Expert-In-The-Loop For Sequential Decision-Making And Predictions

Kianté Brantley | Postdoctoral Scholar | Cornell University

Sequential Decision-Making and Predictions

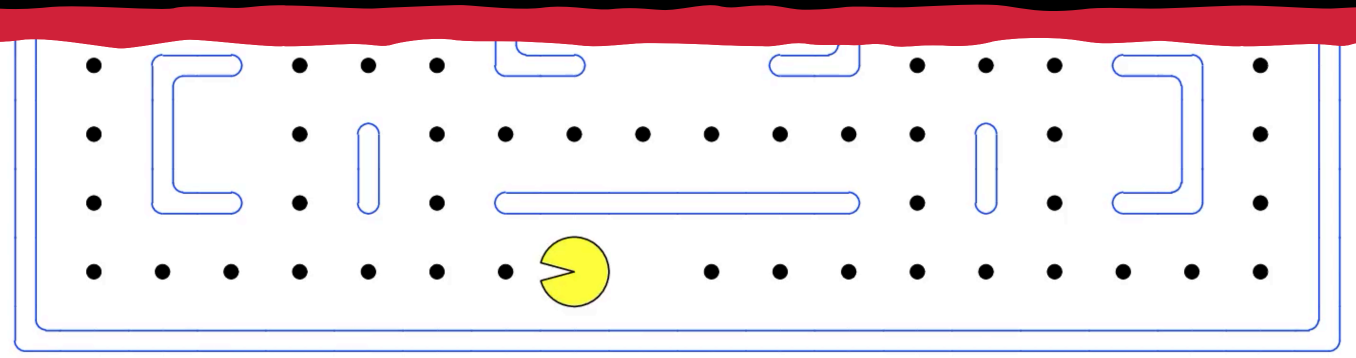
problems

Robotics

Sequence Labeling

Text Generation

Video Games



(Named Entity Recognition)

Source: After completing his Ph.D. , Ellis worked at Bell Labs from 1969 to

Target: O O O O O PER O O ORG ORG O O O

(t_3) , (t_4) Ellis (t_5)

s_4 s_5

Target: How are you?

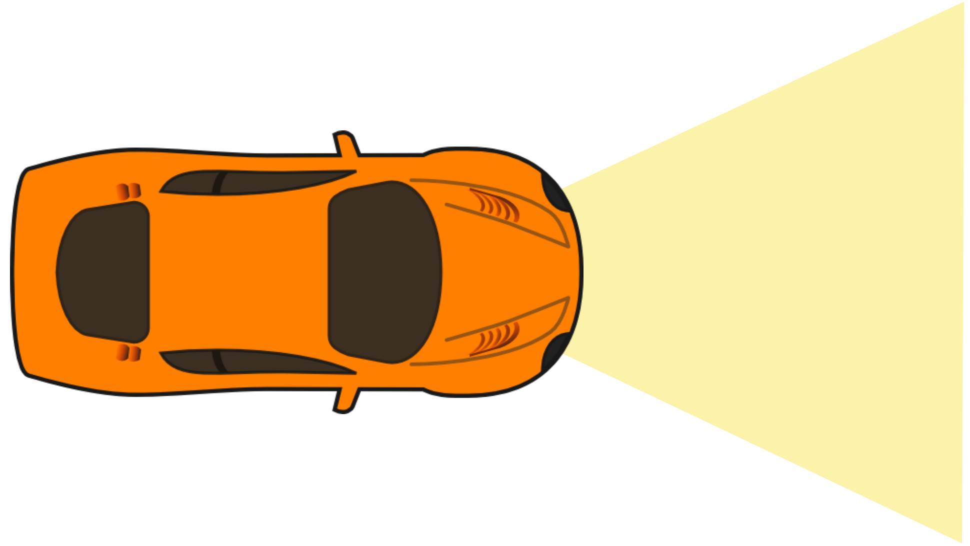
How (w_1) are (w_2) you (w_3) ? (w_4) <eos> (w_5)

