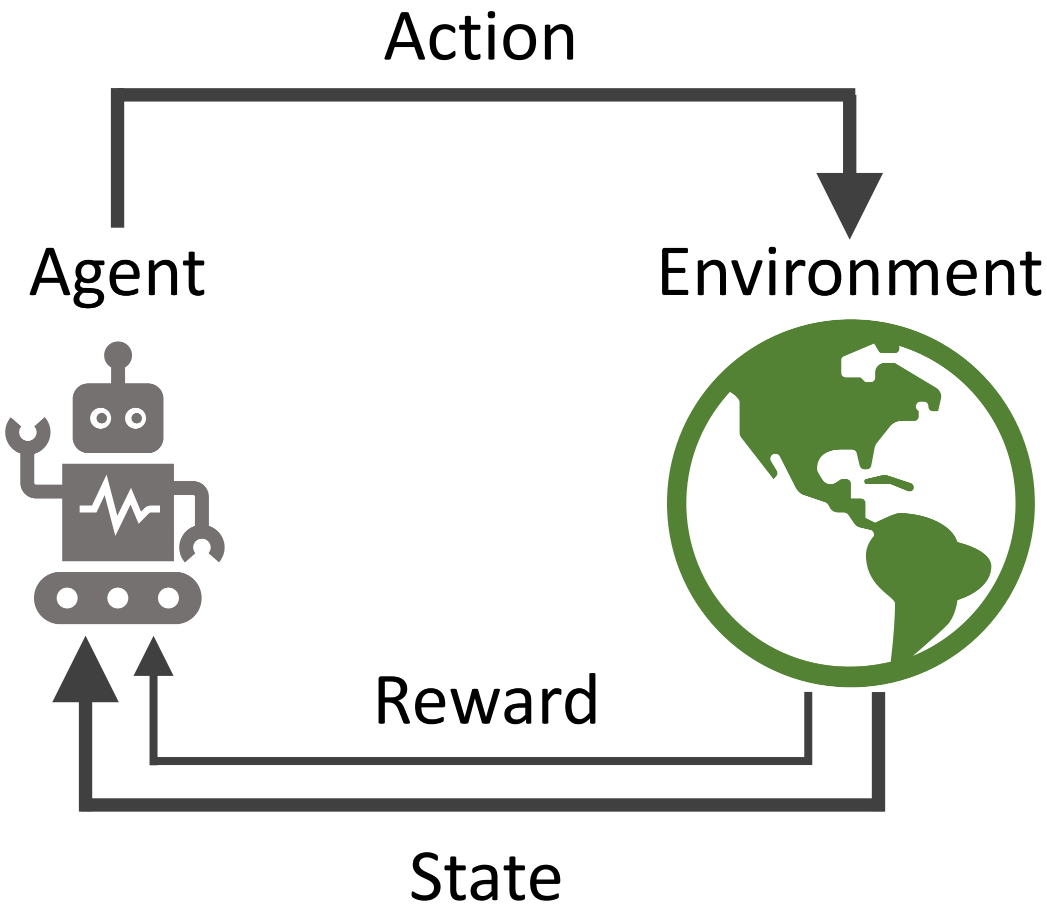
CMPT 729 G100

Jason Peng

Overview

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

Agent-Environment Interface



Action

Agent

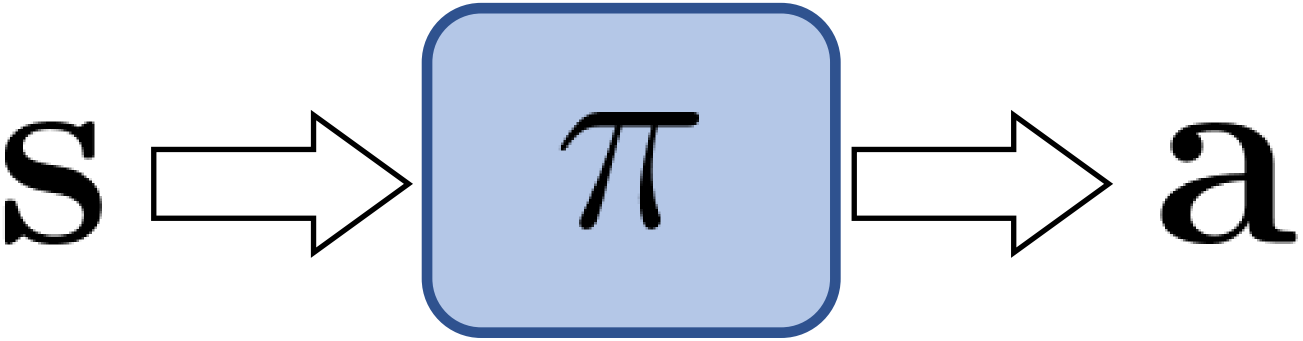
Environment

Reward

State

Policy

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



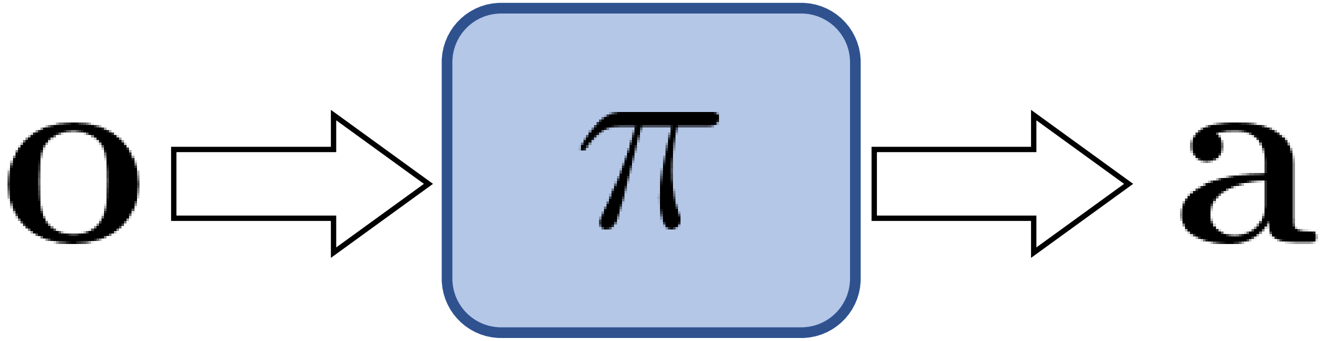
$\mathbf{s}$

π

$\mathbf{a}$

Policy

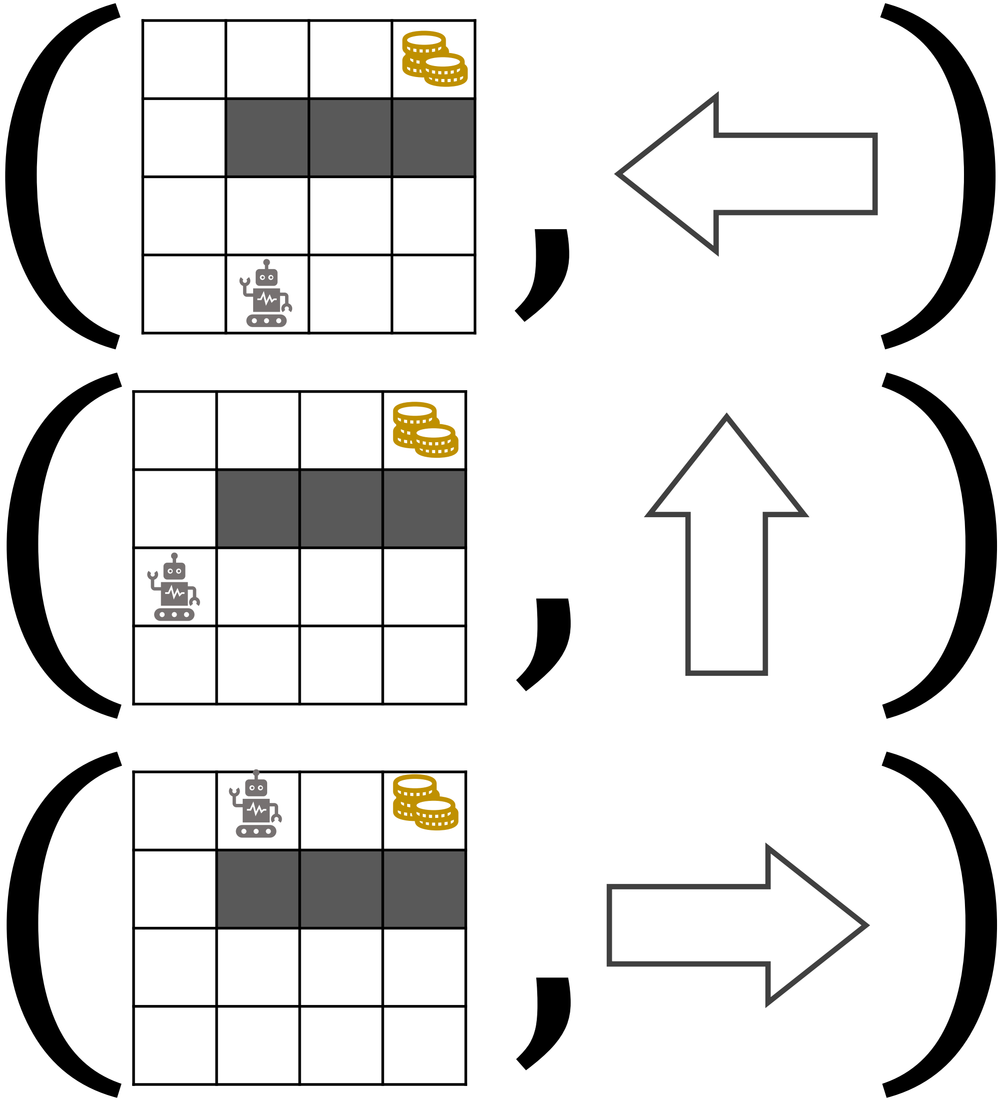
$$\pi(\mathbf{a}|\mathbf{o})$$



Supervised Learning

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

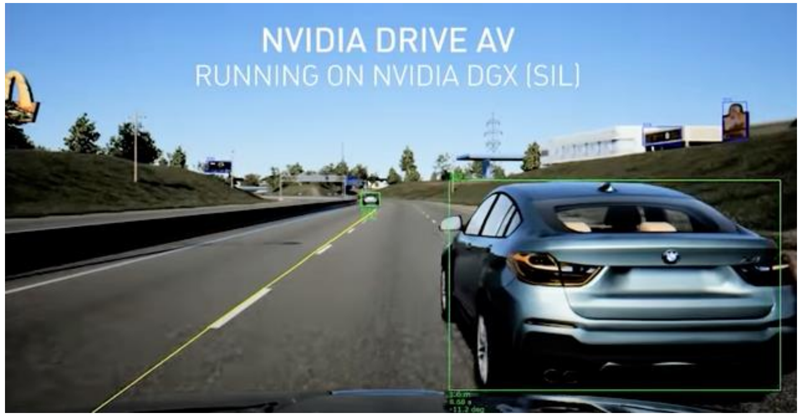
Dataset



Supervised Learning

$$\left\{ \left(\mathbf{o}_0, \mathbf{a}_0\right), \left(\mathbf{o}_1, \mathbf{a}_1\right), \ldots \right\}$$

Dataset



Nvidia Automotive Simulation
[NVIDIA]

$$\left(\mathbf{o}, \mathbf{a} \right)$$

Supervised Learning

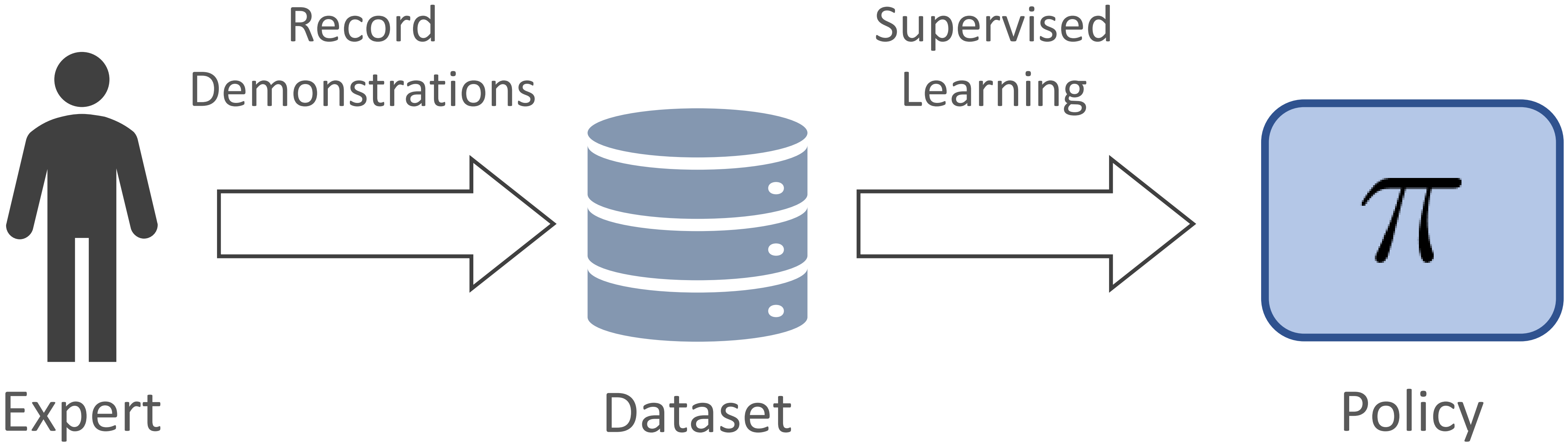
$$\{(\mathbf{o}_0, \mathbf{a}_0), (\mathbf{o}_1, \mathbf{a}_1), \ldots\}$$

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

Dataset

Behavioral Cloning

Behavioral Cloning



Record Demonstrations

Supervised Learning

π

Expert

Dataset

Policy

Behavioral Cloning

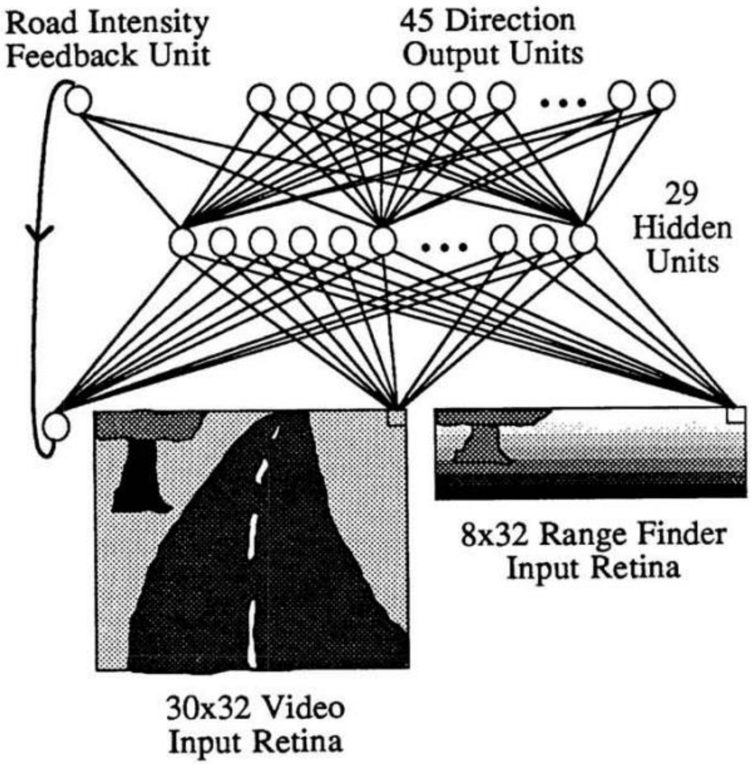
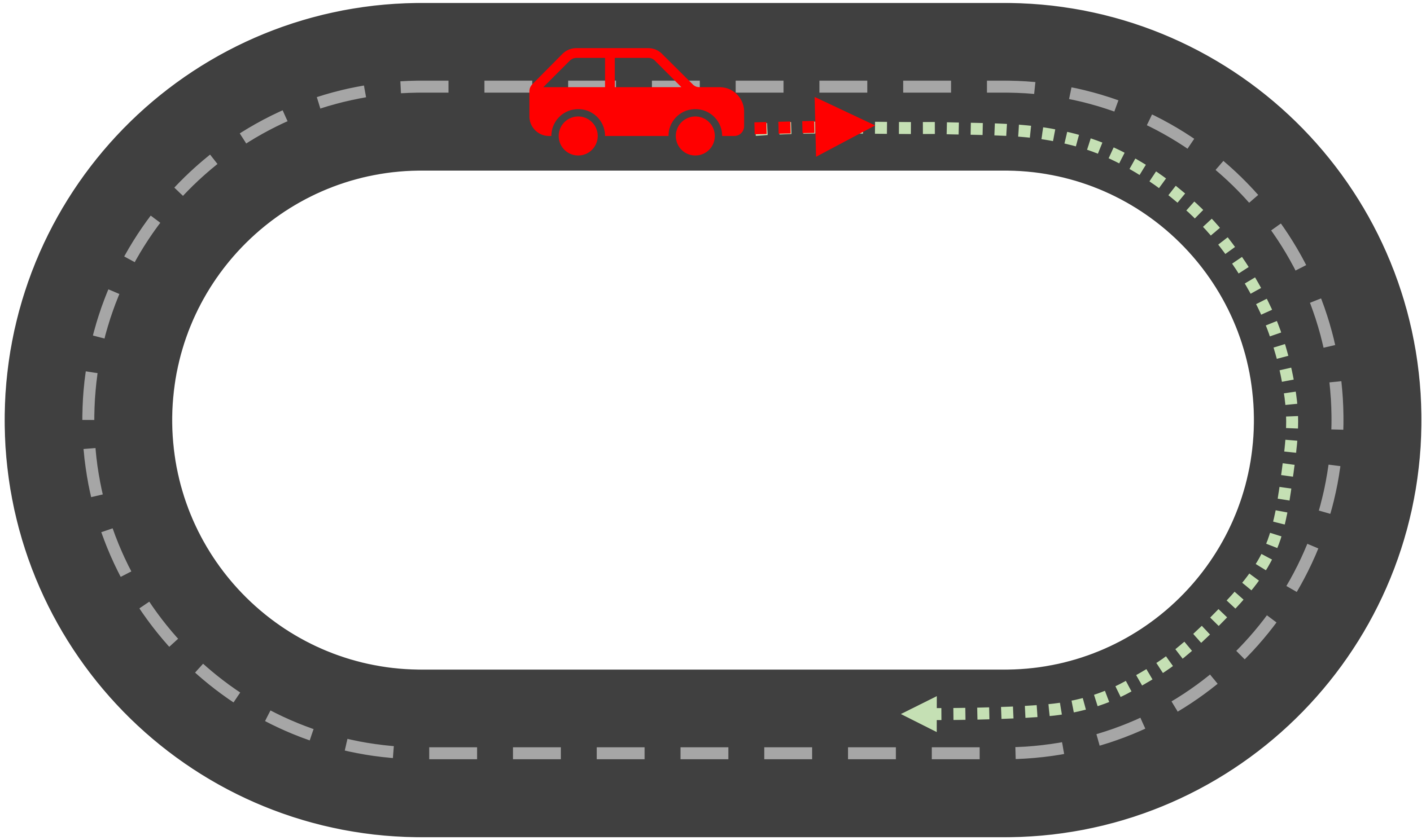