$s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4 \rightarrow s_5$

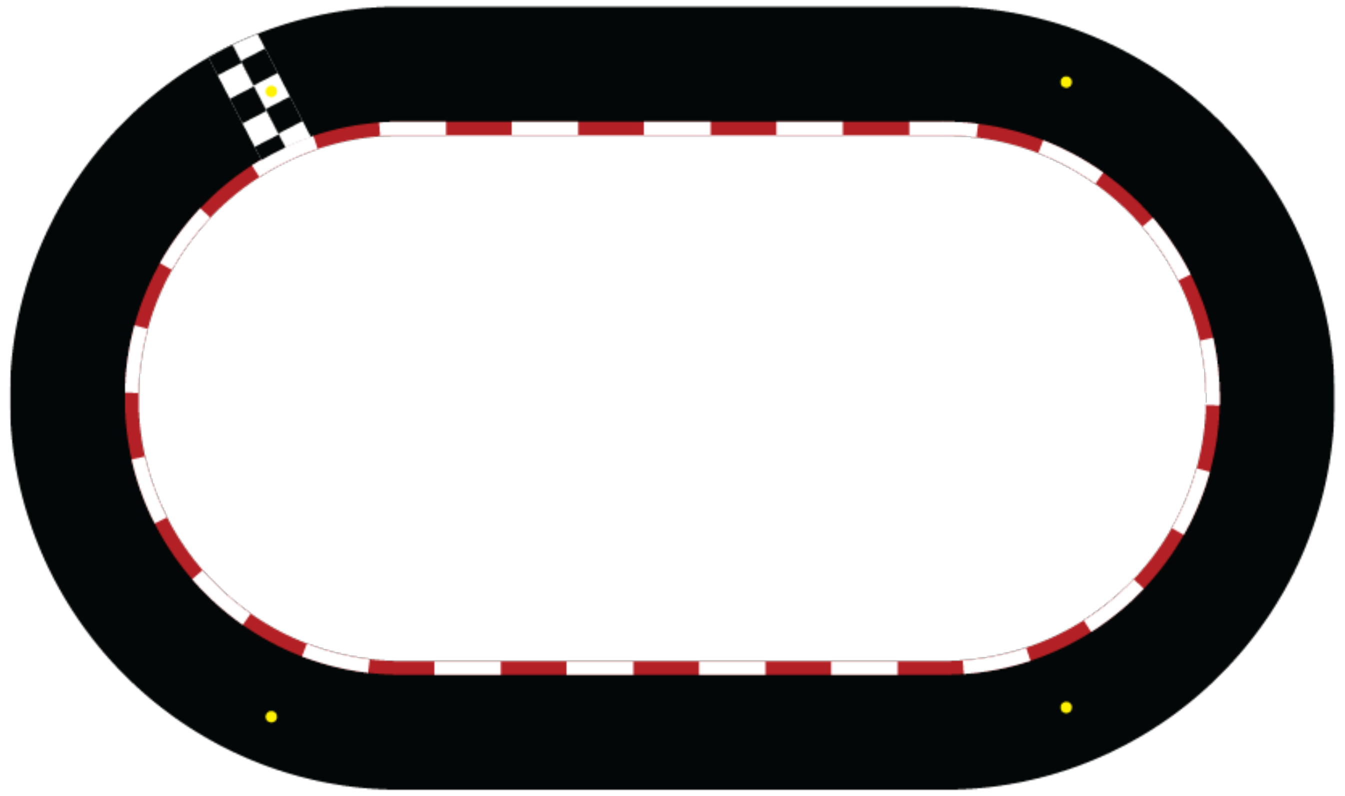
Issue:

Reinforcement Learning requires millions of interactions in an environment to solve sequential decision-making problems.