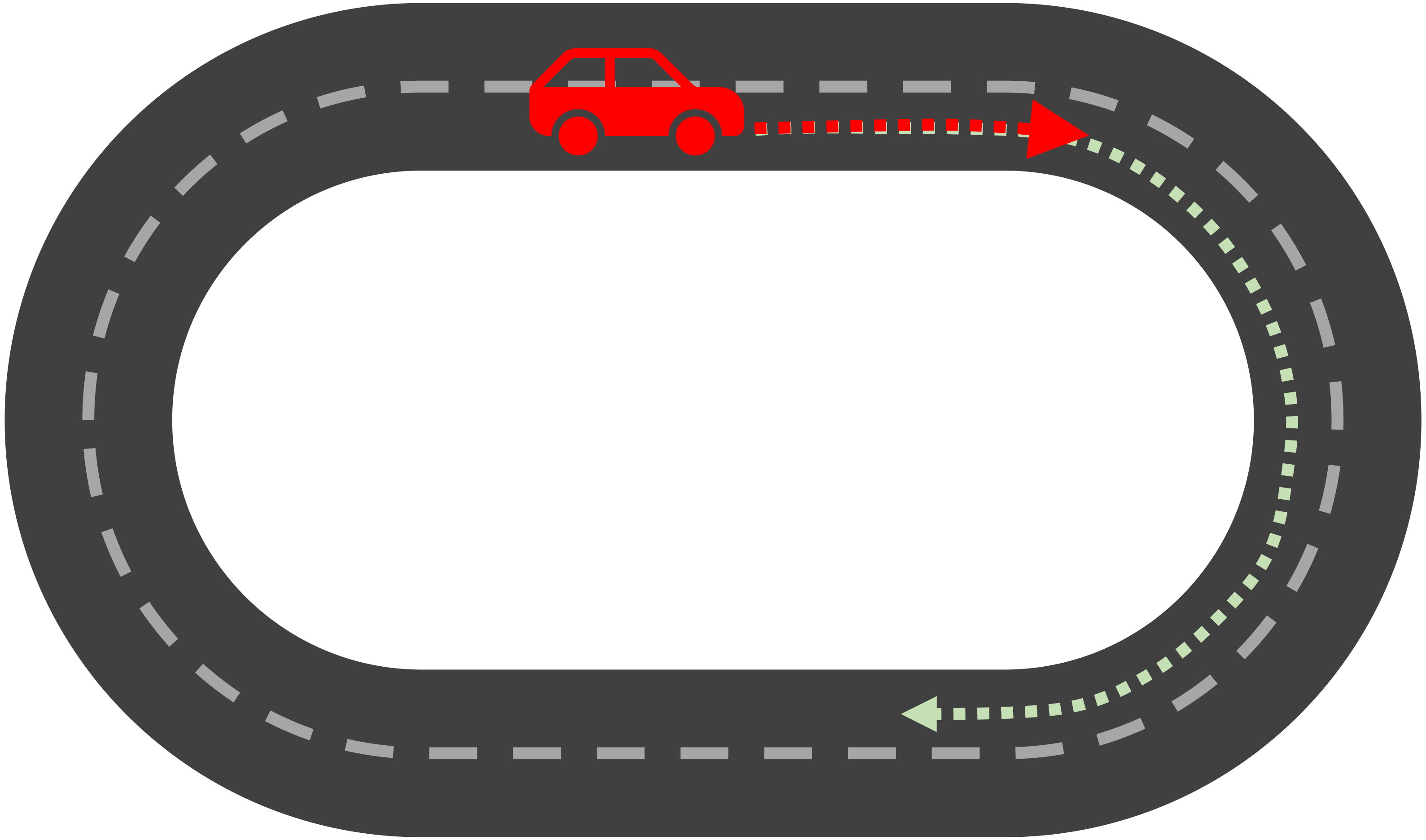
Figure 1: ALVINN Architecture

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

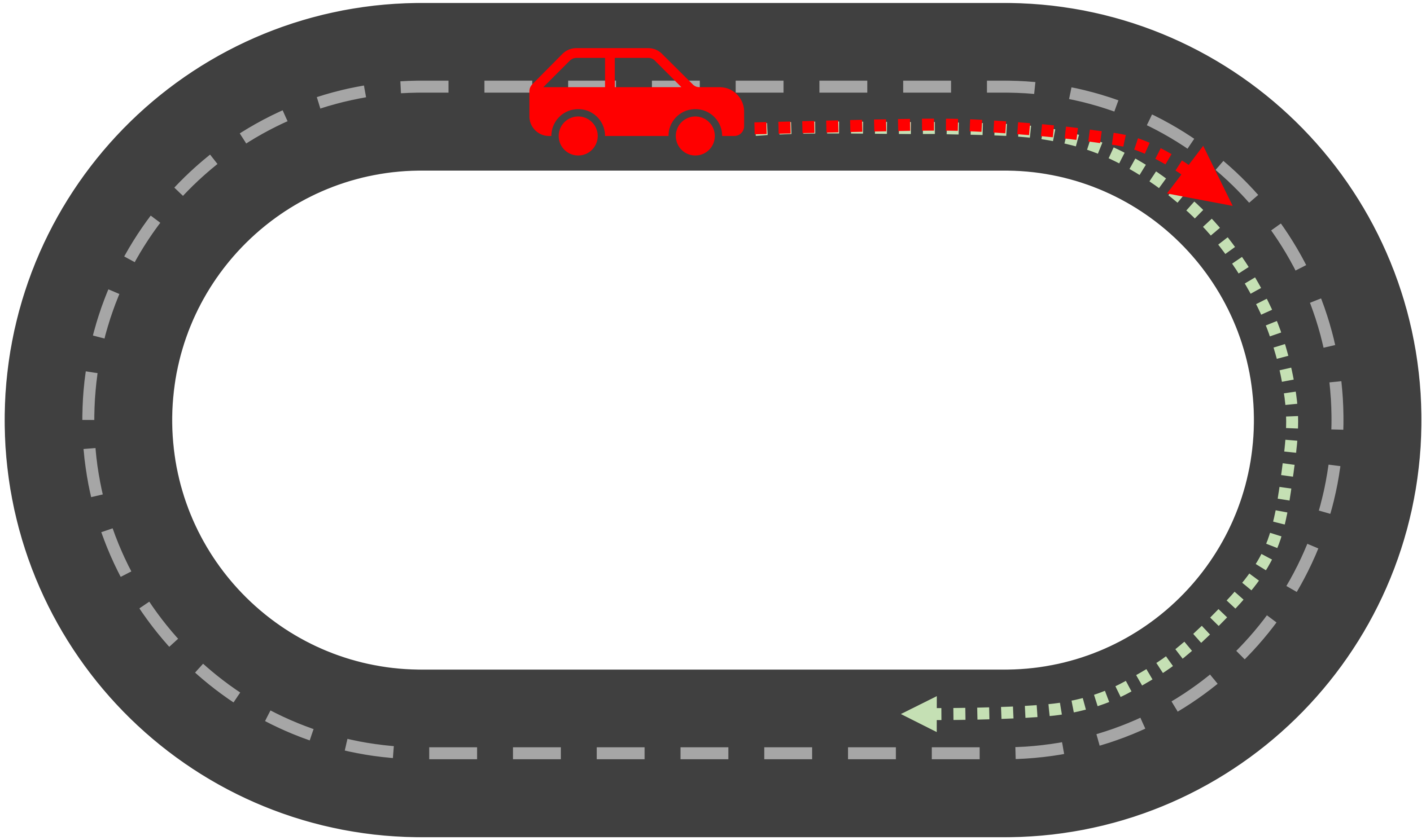
Does it work?



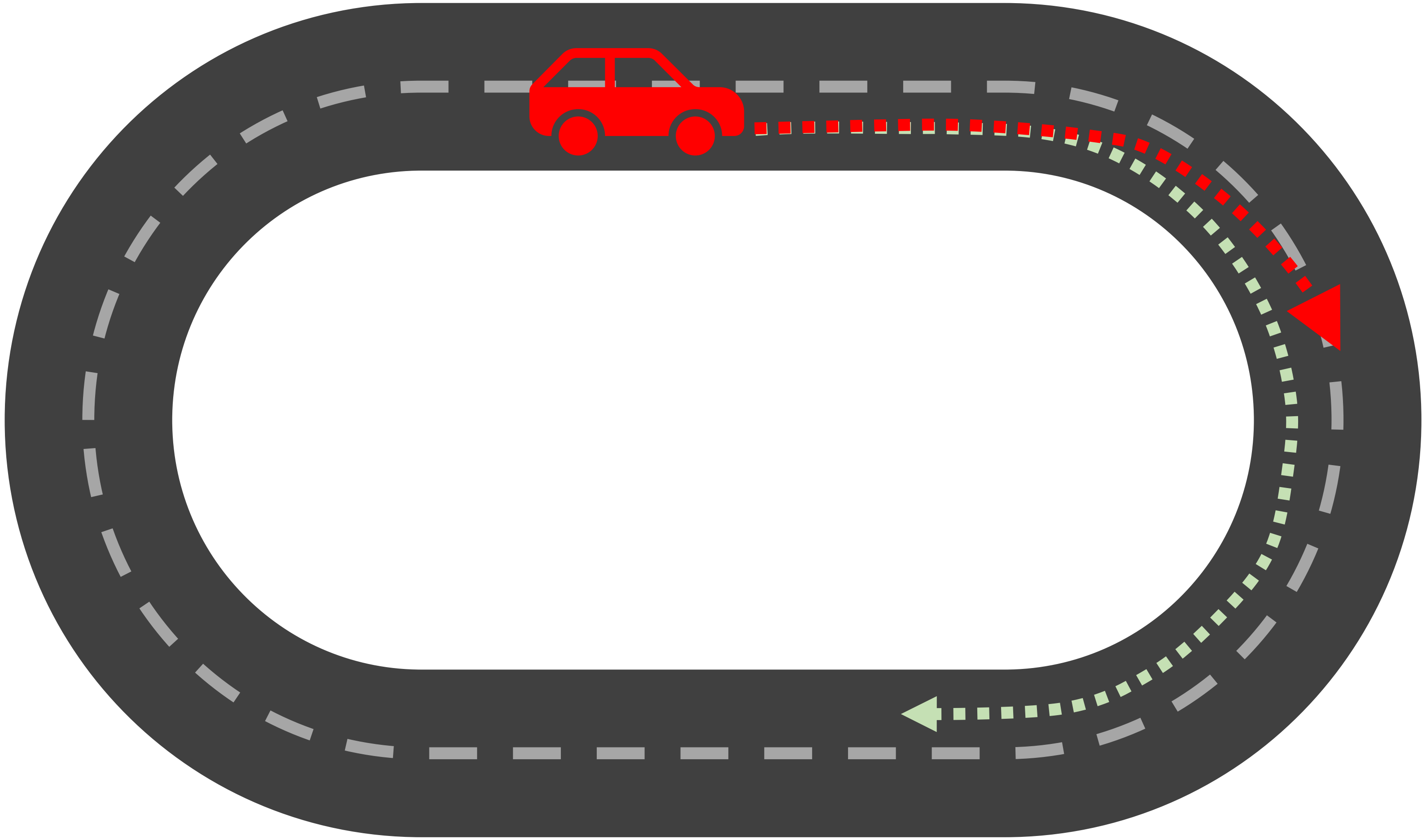
Does it work?



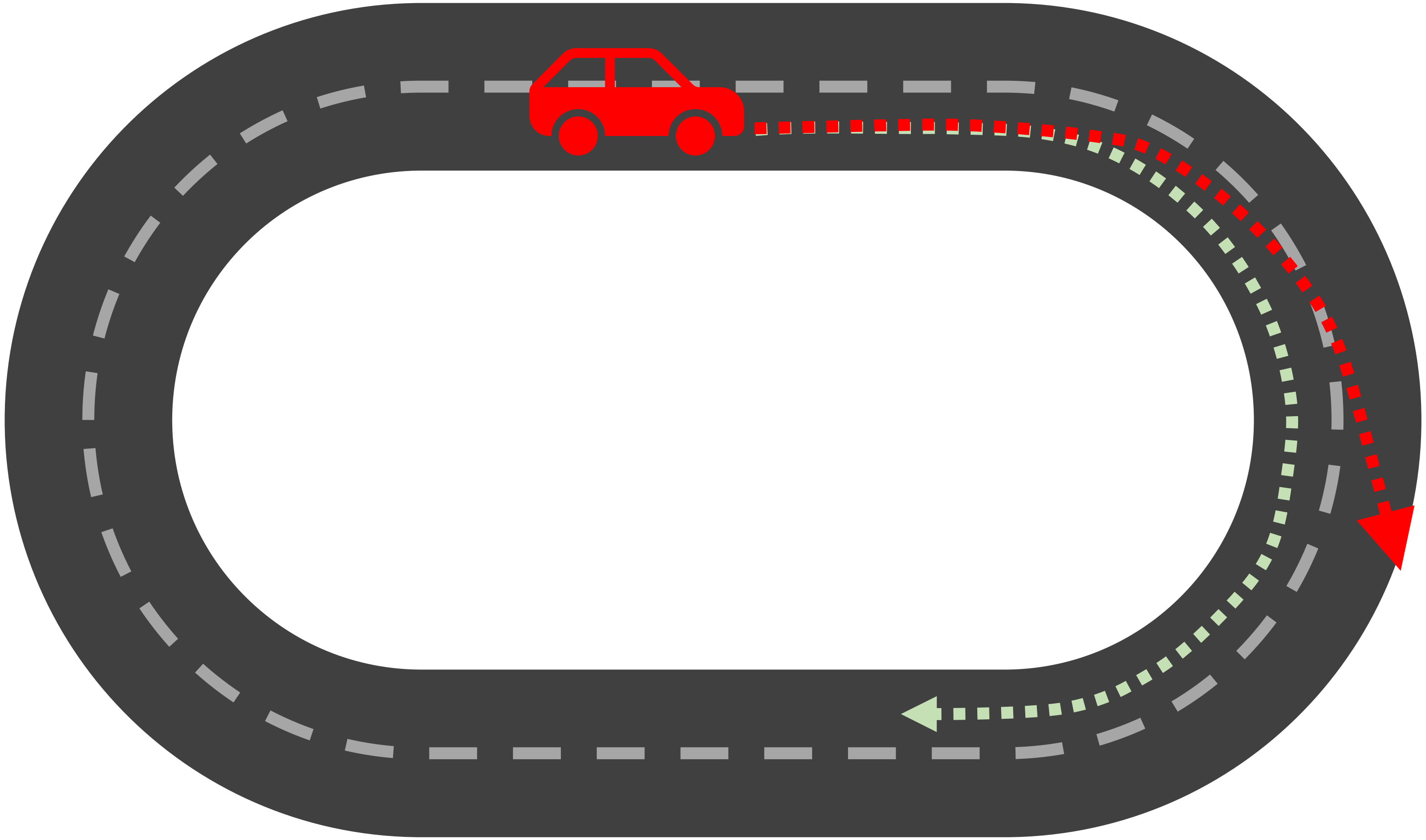
Does it work?



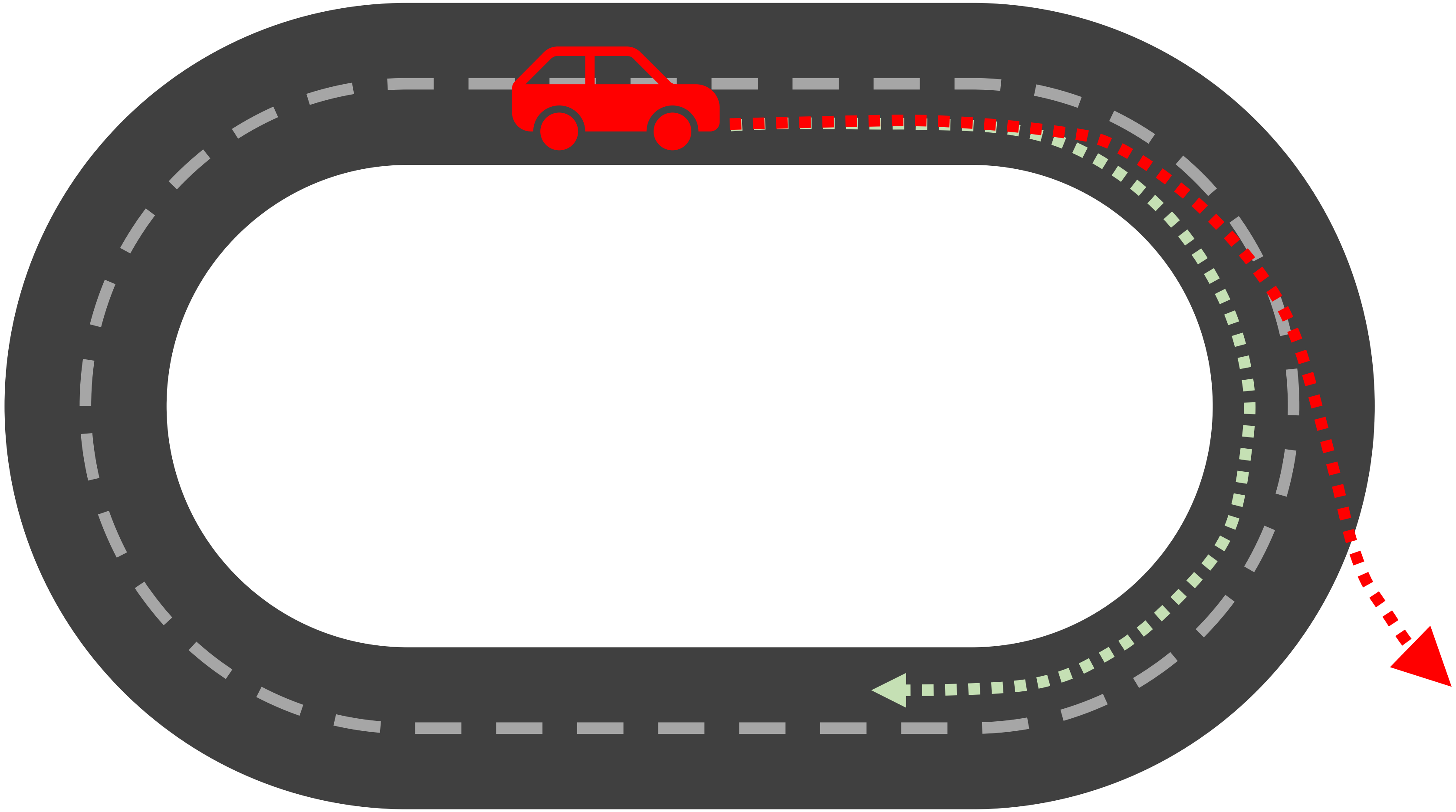
Does it work?



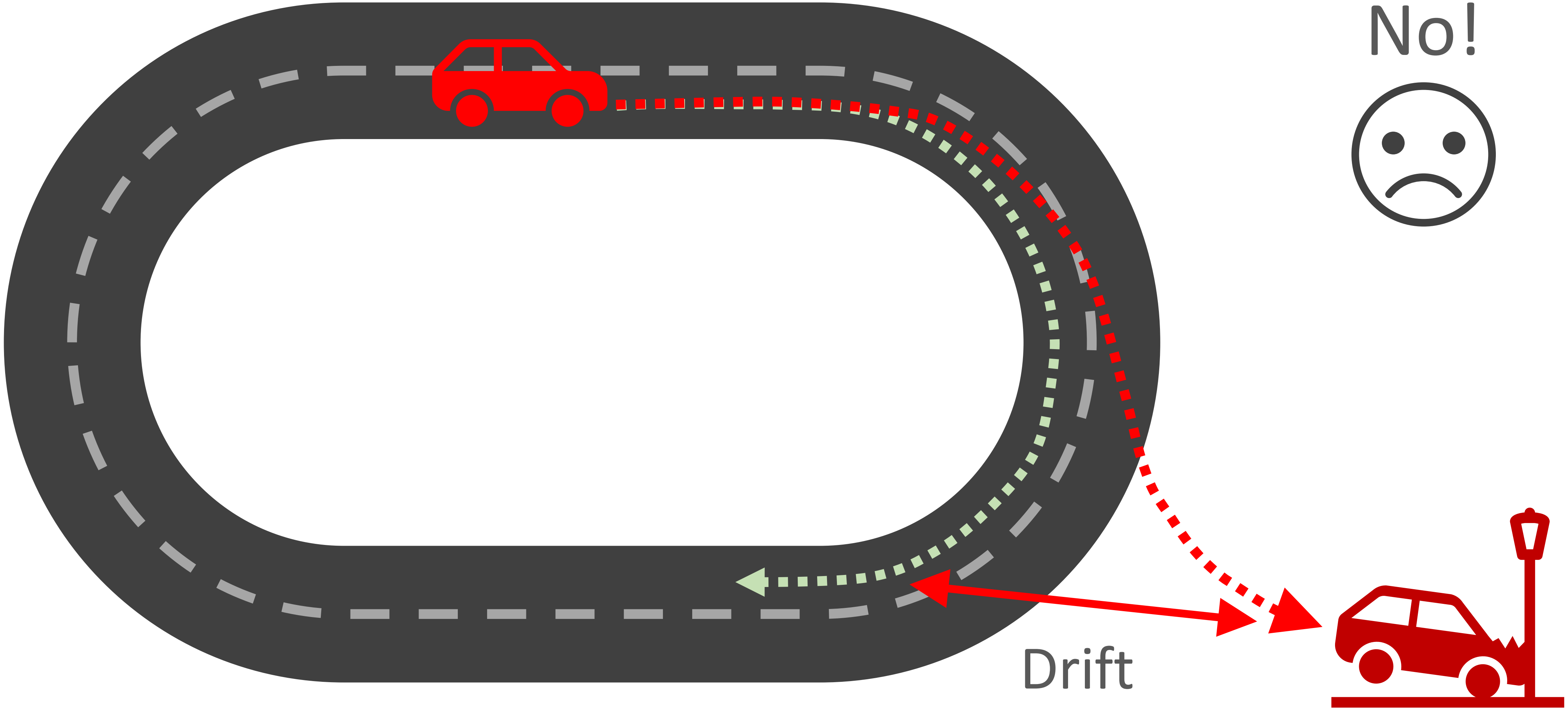
Does it work?



Does it work?



Does it work?



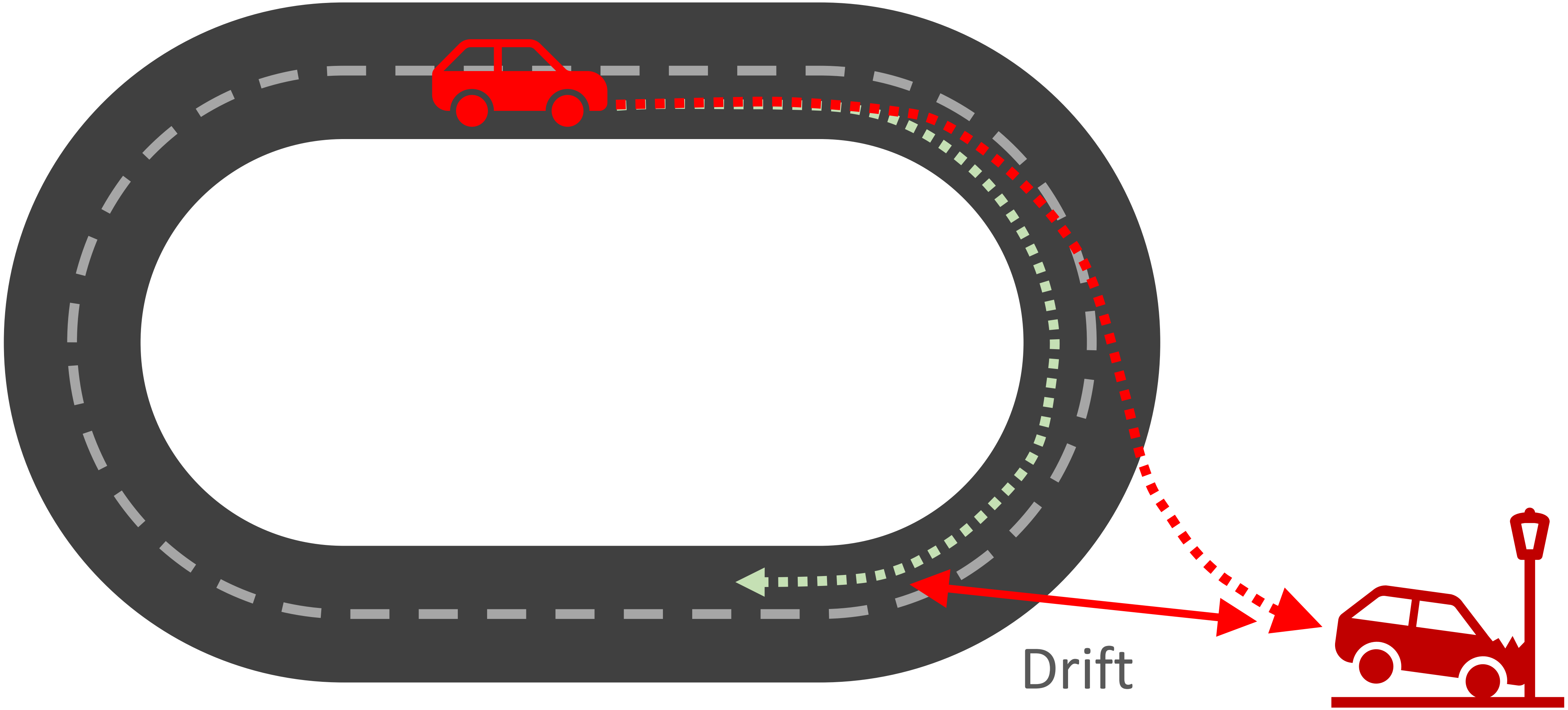
Drift

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

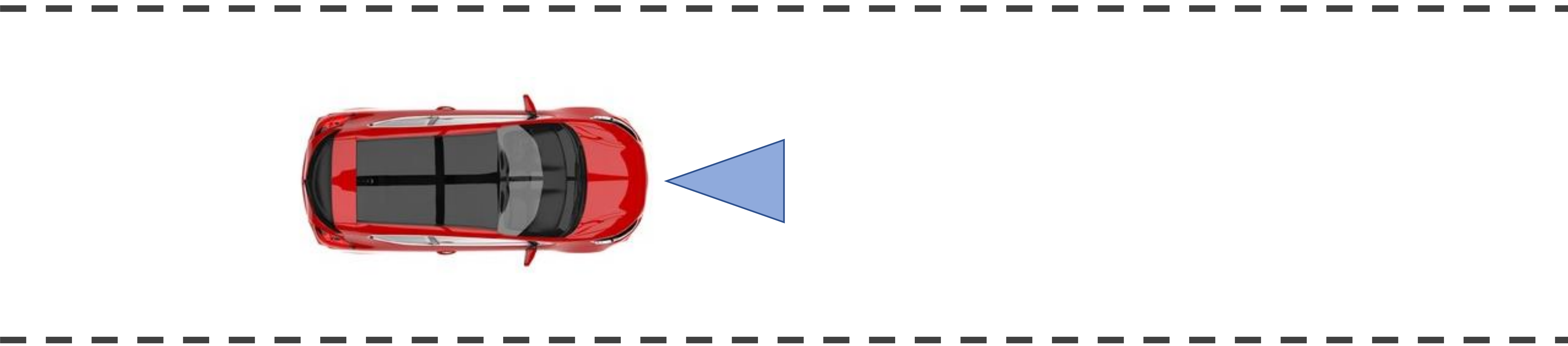
Drift

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

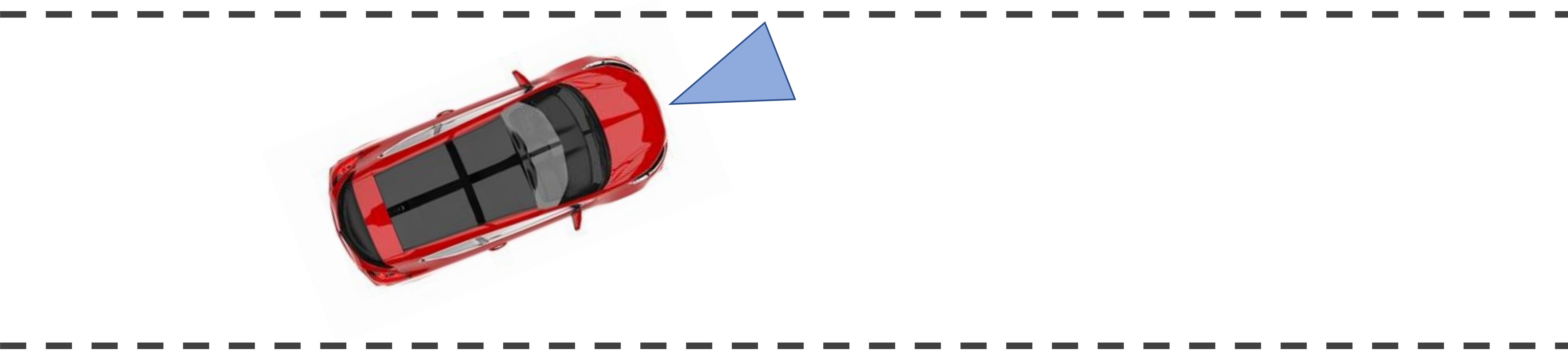
Feedback



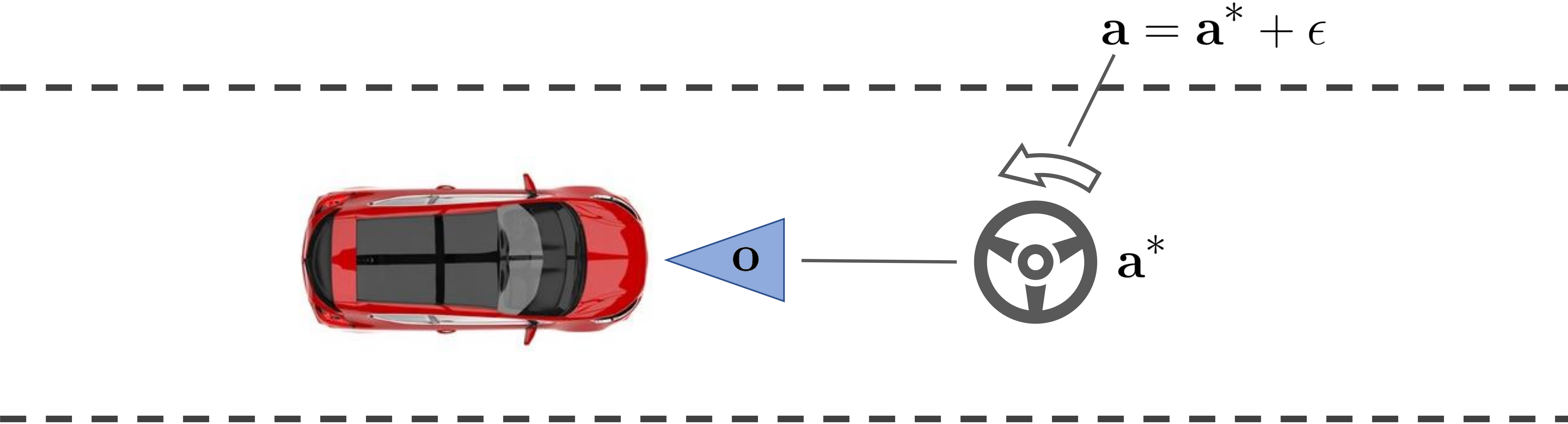
Feedback



Feedback

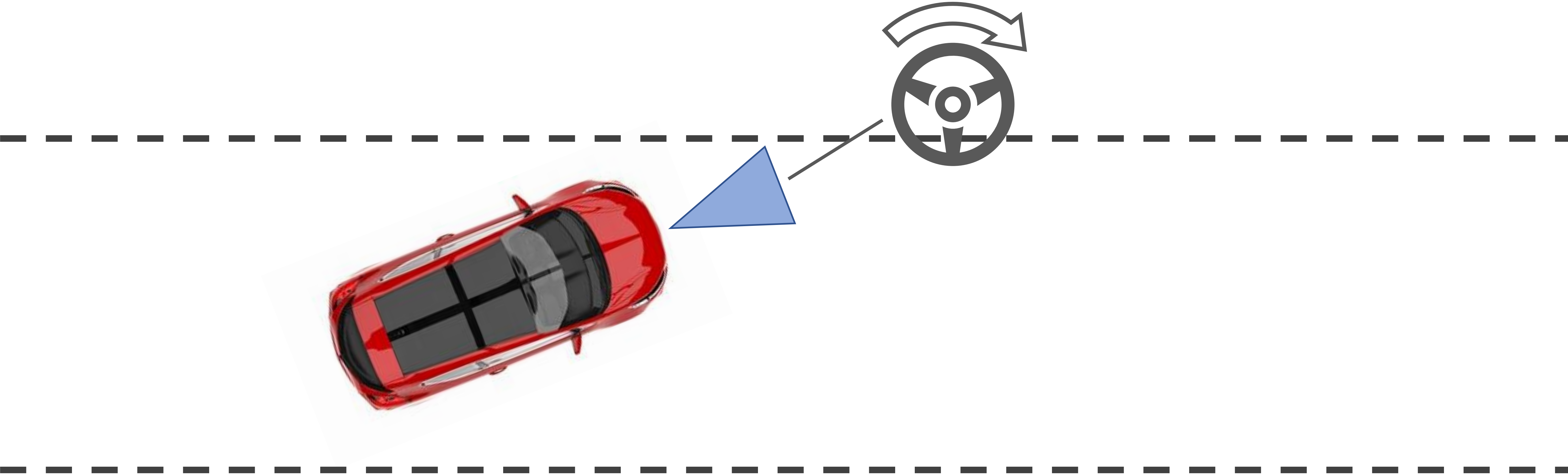


Noise Injection



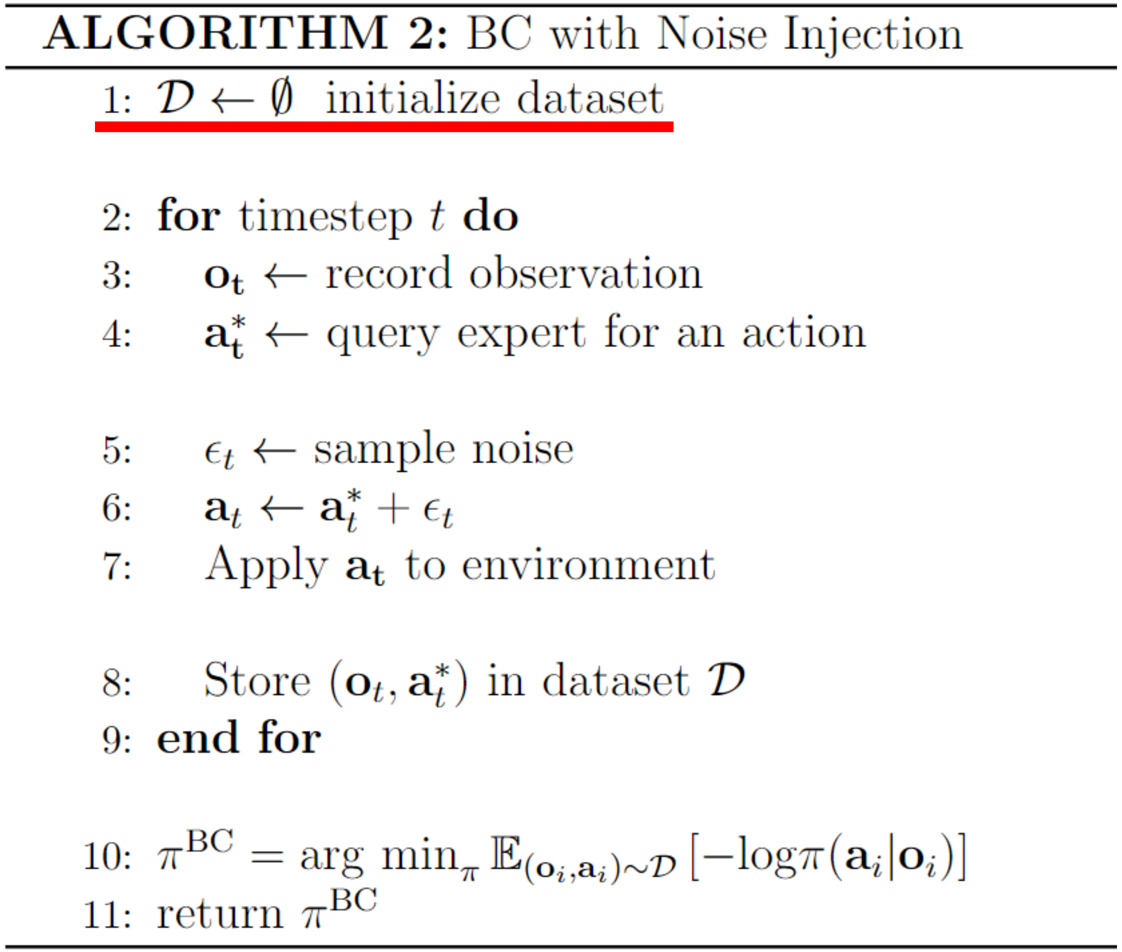
DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