Agent

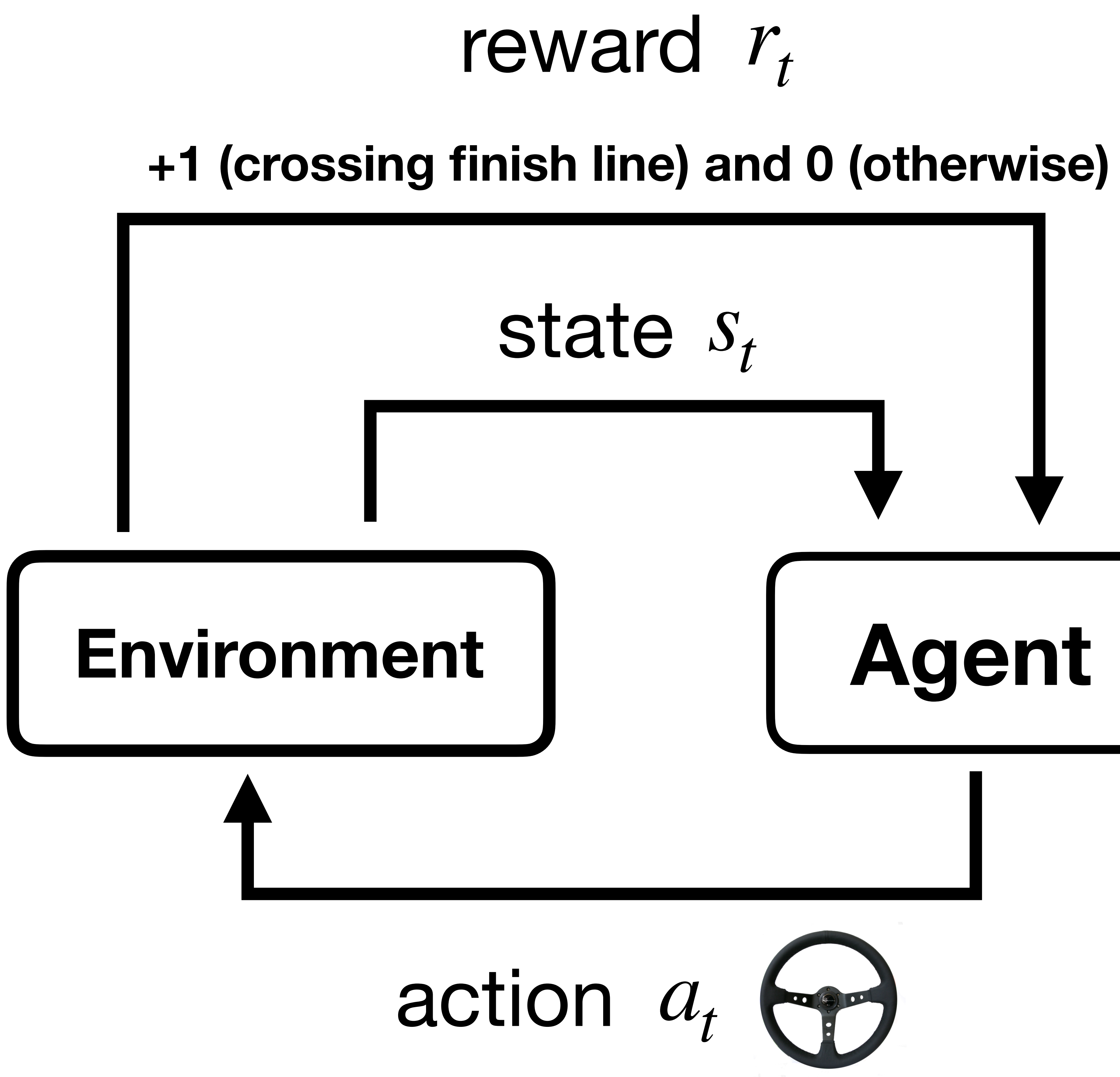


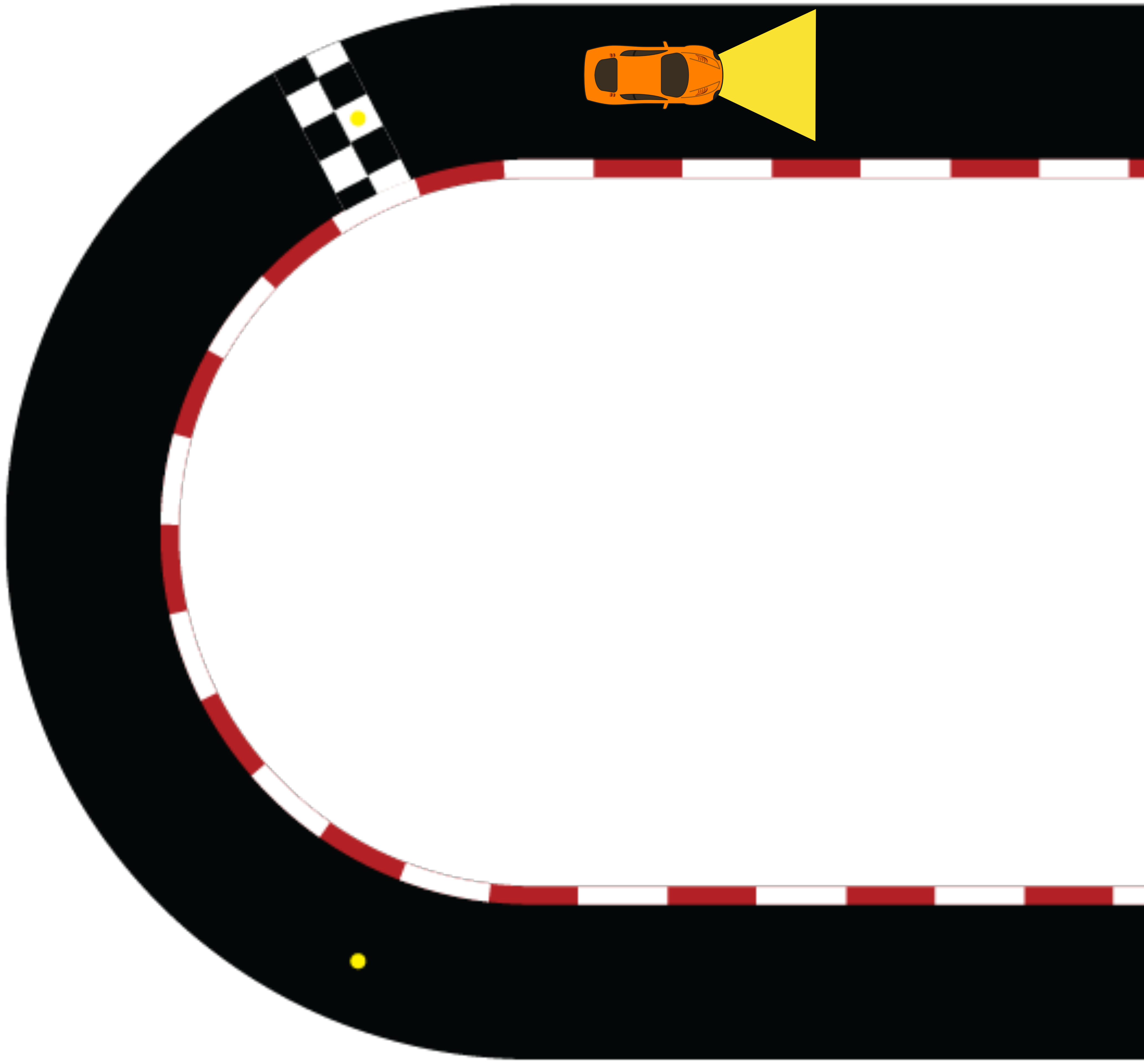
Environment



Reinforcement Learning

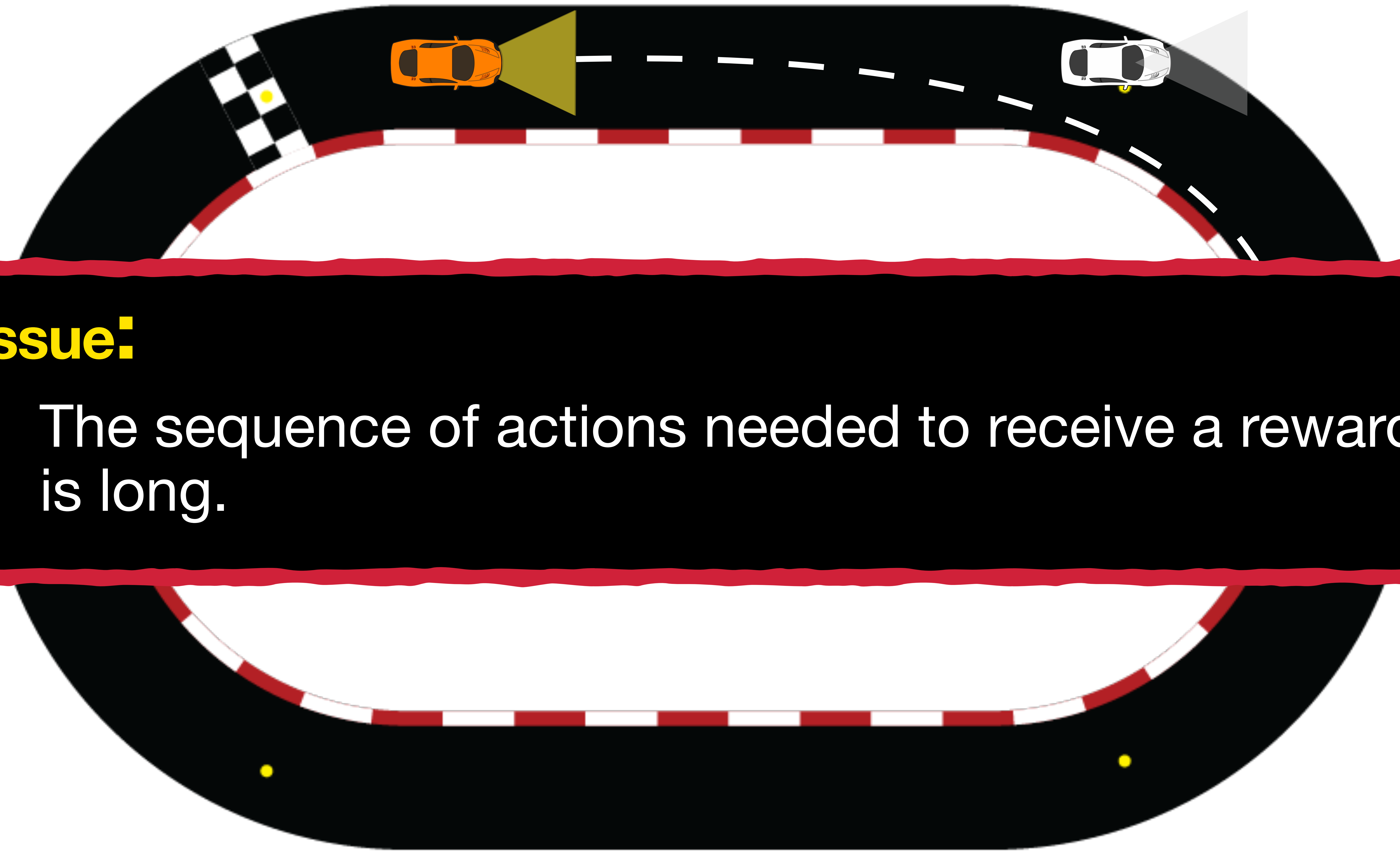