Noise Injection



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

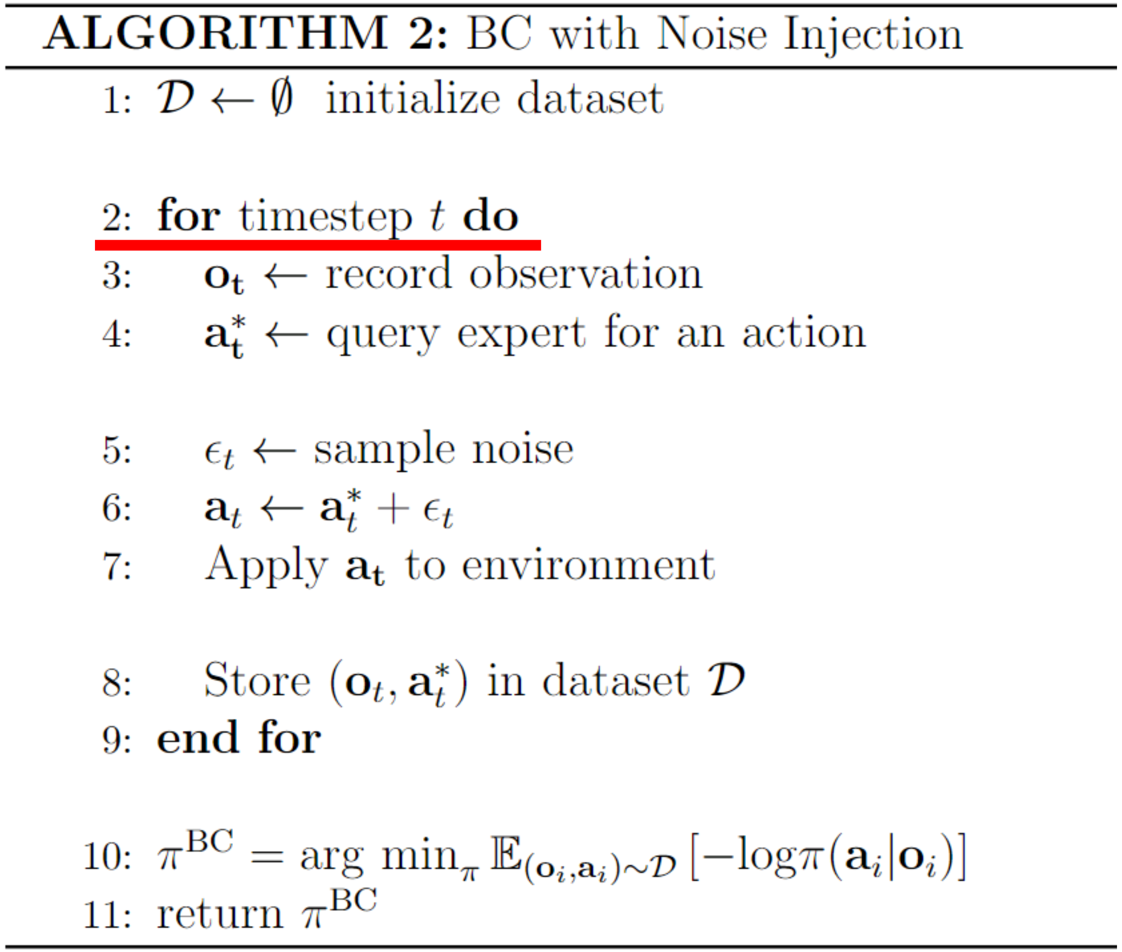
Noise Injection



```
ALGORITHM 2: BC with Noise Injection
$\mathcal{D} \leftarrow \emptyset$  initialize dataset

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection

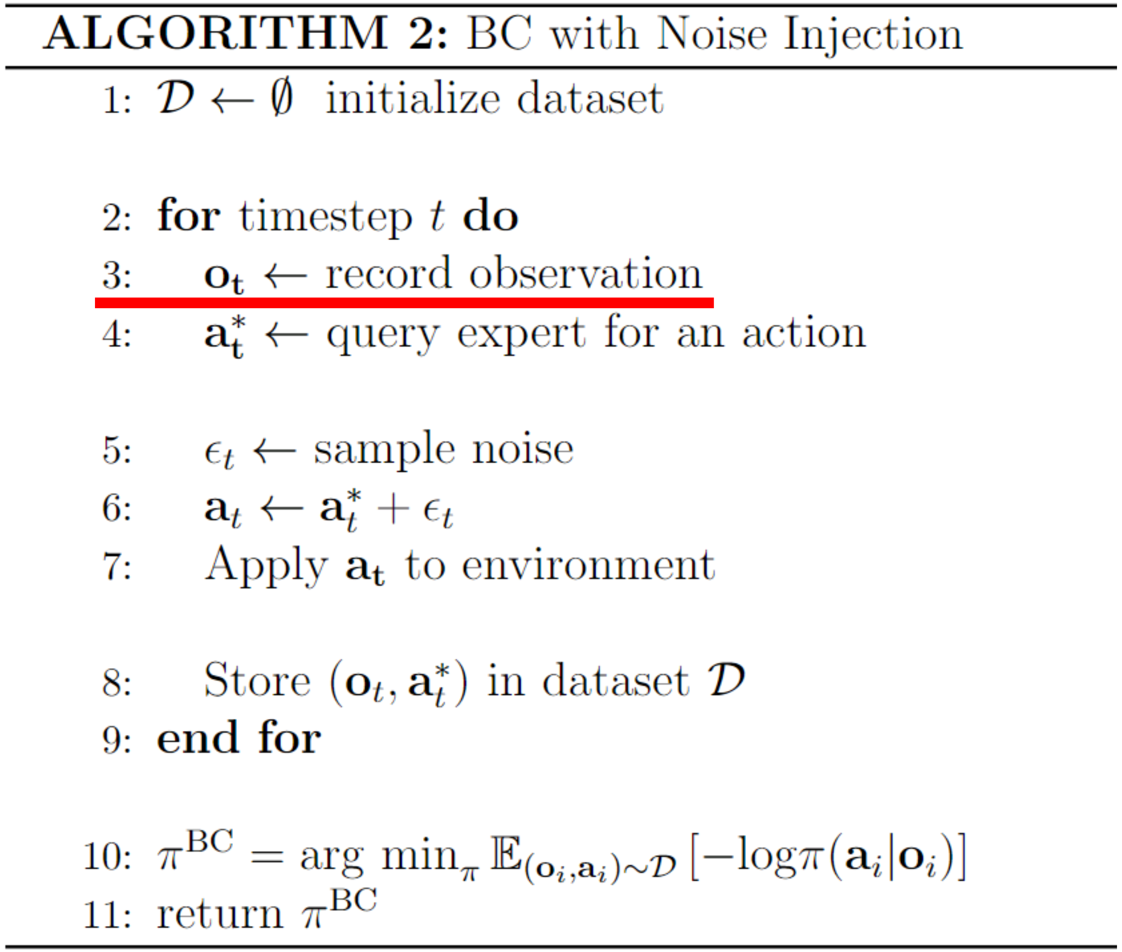


ALGORITHM 2: BC with Noise Injection

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

**for** timestep $t$ **do**
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection

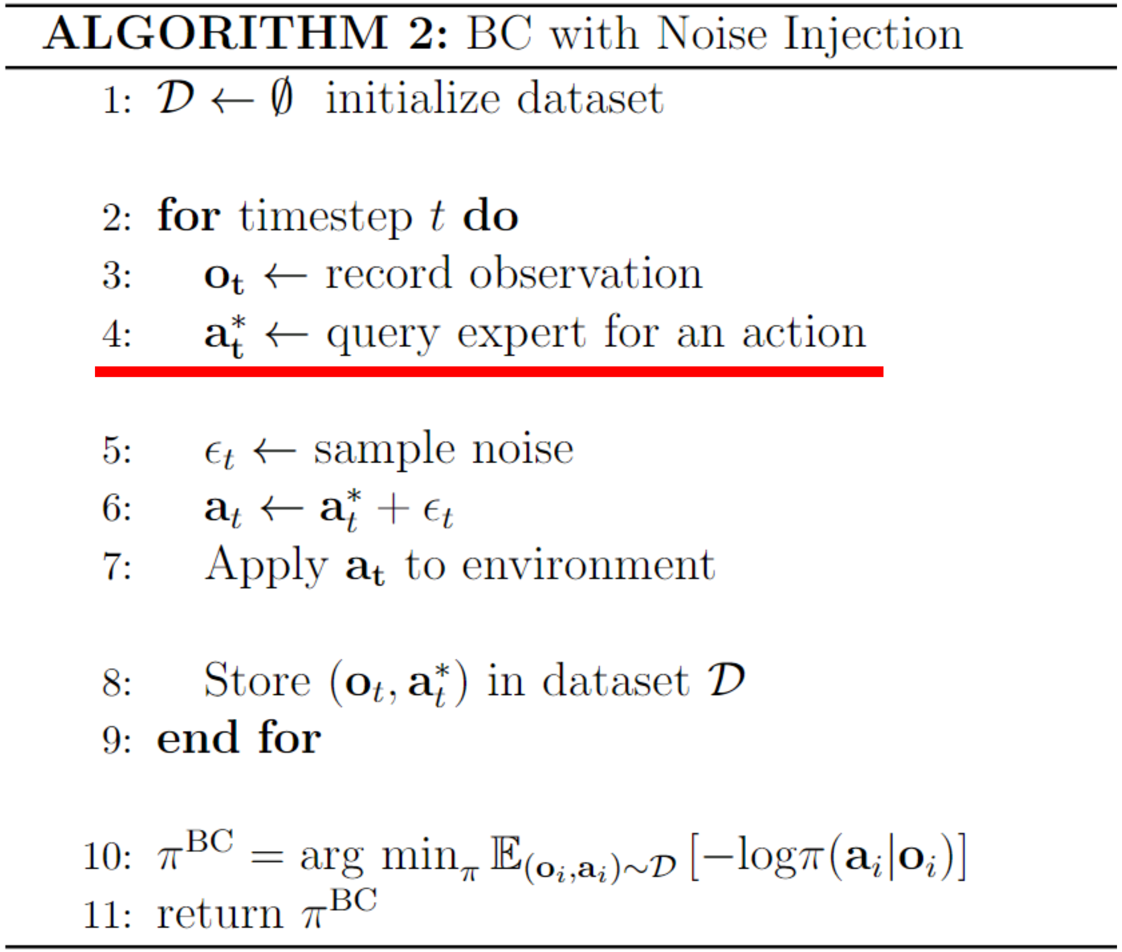


ALGORITHM 2: BC with Noise Injection

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

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection

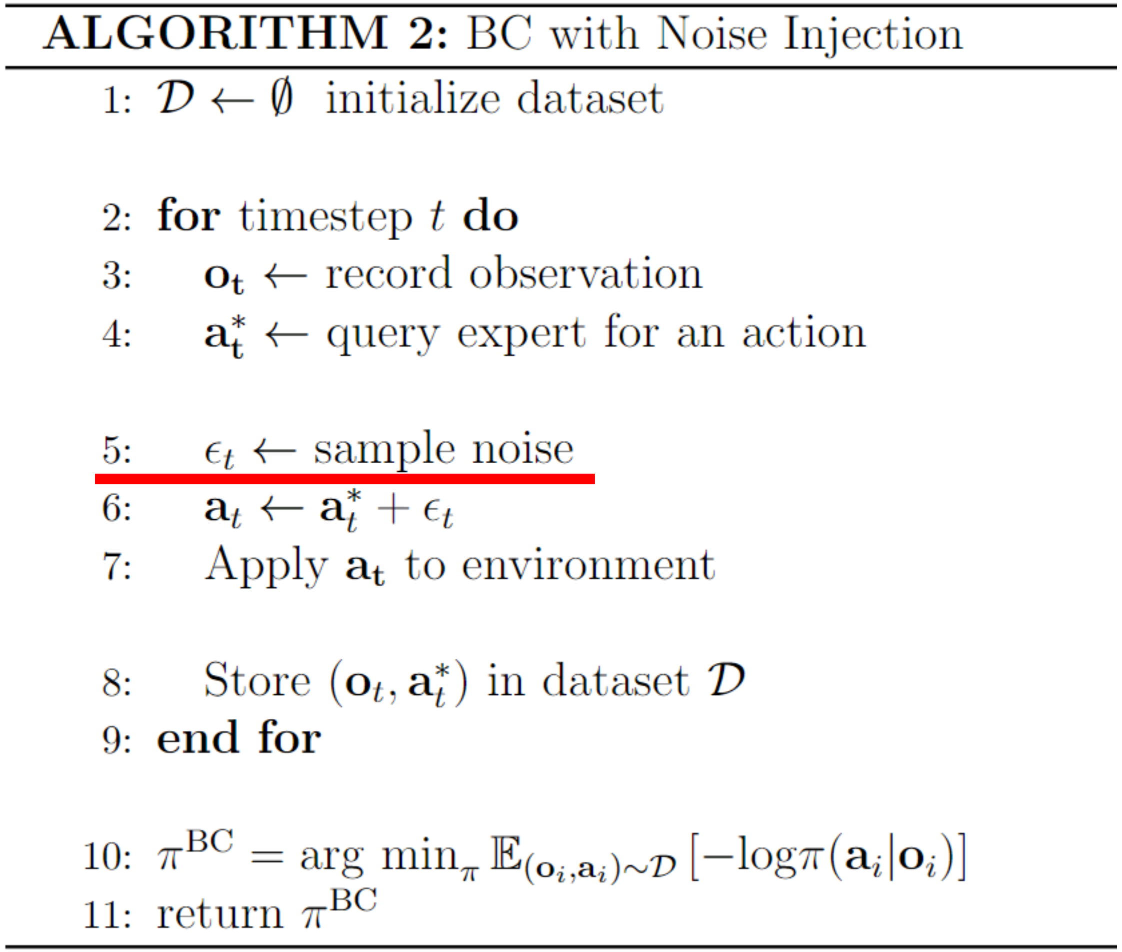


ALGORITHM 2: BC with Noise Injection

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

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection

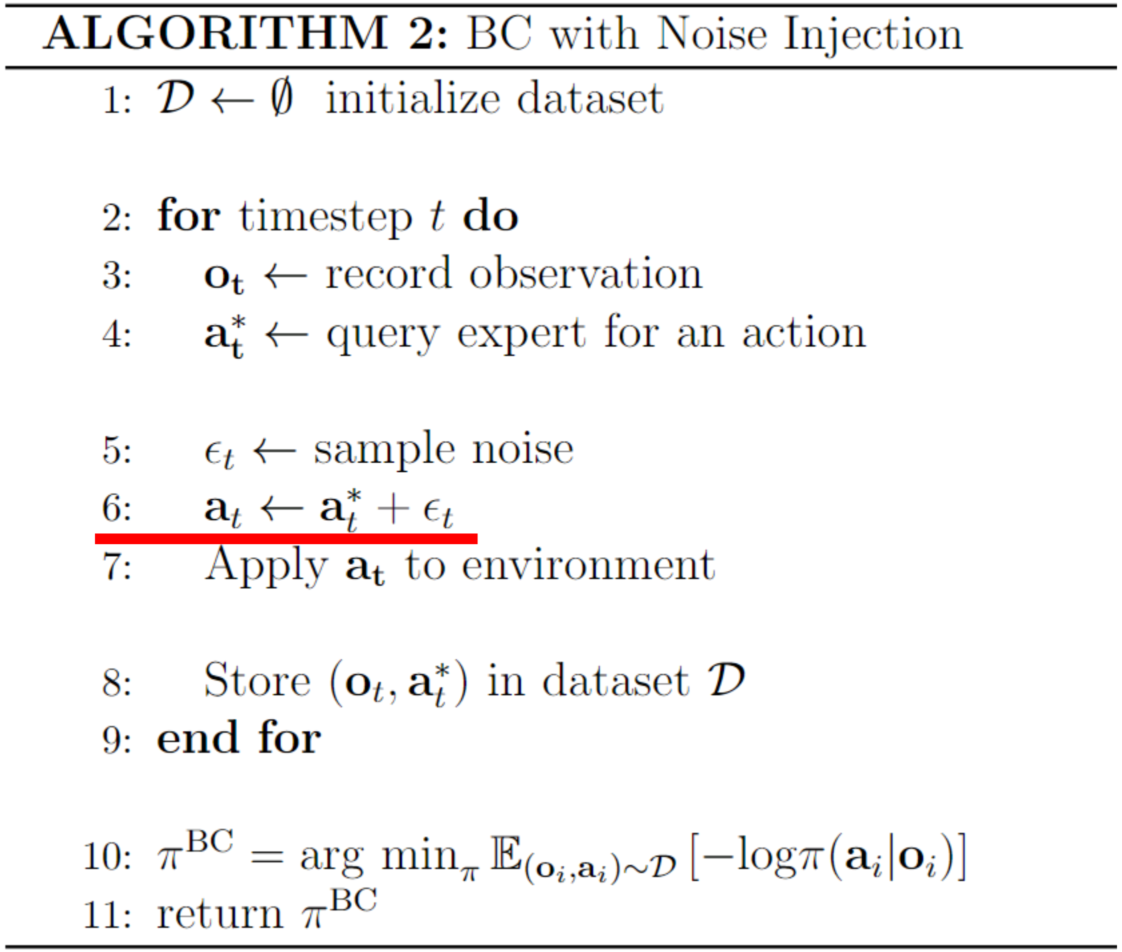


ALGORITHM 2: BC with Noise Injection

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

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection

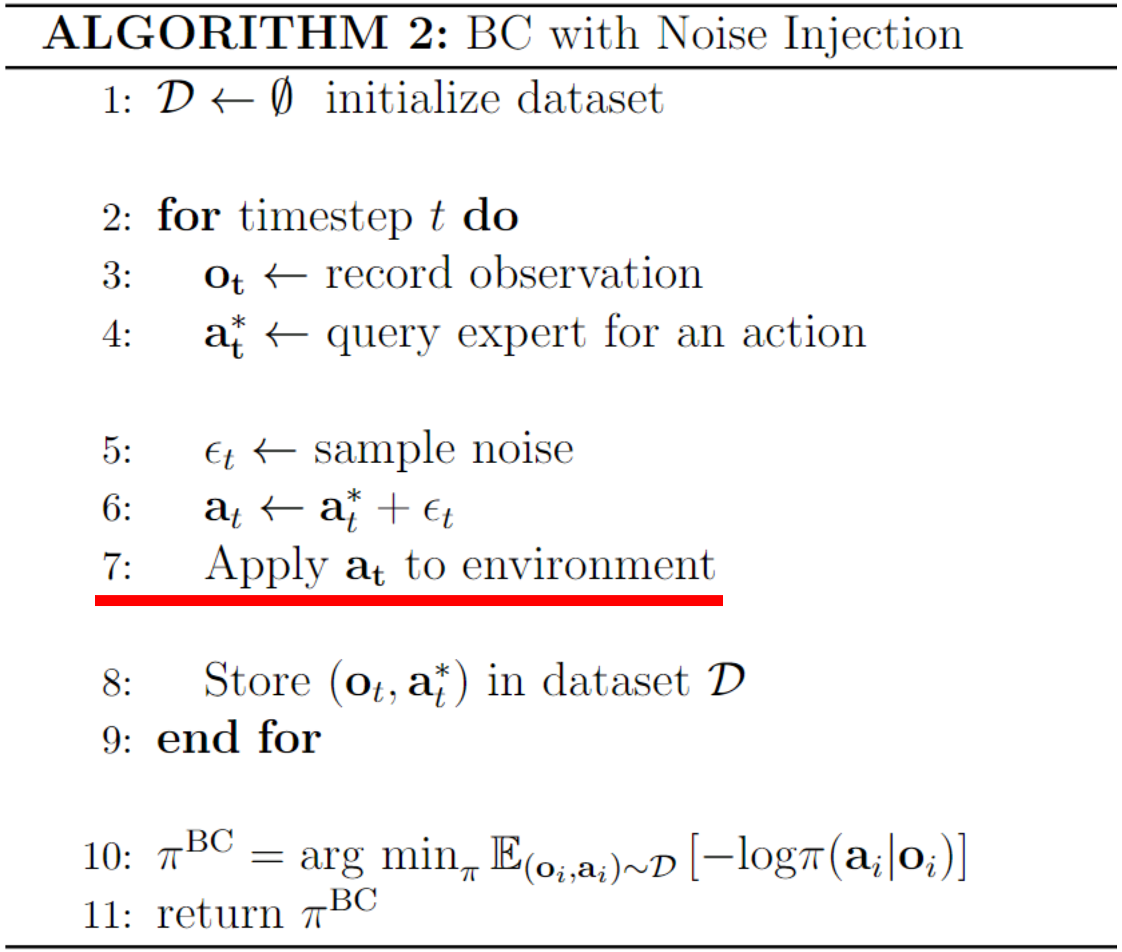


ALGORITHM 2: BC with Noise Injection

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

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

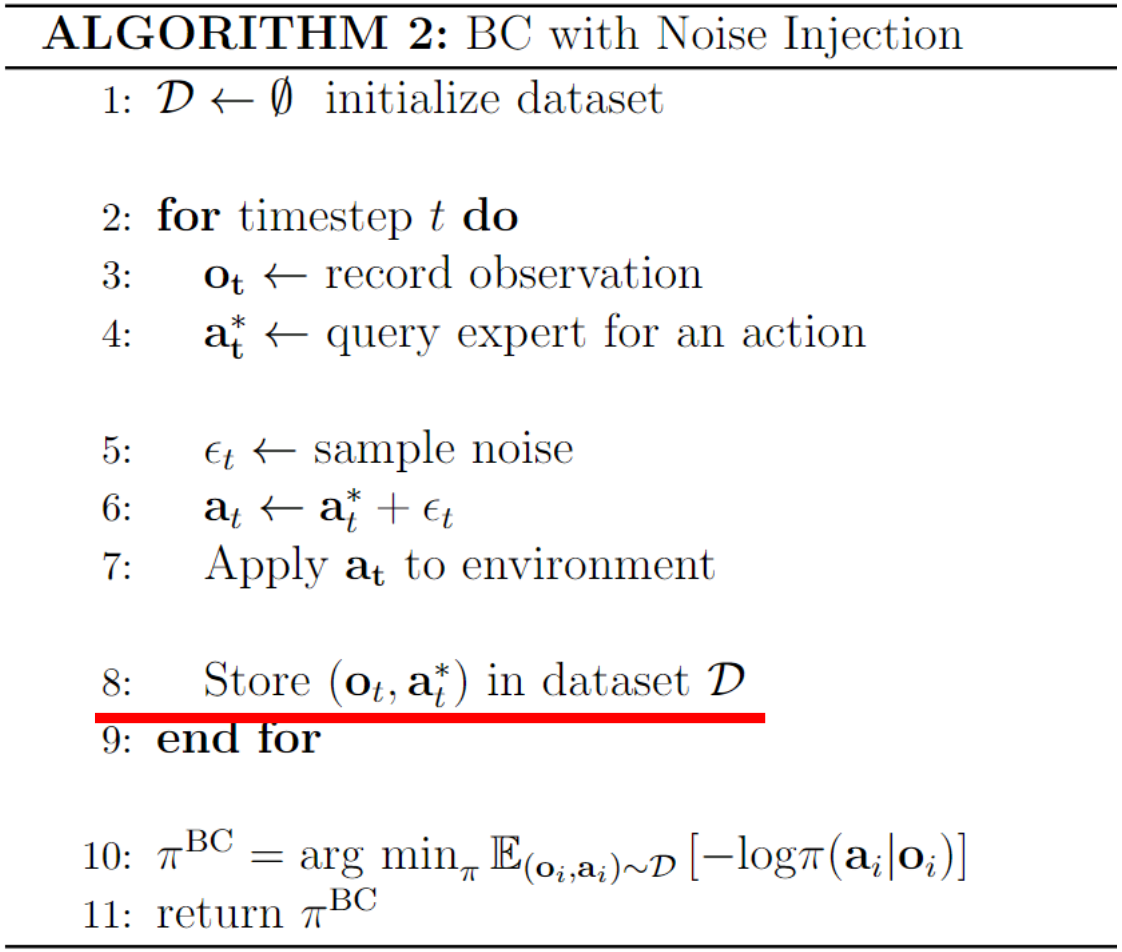
Noise Injection



```
ALGORITHM 2: BC with Noise Injection
$\mathcal{D} \leftarrow \emptyset$ initialize dataset

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection

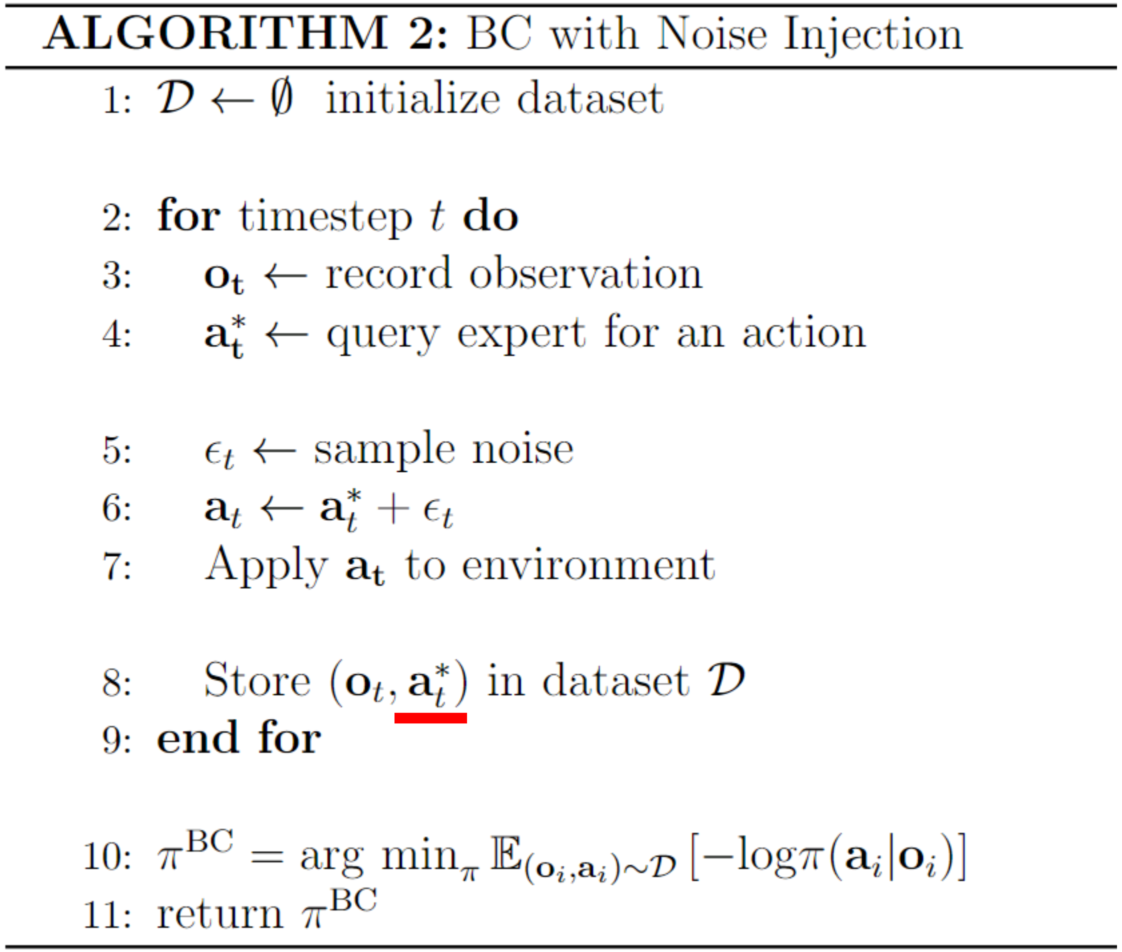


ALGORITHM 2: BC with Noise Injection

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

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection

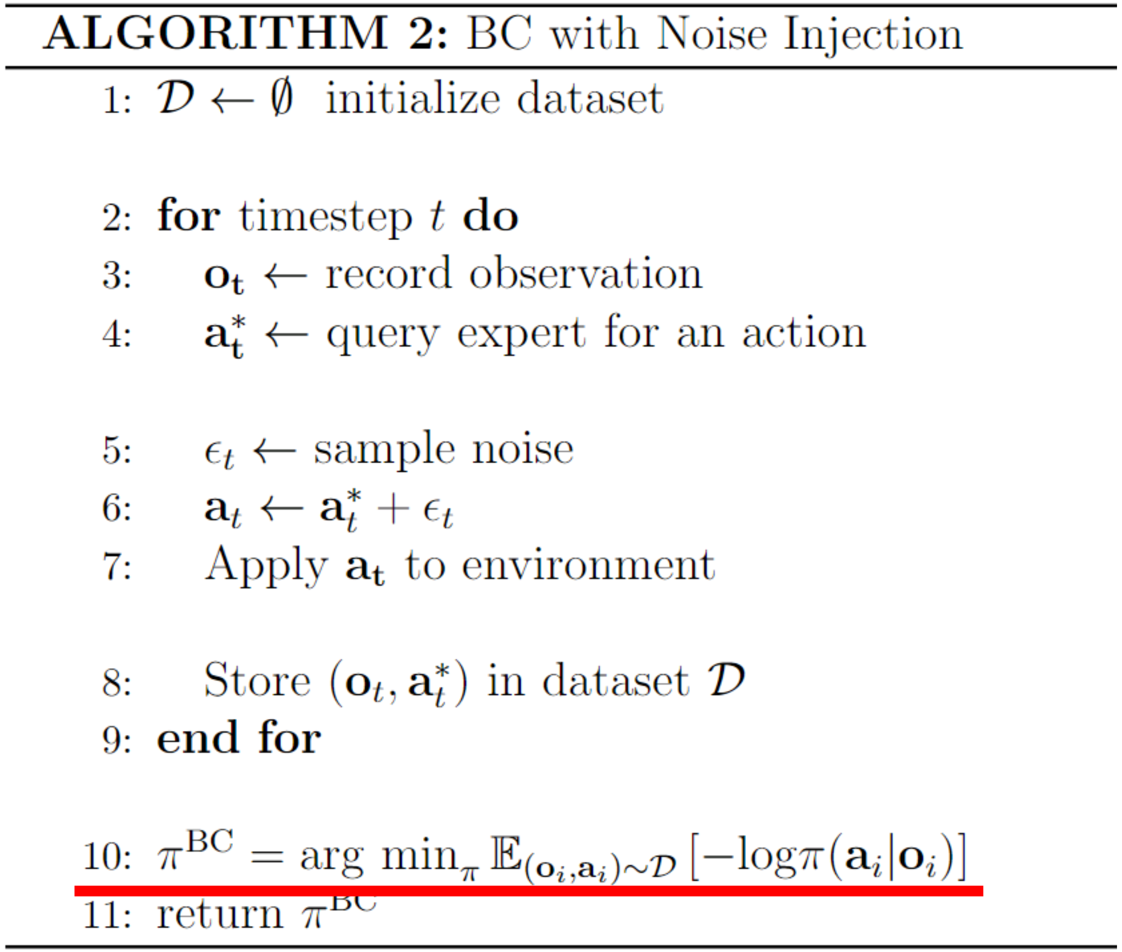


ALGORITHM 2: BC with Noise Injection

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

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection

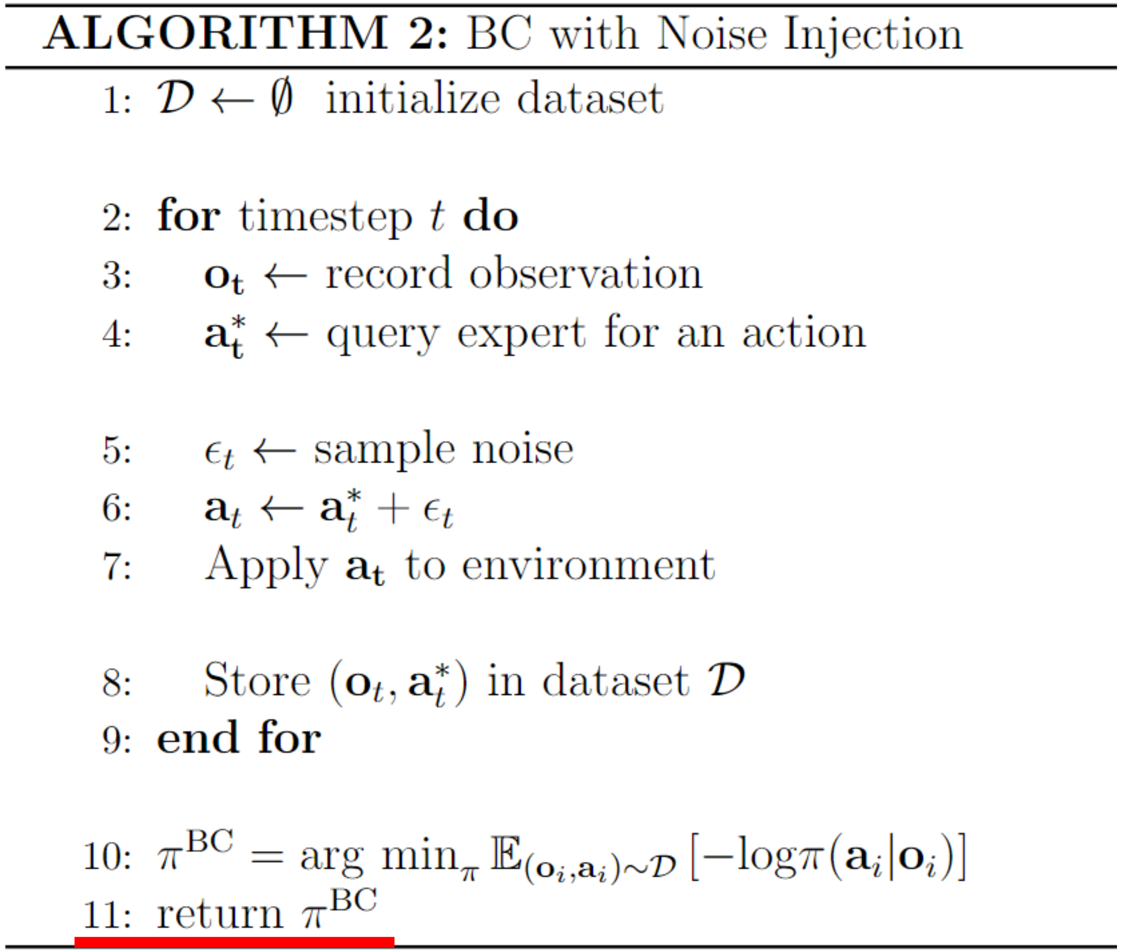


ALGORITHM 2: BC with Noise Injection

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

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection



ALGORITHM 2: BC with Noise Injection

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

for timestep $t$ do
    $\mathbf{o_t} \leftarrow$ record observation
    $\mathbf{a_t^*} \leftarrow$ query expert for an action

    $\epsilon_t \leftarrow$ sample noise
    $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
    Apply $\mathbf{a_t}$ to environment

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

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

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

Noise Injection

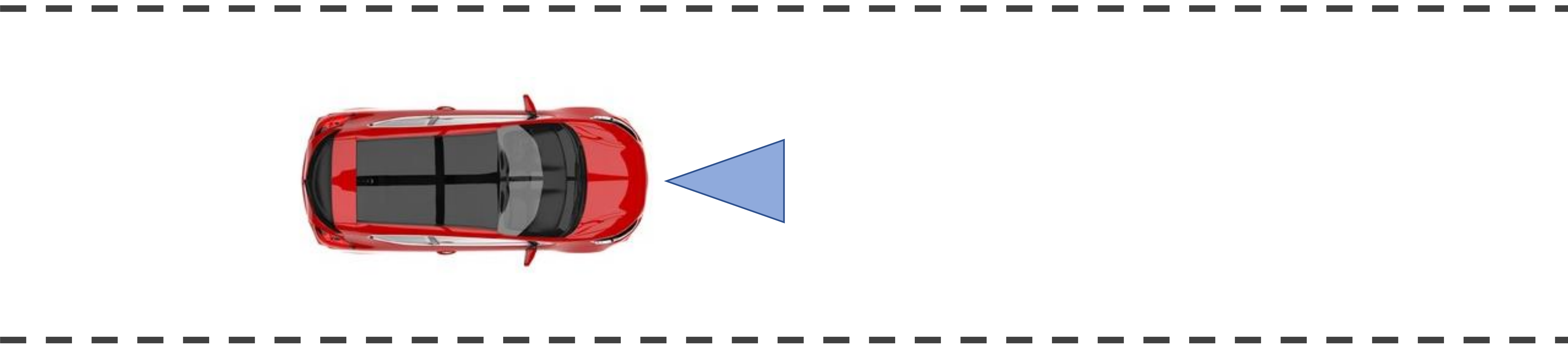
✓ Simple method to get corrective feedback

✓ Can work well in practice

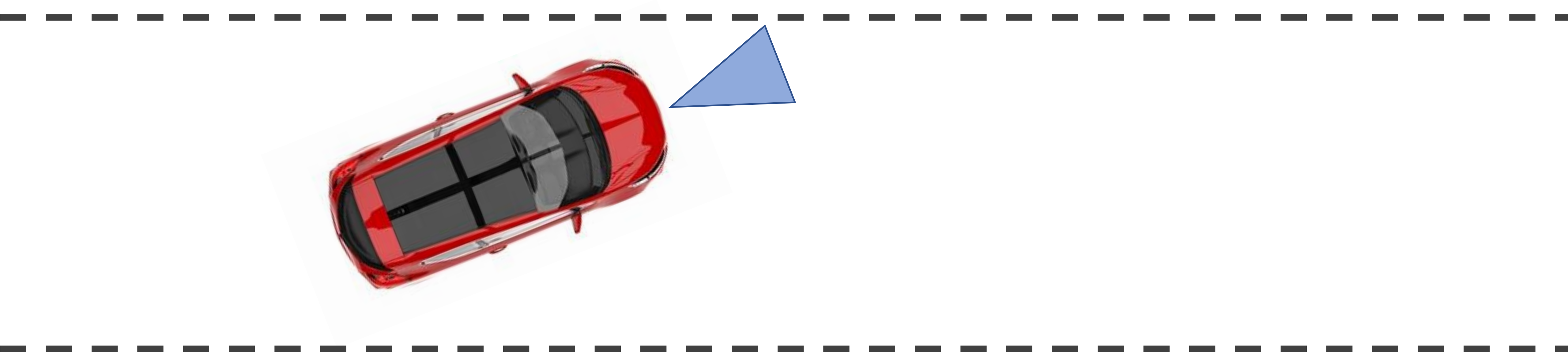
✗ Dangerous for expert!

✗ Difficult to pick effective perturbations

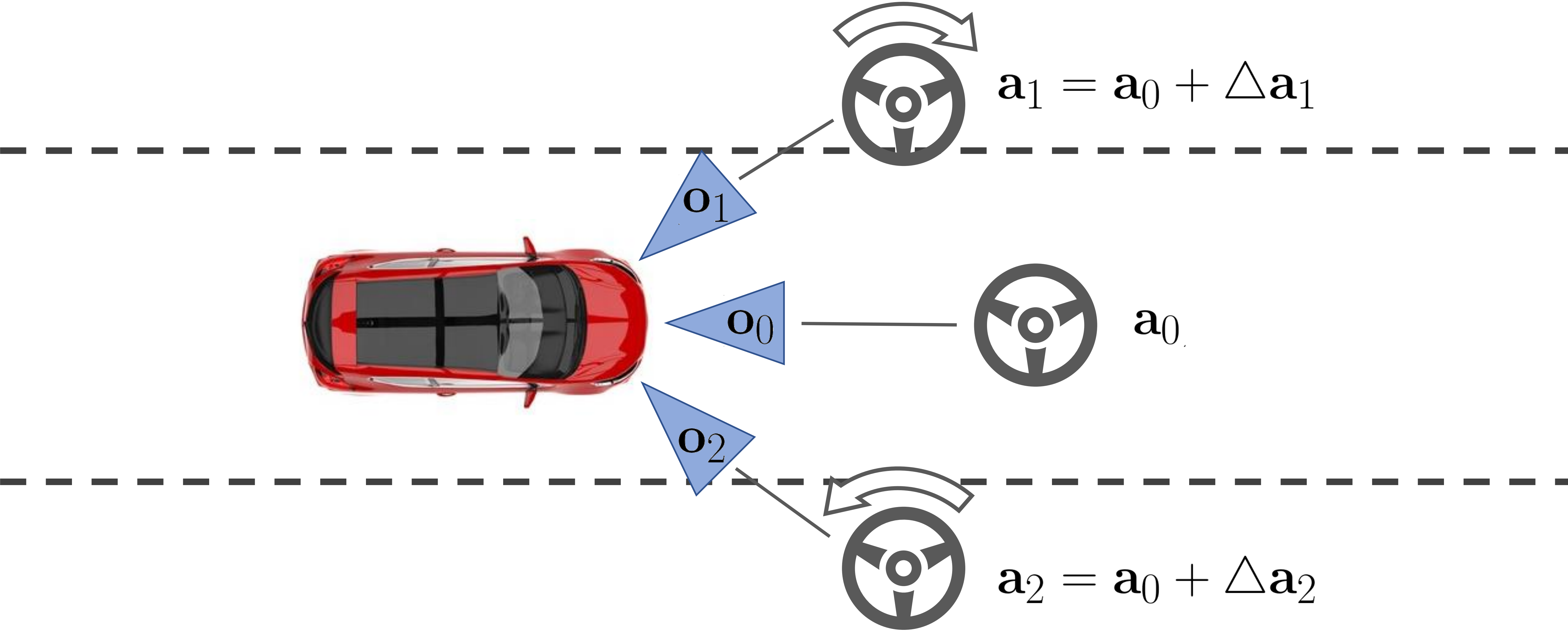
Data Augmentation



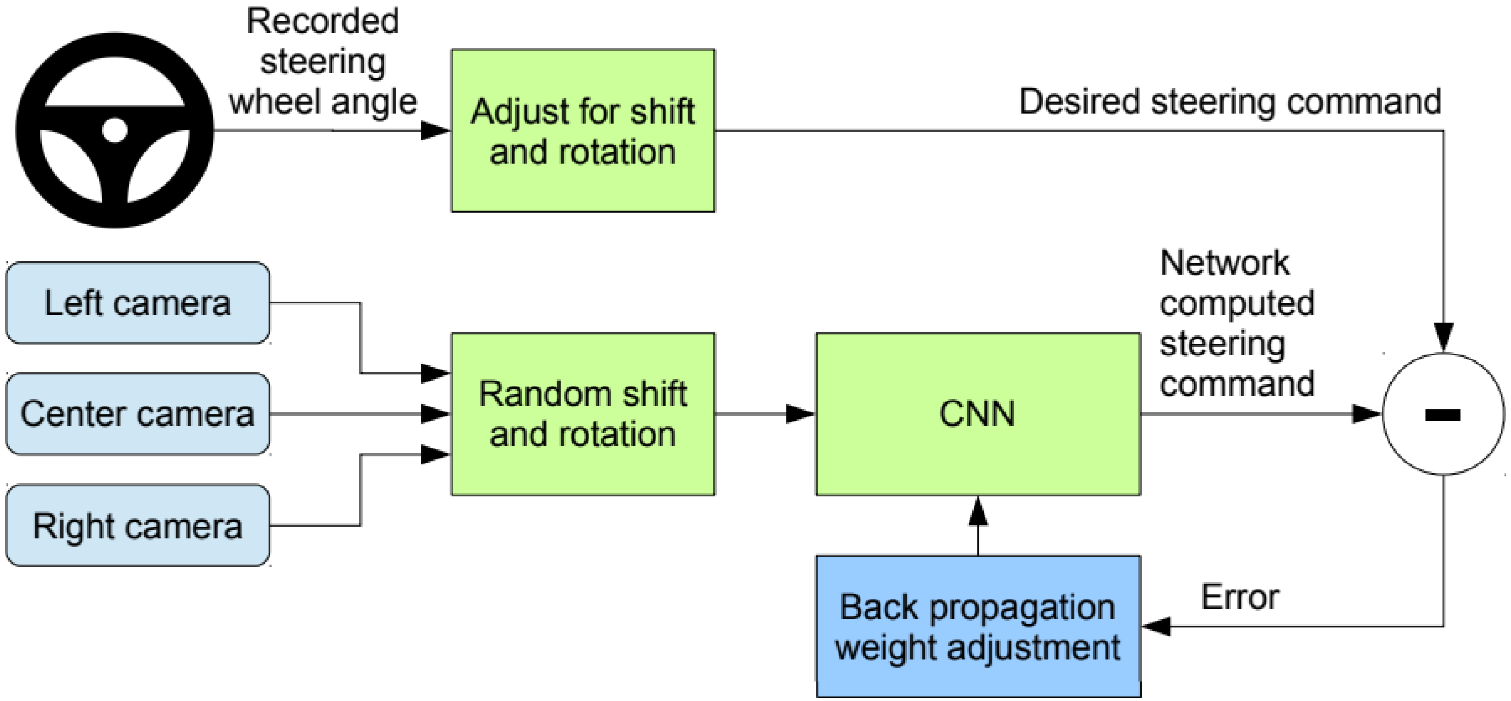
Data Augmentation



Data Augmentation


$\mathbf{a}_1 = \mathbf{a}_0 + \triangle\mathbf{a}_1$
$\mathbf{o}_1$
$\mathbf{o}_0$
$\mathbf{a}_0$
$\mathbf{o}_2$
$\mathbf{a}_2 = \mathbf{a}_0 + \triangle\mathbf{a}_2$