basics





Reinforcement Learning

basics

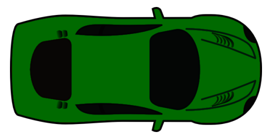


Issue:

The sequence of actions needed to receive a reward is long.

Imitation learning

basics

Expert/Oracle Demonstrator 

- state

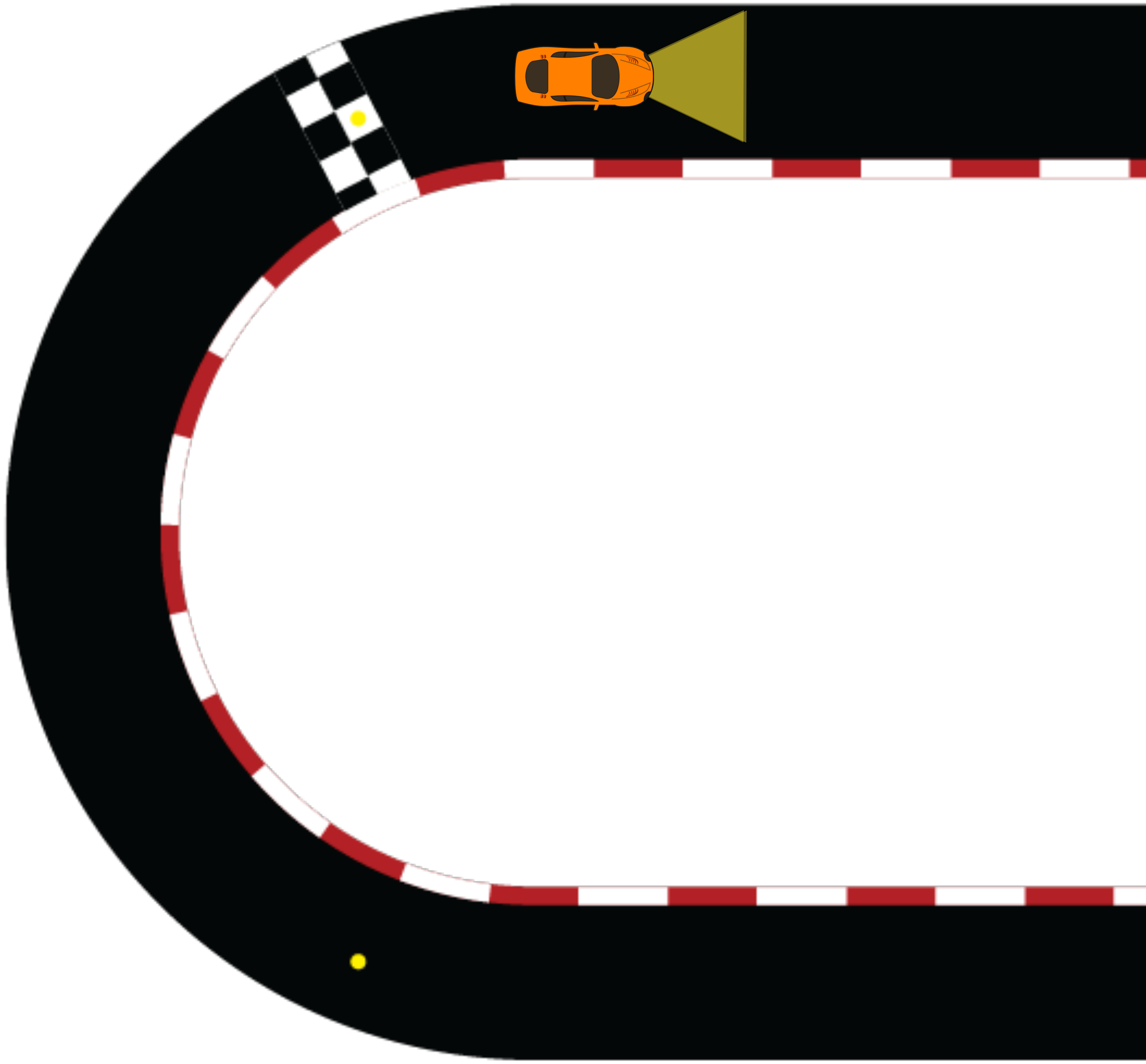
- actions

Training set:

D = {(state, actions)} from expert π^*

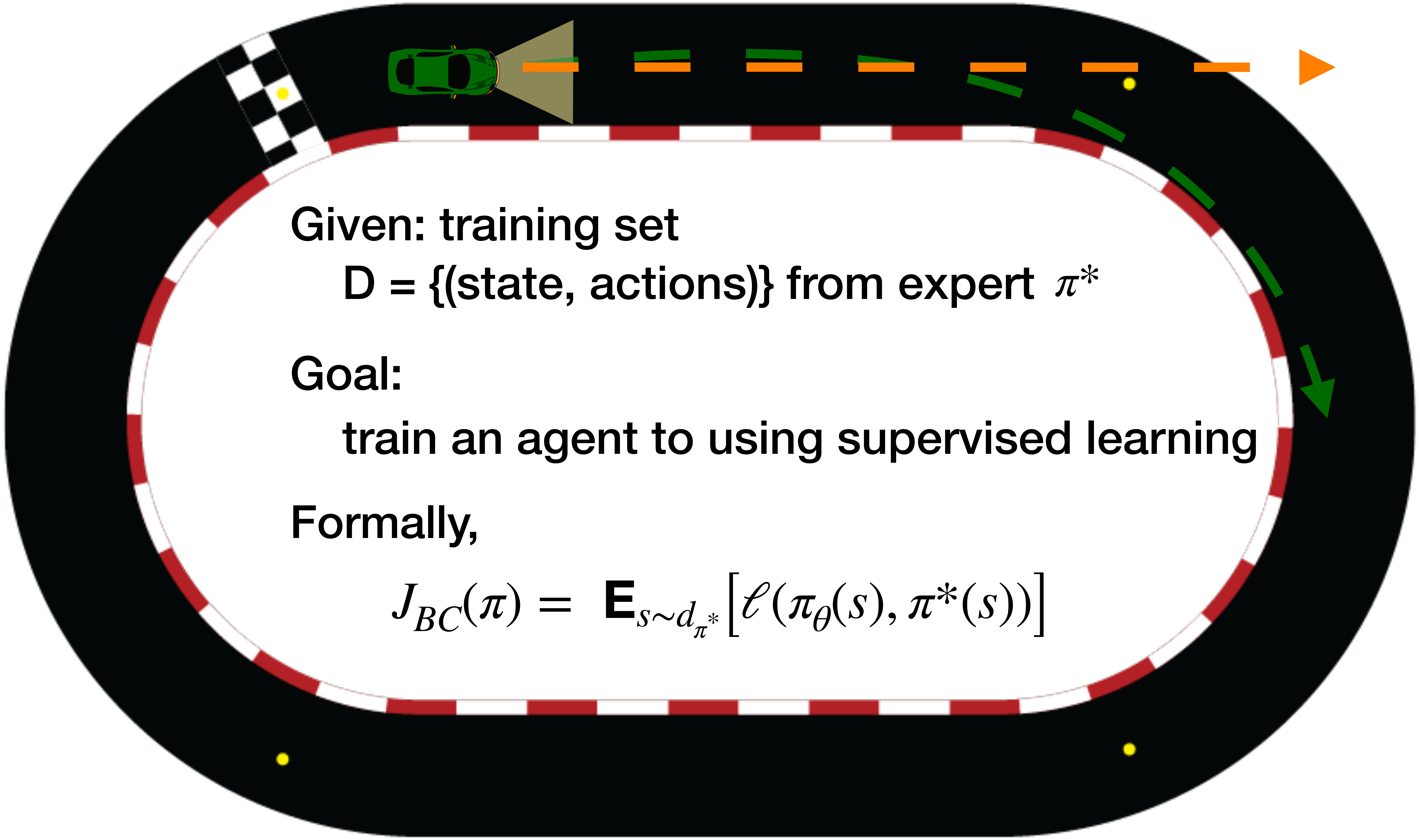
Goal:

learn an agent π_θ(s) -> a



Imitation learning with Behavior Cloning

covariate shift



Given: training set

D = {(state, actions)} from expert π^*

Goal:

train an agent to using supervised learning

Formally,

$$J_{BC}(\pi) = \mathbf{E}_{s \sim d_{\pi^*}}\left[\ell(\pi_\theta(s), \pi^*(s))\right]$$