Data Augmentation



Recorded steering wheel angle

Adjust for shift and rotation

Desired steering command

Left camera

Center camera

Right camera

Random shift and rotation

CNN

Network computed steering command

-

Back propagation weight adjustment

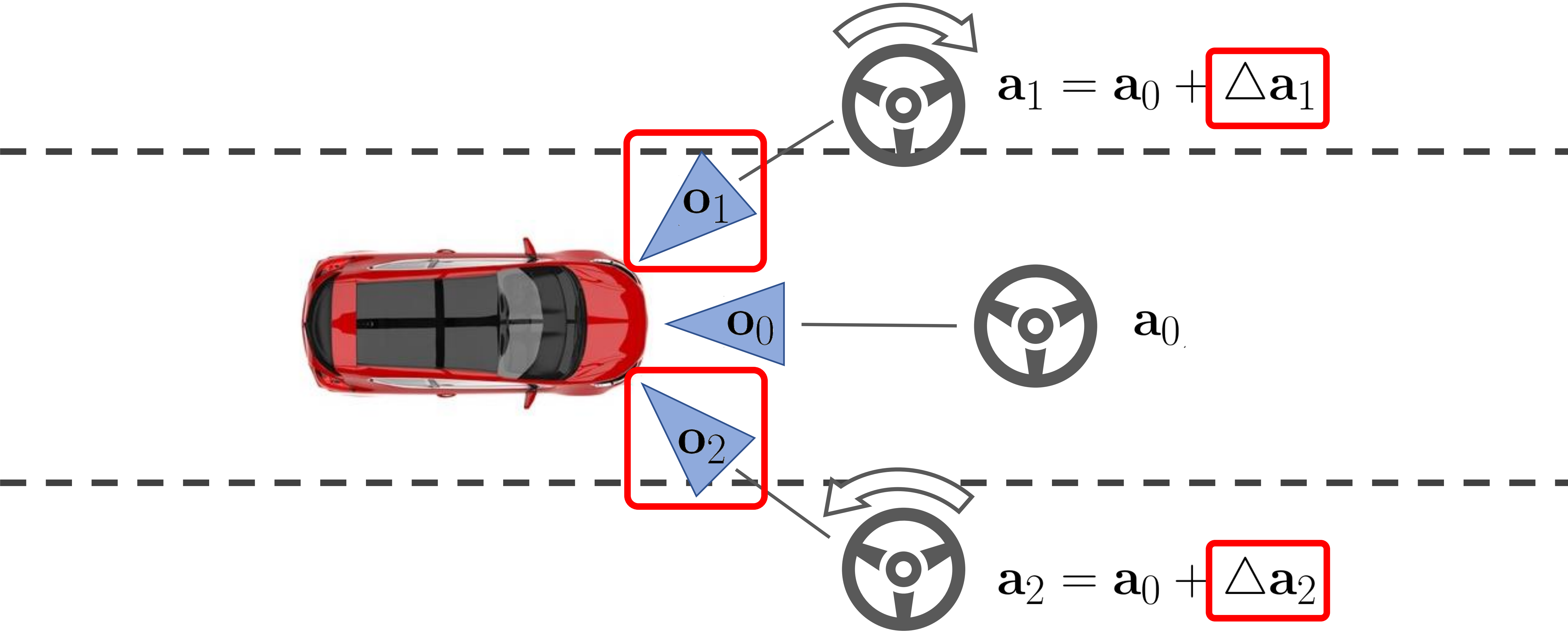
Error

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


$\mathbf{a}_1 = \mathbf{a}_0 + \boxed{\triangle\mathbf{a}_1}$
$\mathbf{o}_1$
$\mathbf{o}_0$
$\mathbf{a}_0$
$\mathbf{o}_2$
$\mathbf{a}_2 = \mathbf{a}_0 + \boxed{\triangle\mathbf{a}_2}$

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 $p_{\text{data}}(\mathbf{s},\mathbf{a})$

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

Such that

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

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

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

Theoretical Analysis

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

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

no mistakes in t timsteps

Theoretical Analysis

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

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

no mistakes in t timsteps

at least 1 mistakes in t timsteps

Theoretical Analysis

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

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

$$\underbrace{\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[c(\mathbf{s}, \mathbf{a})\right]}_{\text{expected cost}}$$

Theoretical Analysis

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

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

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

Theoretical Analysis

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

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

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

Theoretical Analysis

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

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

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

Theoretical Analysis

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

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

$$\sum_t E_{p_\pi^t(\mathbf{s})} E_{\pi(\mathbf{a}|\mathbf{s})}\left[\underline{c(\mathbf{s},\mathbf{a})}\right]$$

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

Theoretical Analysis

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

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

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

Theoretical Analysis

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

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

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

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

Theoretical Analysis

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

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

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

Theoretical Analysis

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

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

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

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

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

Theoretical Analysis

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

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

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

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

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

Theoretical Analysis

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

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

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

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

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

Theoretical Analysis

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

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

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

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

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

Theoretical Analysis

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

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

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

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

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

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

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

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

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

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

Theoretical Analysis

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

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

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

Theoretical Analysis

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

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

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

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

Theoretical Analysis

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

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

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

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

Theoretical Analysis

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

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

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

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

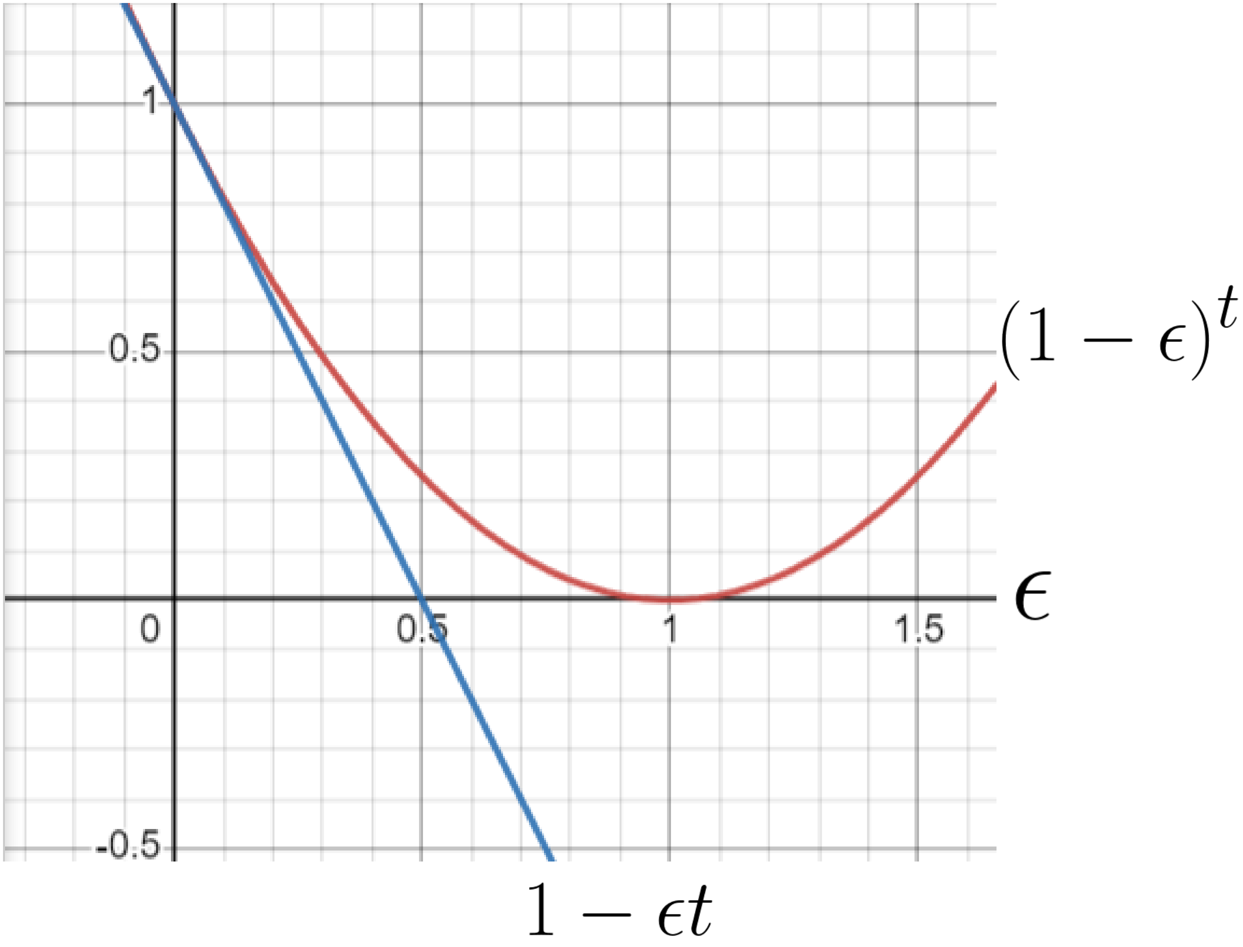
Note: $(1-\epsilon)^t \geq 1-\epsilon t \qquad \text{for } \epsilon \in [0,1]$

Theoretical Analysis

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

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

Note: $(1-\epsilon)^t \ge 1-\epsilon t \qquad \text{for } \epsilon \in [0,1]$



Theoretical Analysis

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

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

Theoretical Analysis

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

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

$$\leq \epsilon T + \underline{\sum_t 2\epsilon t}$$

Theoretical Analysis

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

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

$$\leq \epsilon T + \sum_t 2\epsilon t$$

$$\leq \epsilon T + 2\epsilon T^2 \in \boxed{O(\epsilon T^2)}$$

Worst Case

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

Worst Case

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

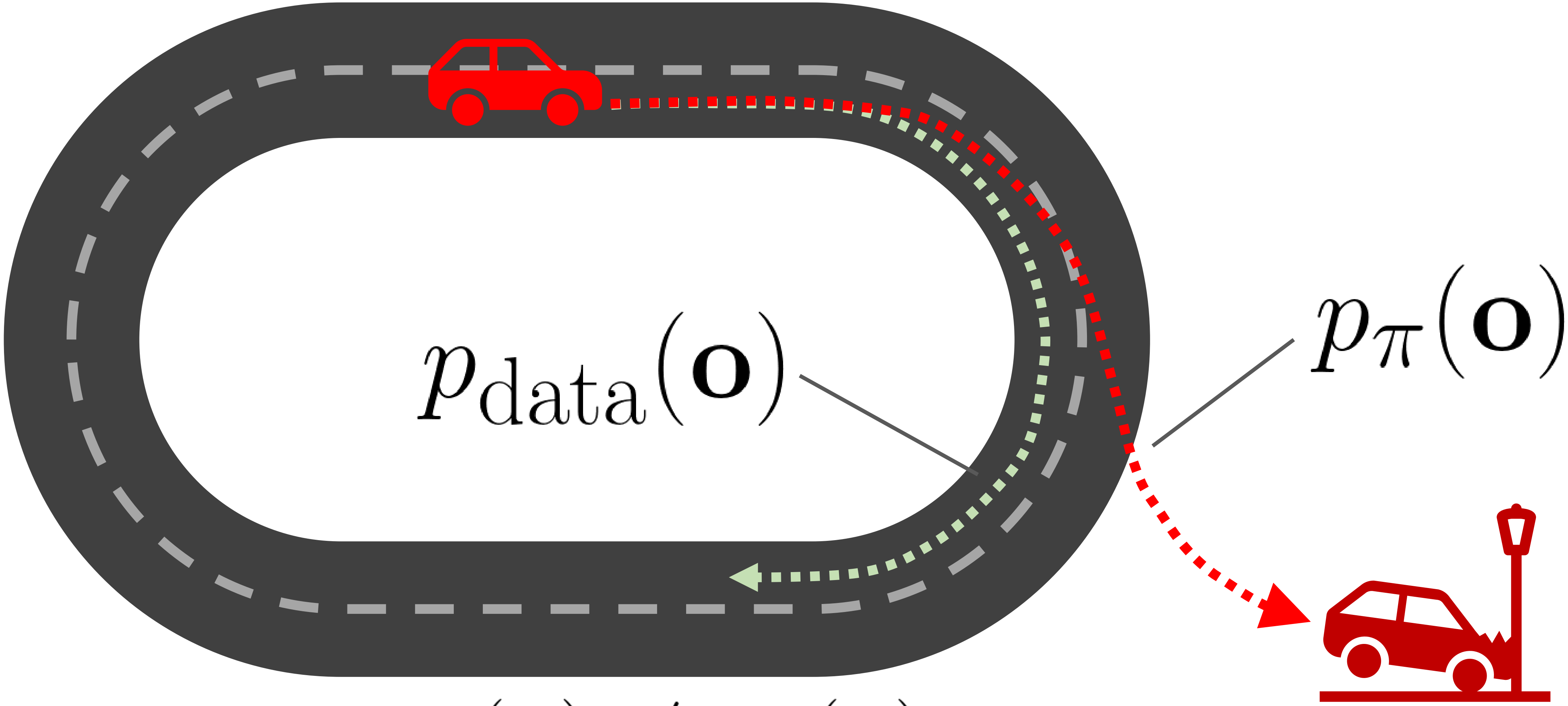
Distribution Shift

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

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

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

Distribution Shift



$$p_{\text{data}}(\mathbf{o}) \neq p_\pi(\mathbf{o})$$

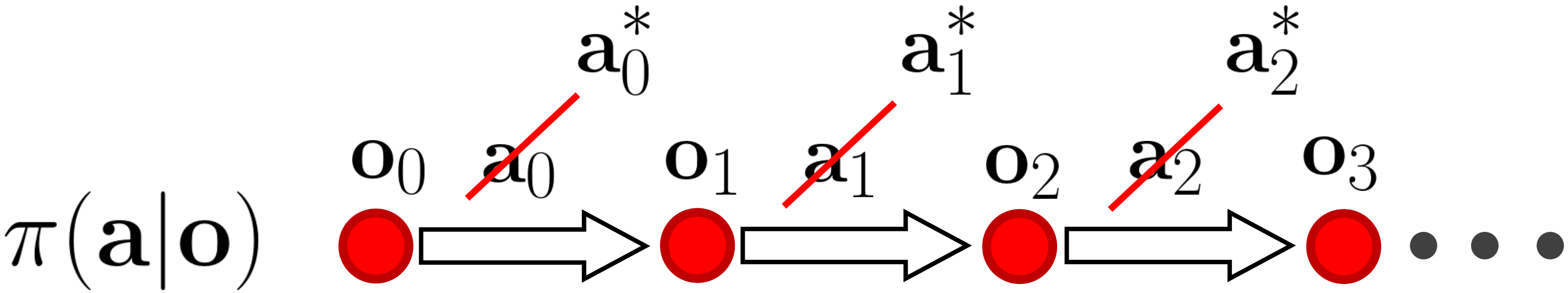
Dataset Aggregation

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

Key idea:

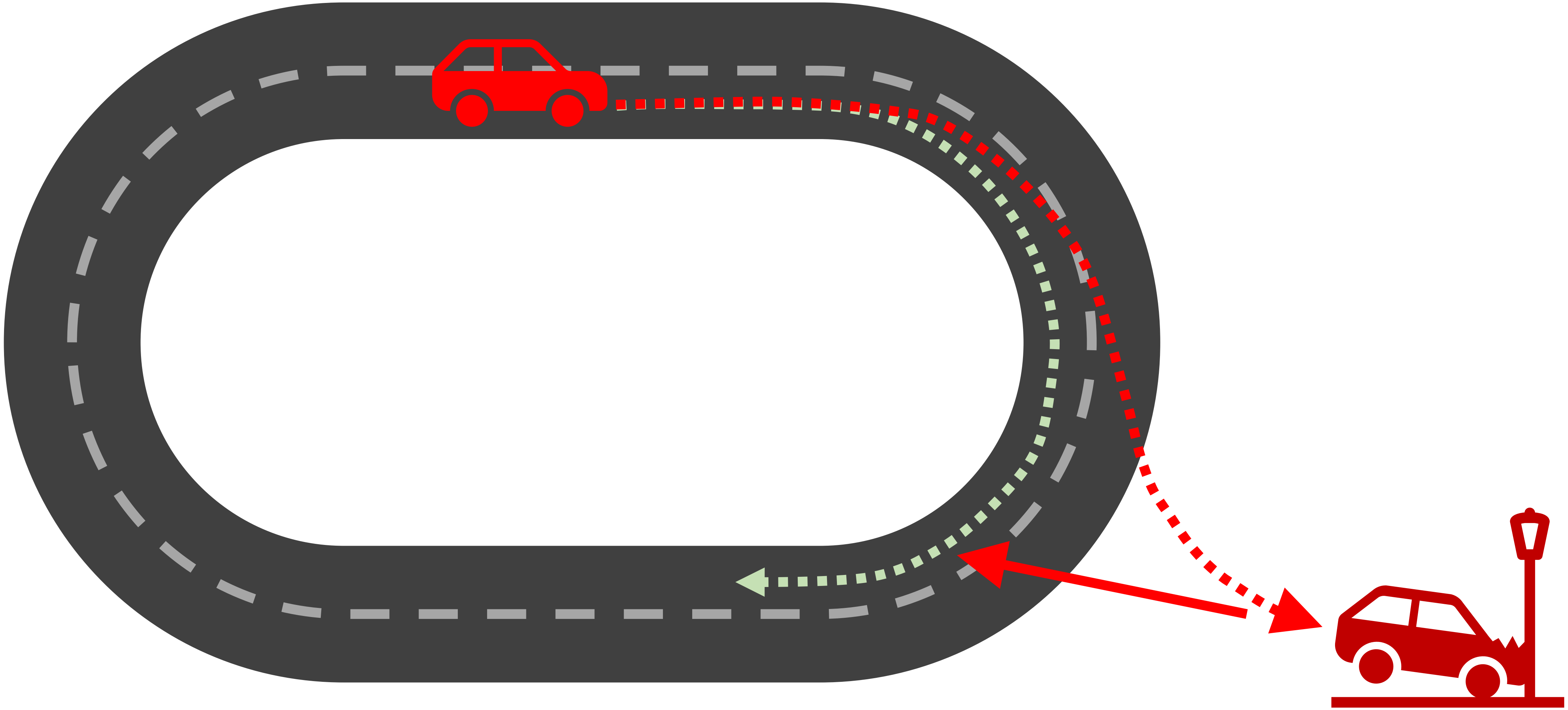
- Collect observations from $p_\pi(\mathbf{o})$ instead of $p_{\text{data}}(\mathbf{o})$
- Label actions with expert
- DAgger: Dataset Aggregation [Ross et al. 2011]