[ALVINN: An Autonomous Land Vehicle in a Neural Network, Dean Pomerleau Neurips 1989]
[An Invitation to Imitation - Semantic Scholar, Bagnell]

Imitation learning with Behavior Cloning

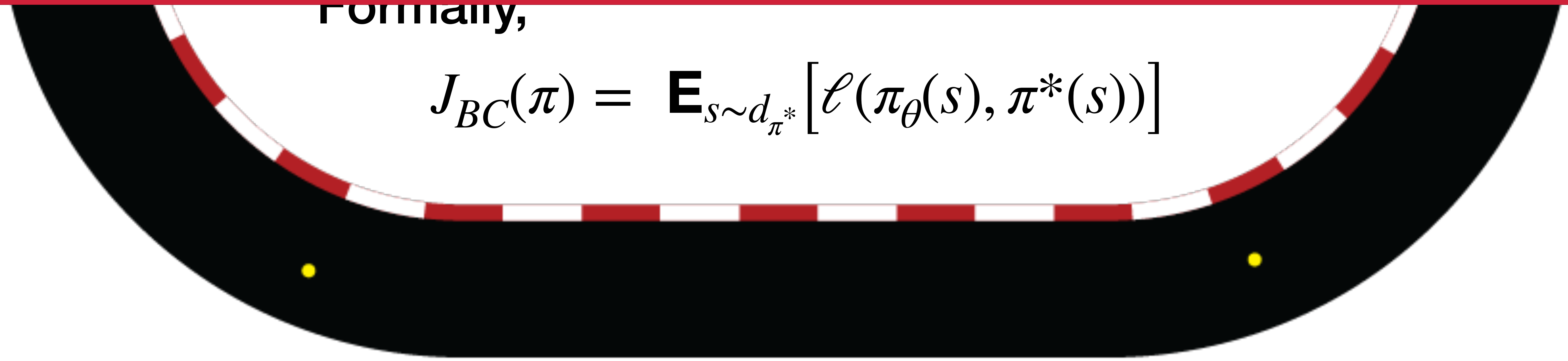
covariate shift

Issue:

The assumptions underlying supervised learning no longer hold, resulting in a covariate shift issue.

	Supervised Learning	Behavior Cloning
Train	$(x, y) \sim D$	$(s, a) \sim d_{\pi^*}$
Test	$(x, y) \sim D$	$(s, a) \sim d_{\pi}$

Formally,



$$J_{BC}(\pi) = \mathbf{E}_{s \sim d_{\pi^*}}\left[\ell(\pi_\theta(s), \pi^*(s))\right]$$

[ALVINN: An Autonomous Land Vehicle in a Neural Network, Dean Pomerleau Neurips 1989]
[An Invitation to Imitation - Semantic Scholar, Bagnell]

Structured Prediction with Behavior Cloning

exposure bias in nlp

Task: Word Descrambling Text-Generation

How are you ?

Source:

Target: How are you?

you	how	?	are	<eos>

context: previous ground truth words

Train

$f(\cdot|X)$ | $f(\cdot|X,$ **how**$)$ | $f(\cdot|X,$ **how,are**$)$ | $f(\cdot|X,$ **how,are, you**$)$ | $f(\cdot|X,$ **how,are, you, ?**$)$

How (w_1) → are(w_2) → you(w_3) → ? (w_4) → <eos> (w_5)

s_1 → s_2 → s_3 → s_4 → s_5

context: previous model predicted words

Test

$f(\cdot|X)$ | $f(\cdot|X,$ **can**$)$ | $f(\cdot|X,$ **can,you**$)$ | $f(\cdot|X,$ **can,you, you**$)$ | $f(\cdot|X,$ **can,you, you, ?**$)$

Can (w_1) → you(w_2) → you(w_3) → ? (w_4) → <eos> (w_5)

s_1 → s_2 → s_3 → s_4 → s_5

Structured Prediction with Behavior Cloning

exposure bias in nlp

Task: Word Descrambling Text-Generation

Issue:

The assumptions underlying supervised learning no longer hold, resulting in the covariate issue/exposure bias.