DAgger

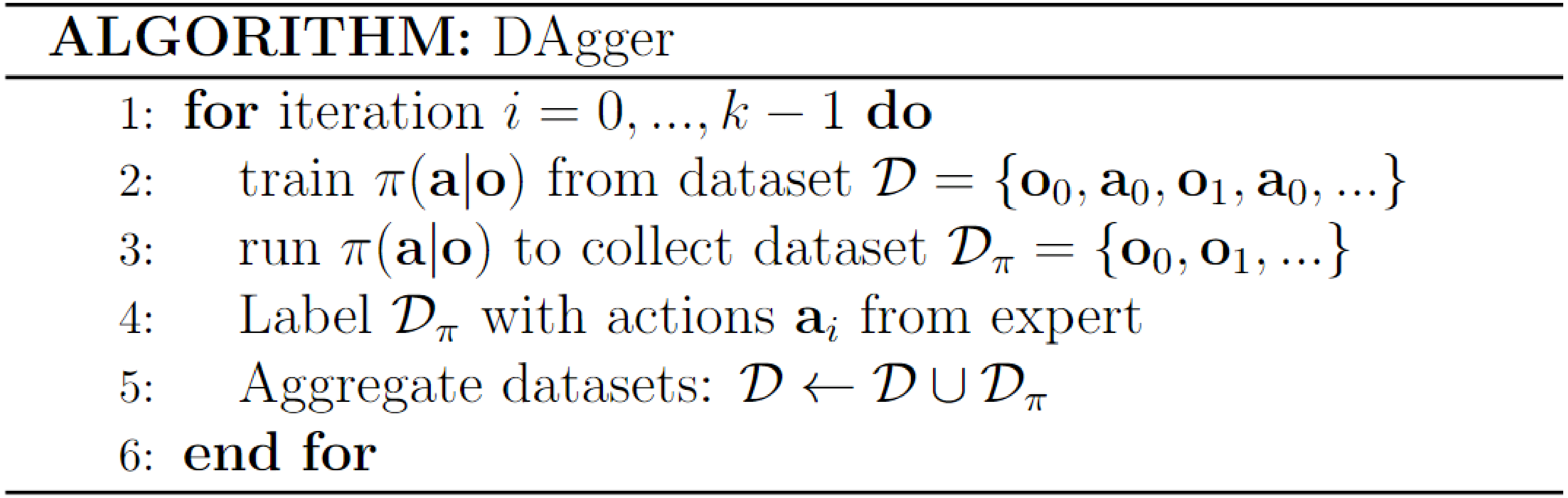


Train with $(\mathbf{o}_i, \mathbf{a}_i^*)$

DAgger



DAgger

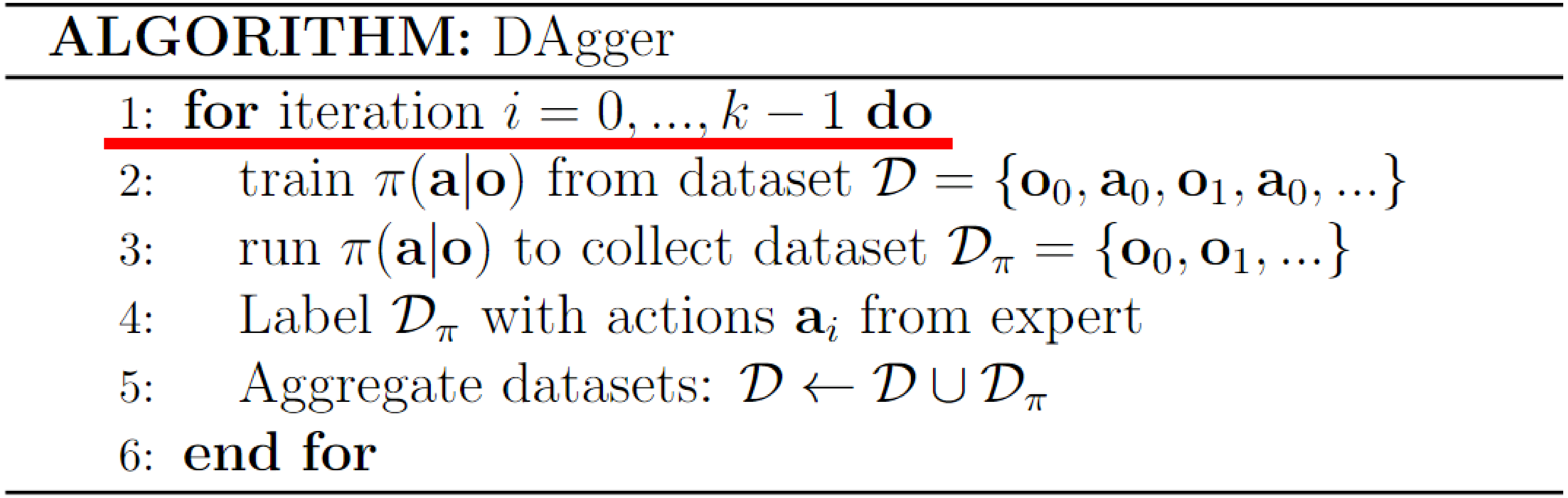


ALGORITHM: DAgger

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

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

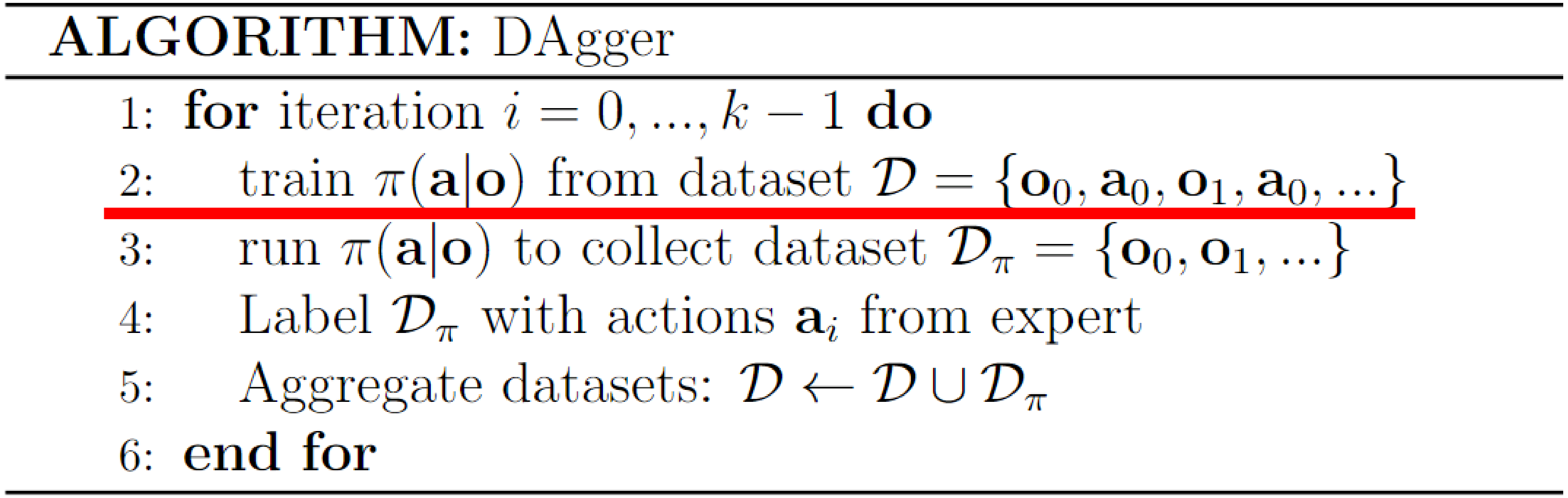
DAgger



```
ALGORITHM: DAgger
for iteration $i = 0, ..., k-1$ do
    train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
    run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
    Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
    Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
end for
```

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

DAgger

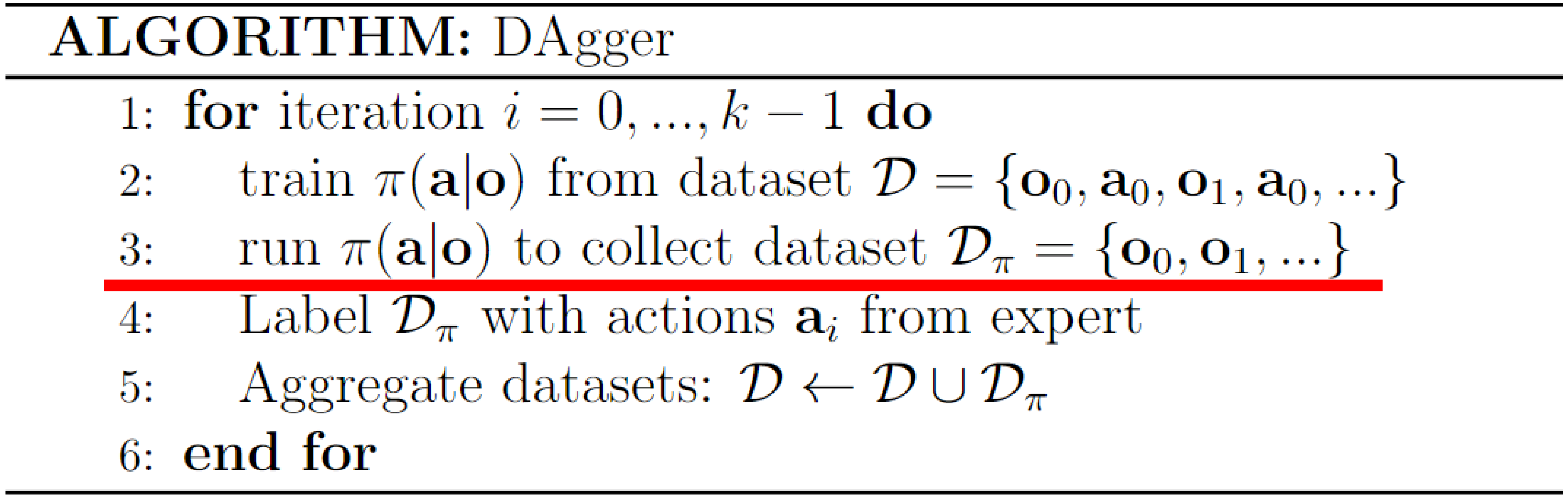


ALGORITHM: DAgger

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

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

DAgger

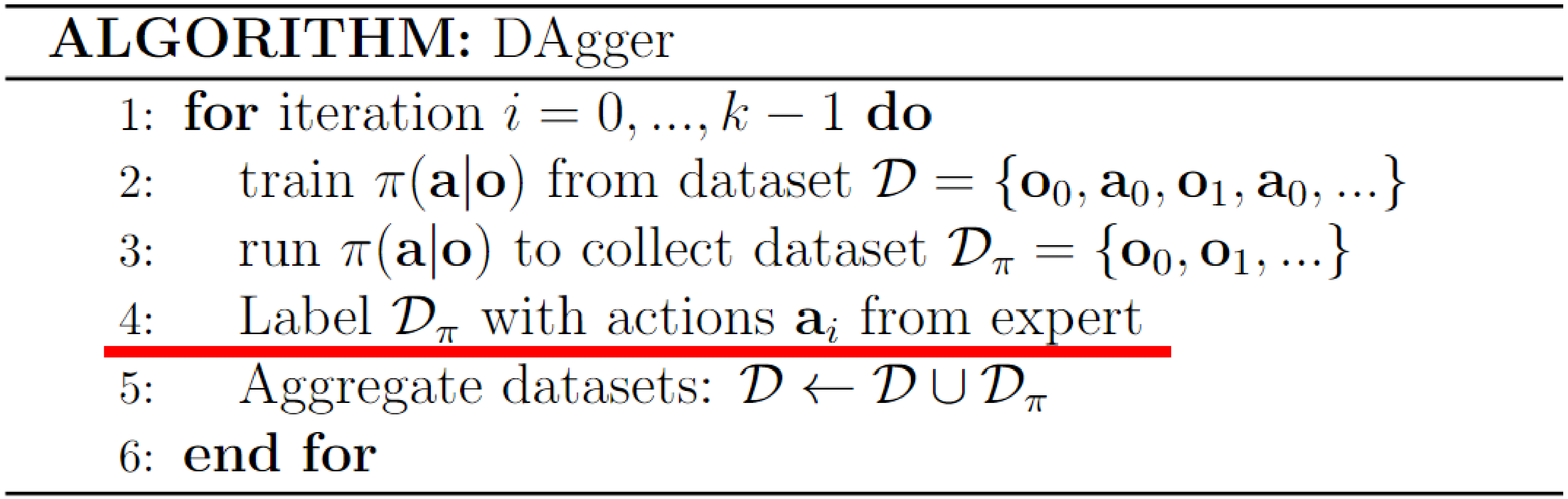


ALGORITHM: DAgger

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

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

DAgger

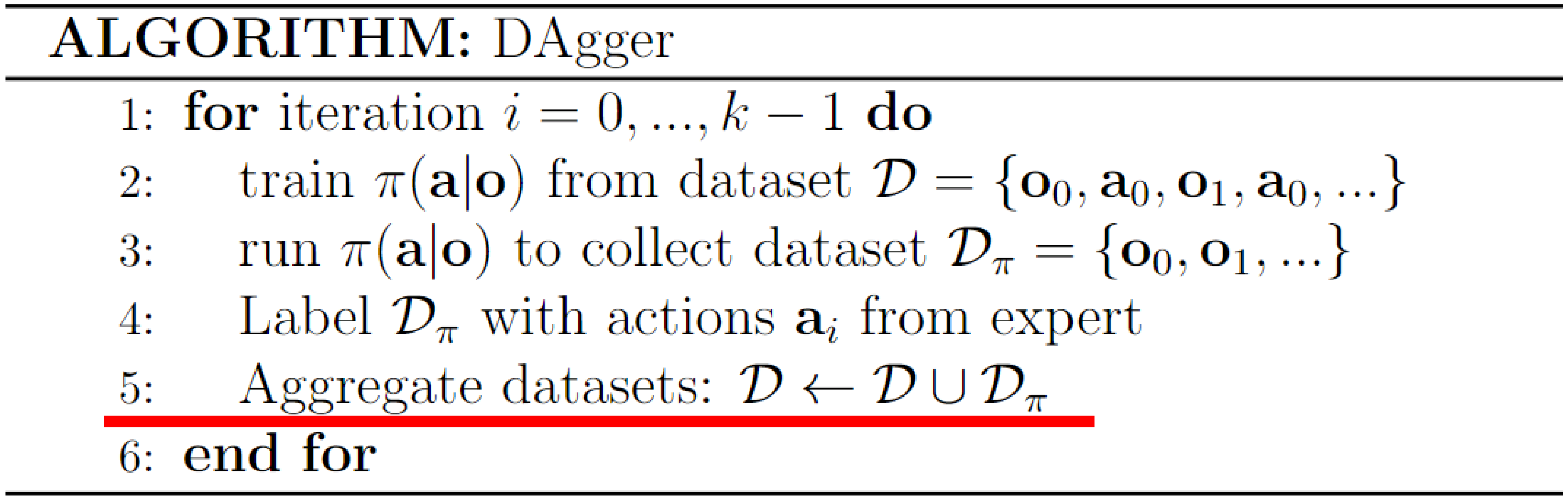


ALGORITHM: DAgger

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

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

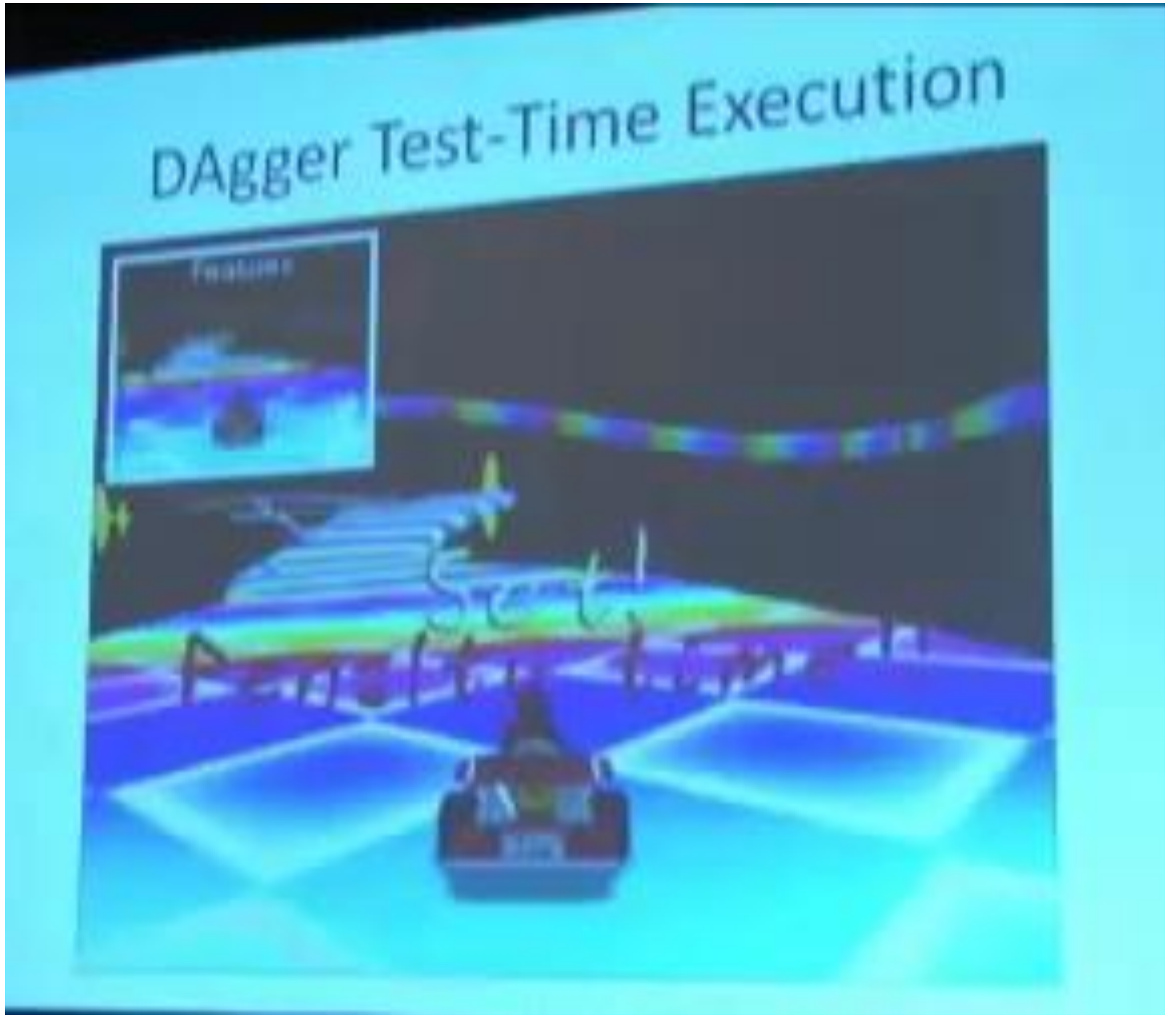
DAgger



```
ALGORITHM: DAgger
for iteration $i = 0, ..., k-1$ do
    train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
    run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
    Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
    Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
end for
```

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

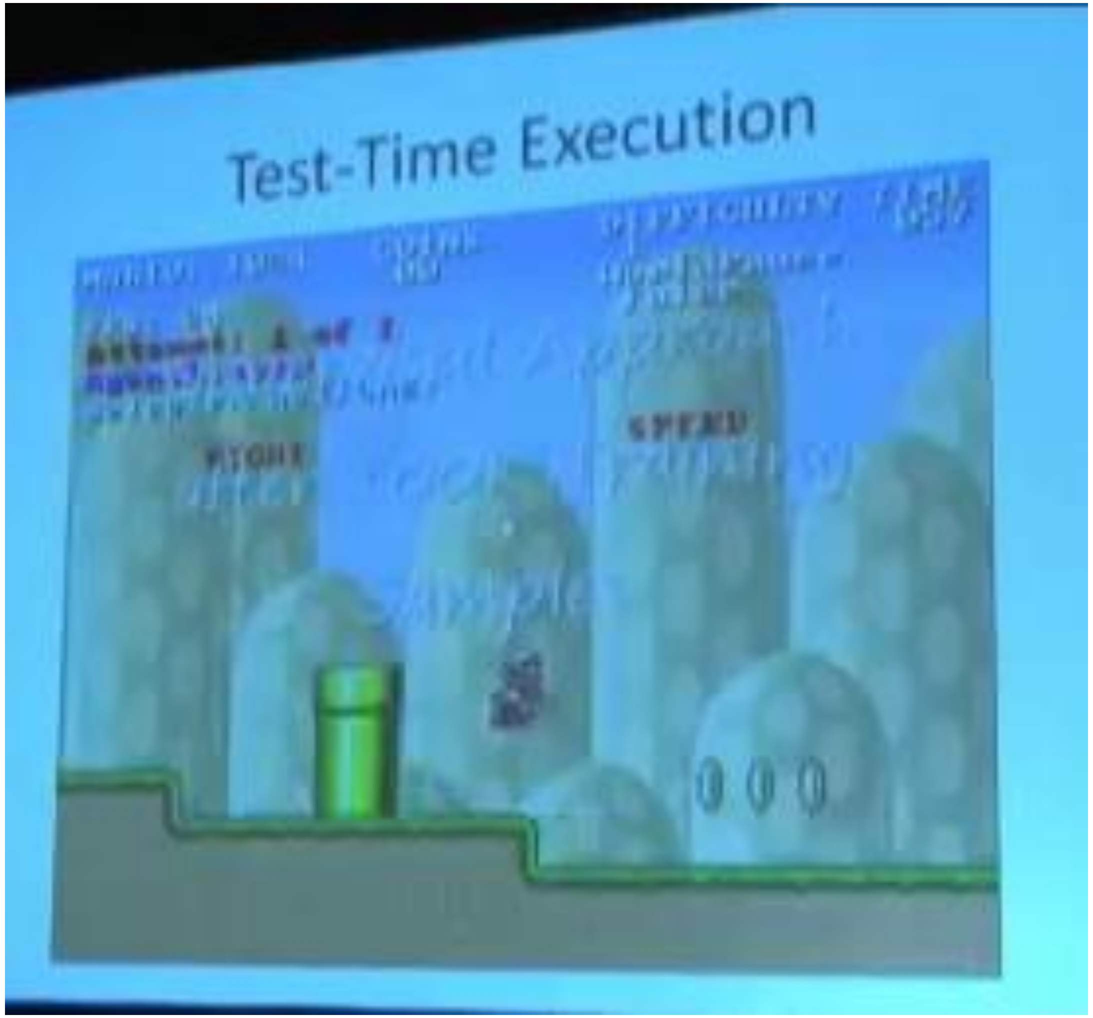
DAgger



DAgger Test-Time Execution

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

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

$$p^t_\pi(\mathbf{s}) = (1-\epsilon)^t p^t_{\text{data}}(\mathbf{s}) + (1-(1-\epsilon)^t)\underbrace{p^t_{\text{mistake}}(\mathbf{s})}_{= p^t_{\text{data}}(\mathbf{s})}$$

DAgger Analysis

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

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

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

$$= p^t_{\text{data}}(\mathbf{s})$$

DAgger Analysis

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

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

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

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

DAgger Analysis

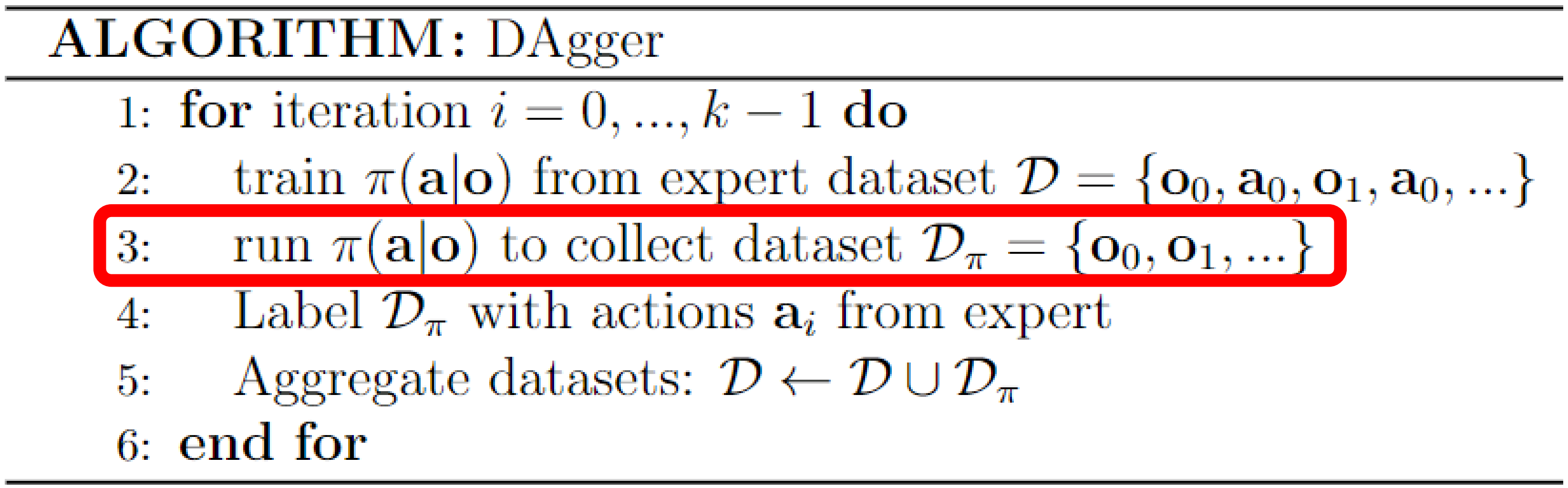
Assume: $\underline{\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}\right) \leq \epsilon \quad \text{for all } \mathbf{s} \sim p_{\text{data}}(\mathbf{s})}$

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

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

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

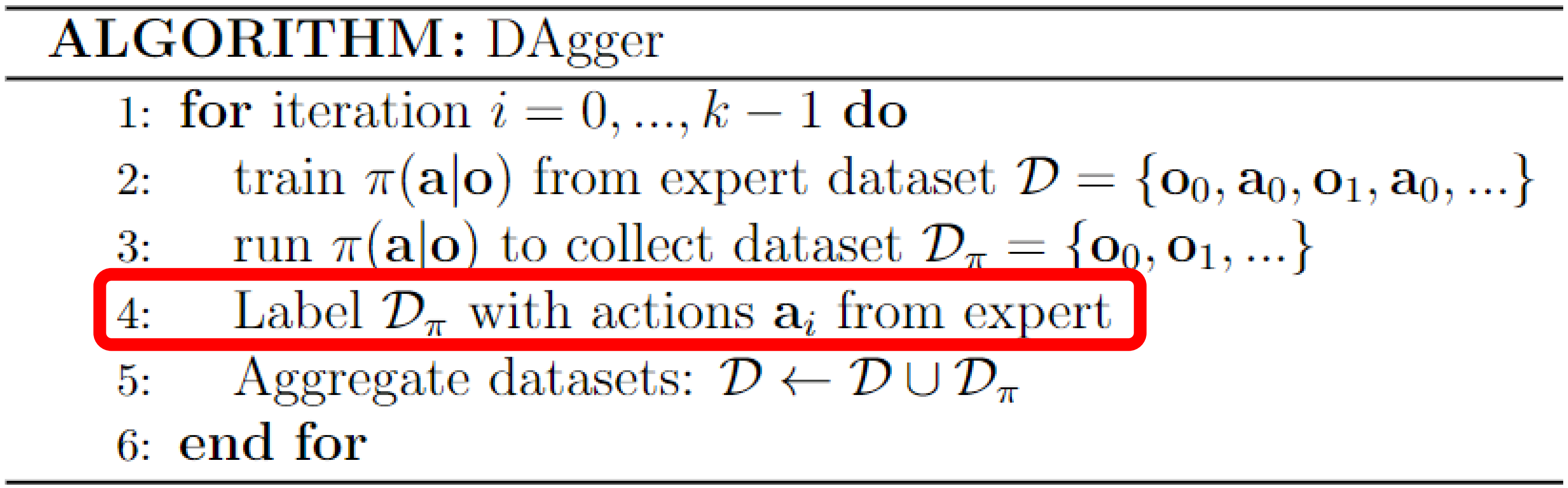
DAgger



```
ALGORITHM: DAgger
for iteration $i = 0, ..., k - 1$ do
    train $\pi(\mathbf{a}|\mathbf{o})$ from expert dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
    run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
    Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
    Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
end for
```

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

DAgger

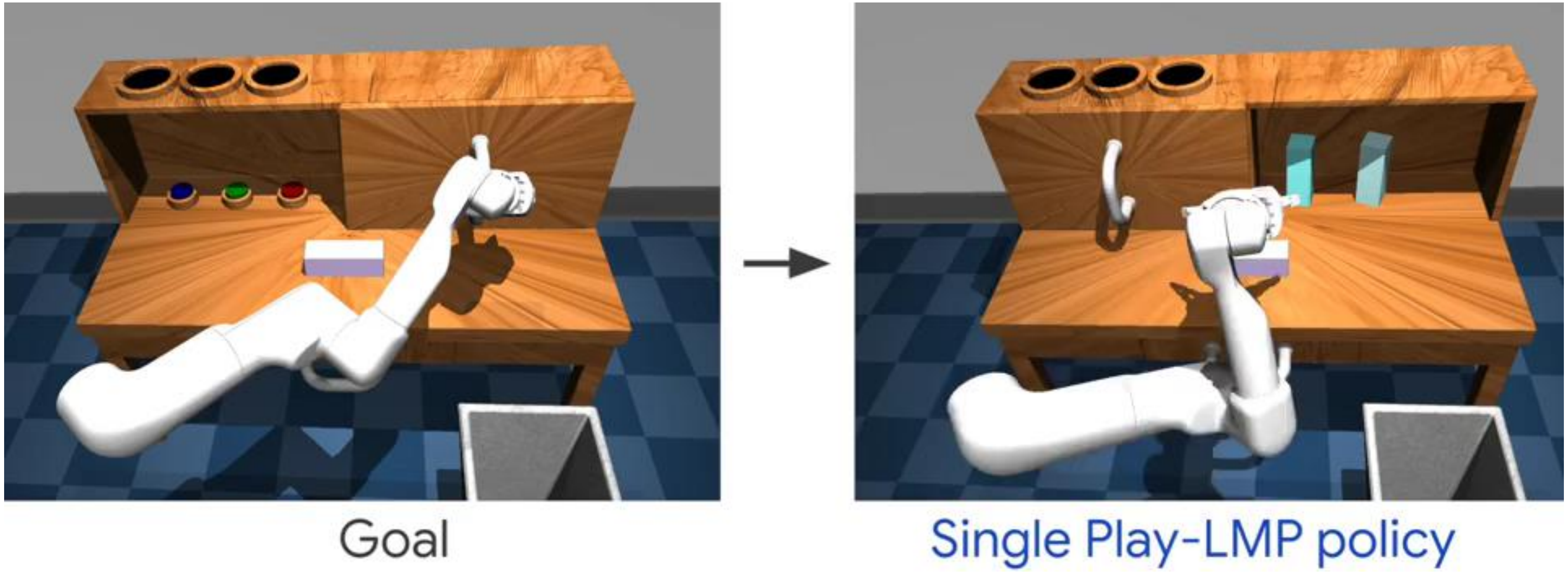


```
ALGORITHM: DAgger
for iteration $i = 0, ..., k-1$ do
    train $\pi(\mathbf{a}|\mathbf{o})$ from expert dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
    run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
    Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
    Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
end for
```

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

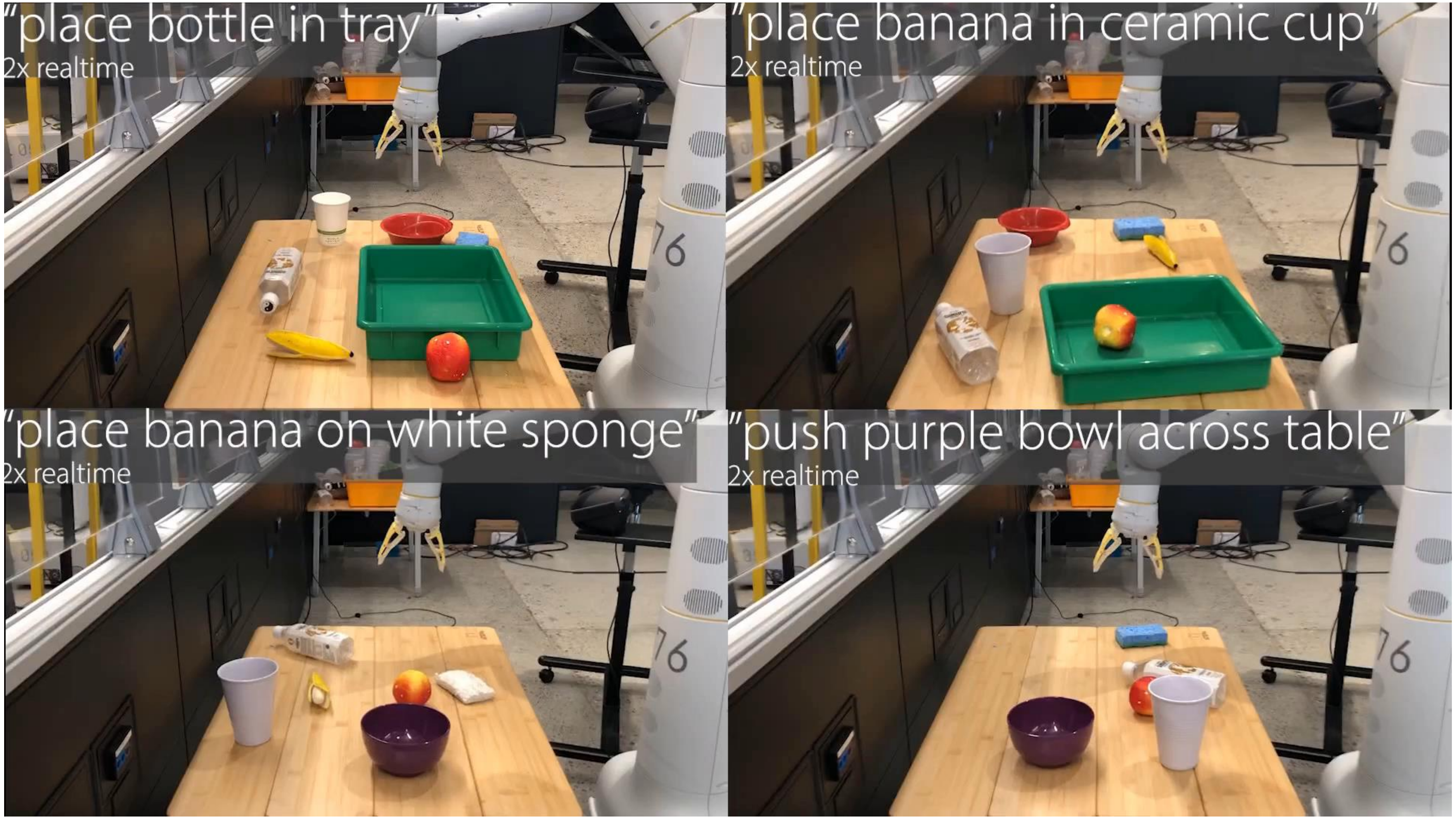
Applications

Applications



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

Applications


"place bottle in tray"
2x realtime
"place banana in ceramic cup"
2x realtime
"place banana on white sponge"
2x realtime
"push purple bowl across table"
2x realtime

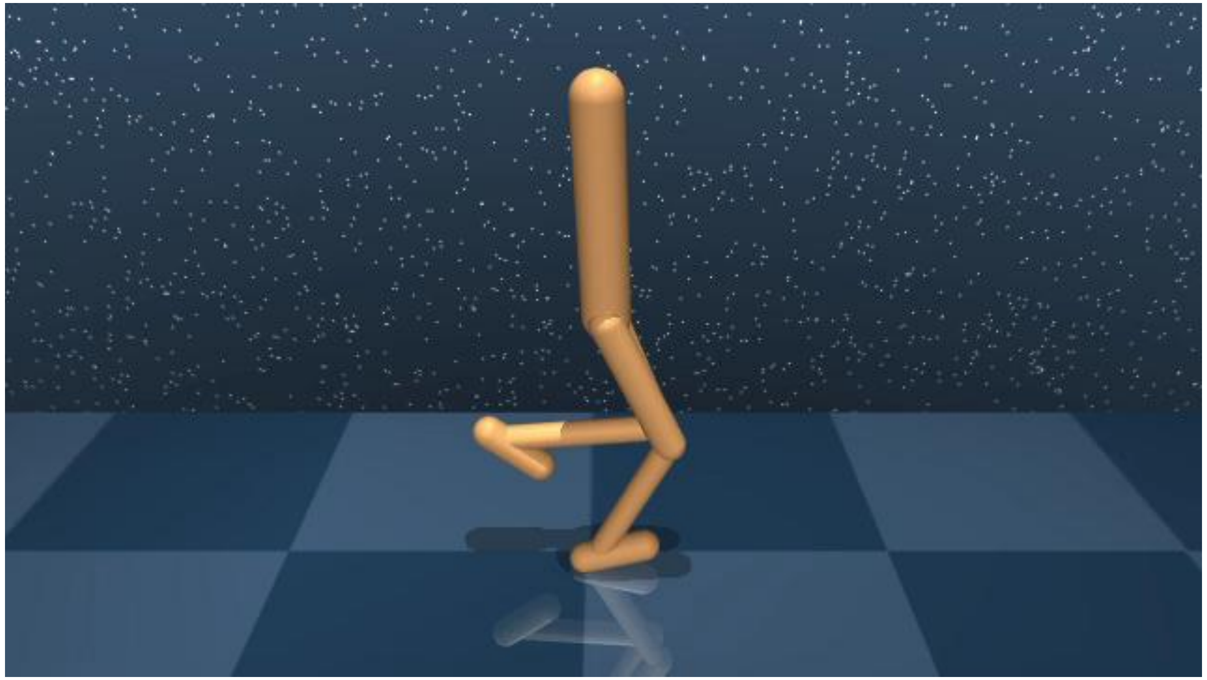
BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning
[Jang et al. 2021]

Summary

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

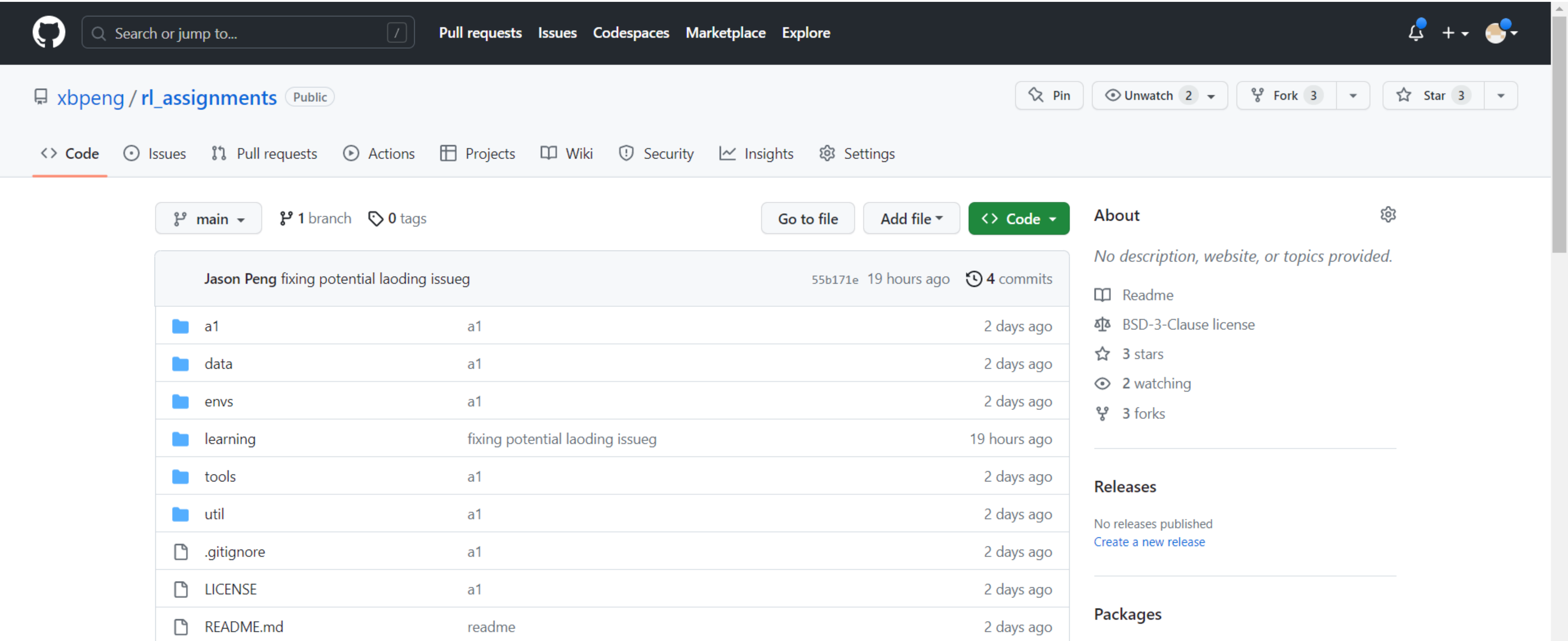
Assignment 1: Behavioral Cloning

Cheetah



Walker

Assignment 1: Behavioral Cloning



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

a1	a1	2 days ago
data	a1	2 days ago
envs	a1	2 days ago
learning	fixing potential laoding issueg	19 hours ago
tools	a1	2 days ago
util	a1	2 days ago
.gitignore	a1	2 days ago
LICENSE	a1	2 days ago
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