	Supervised Learning	Behavior Cloning
Train	$(x, y) \sim D$	$(s, a) \sim d_{\pi^*}$
Test	$(x, y) \sim D$	$(s, a) \sim d_{\pi}$

Talk Overview

Background

- Behavior Cloning
- Interactive Imitation Learning with Dagger (Ross et al. 2011)

Modern Imitation Learning

- Uncertainty-Based Learning (ICLR’ 20)
- An Empirical Study of Imitation Learning (Under Review)

Talk Overview

Background

- Behavior Cloning
- Interactive Imitation Learning with Dagger (Ross et al. 2011)

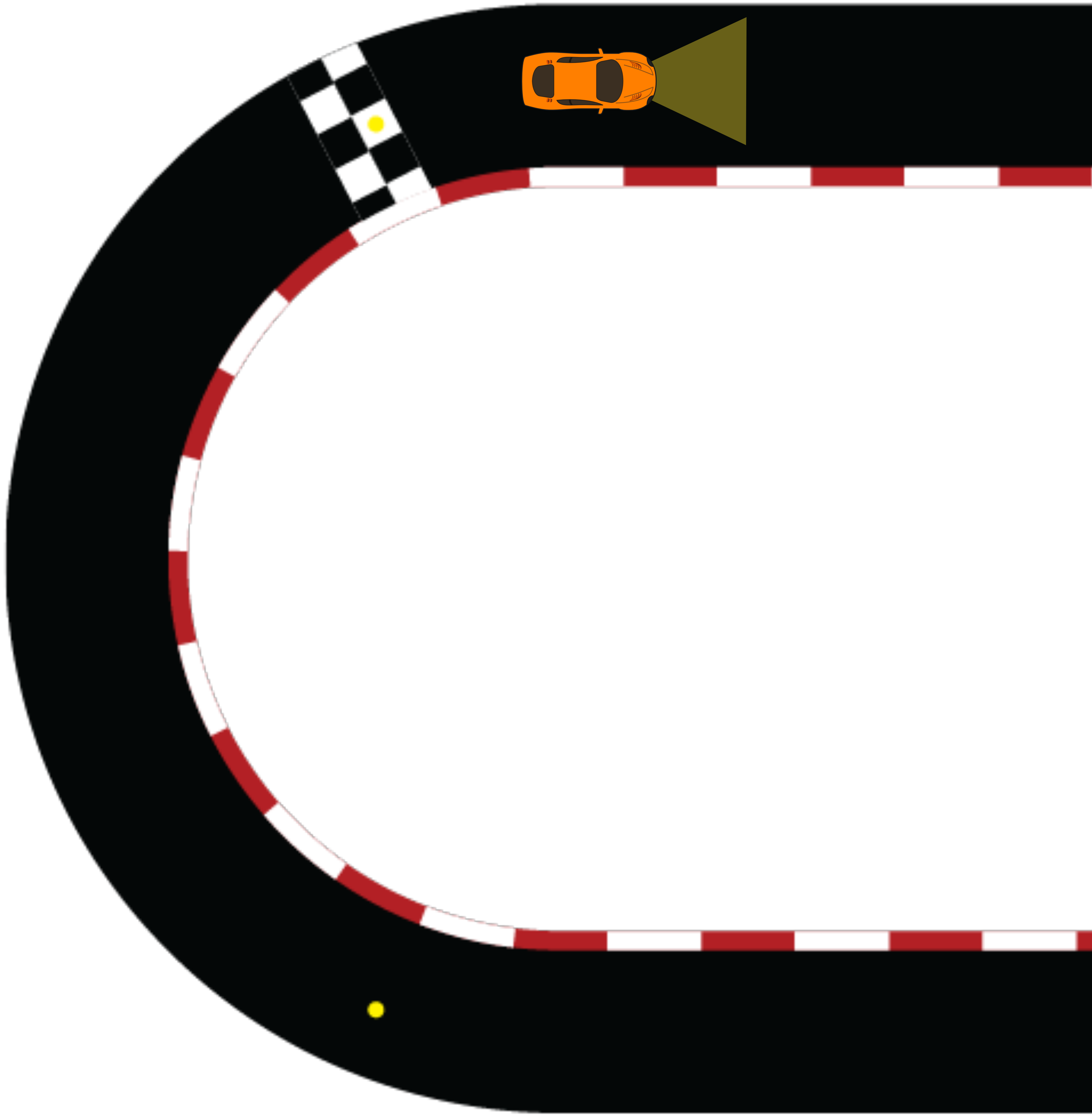
Research Question:

Can we design algorithms to deal with the exposure bias/covariate shift issue?

Interactive Imitation Learning

with dagger

Uses an online queryable expert



Initialize Dataset D

Initialize $\hat{\pi}_1$

For $i = 1$ **to** N **do**

$\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$

Sample T-step trajectory from π_i

Get dataset $D_i = \{(s, \pi^*(s))\}$

Aggregate dataset $D \leftarrow D \cup D_i$

Train classifier $\hat{\pi}_{i+1}$ **on** D

[Stéphane Ross, Geoff J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In AI-Stats.]

Interactive Imitation Learning

with dagger

Uses an online queryable expert

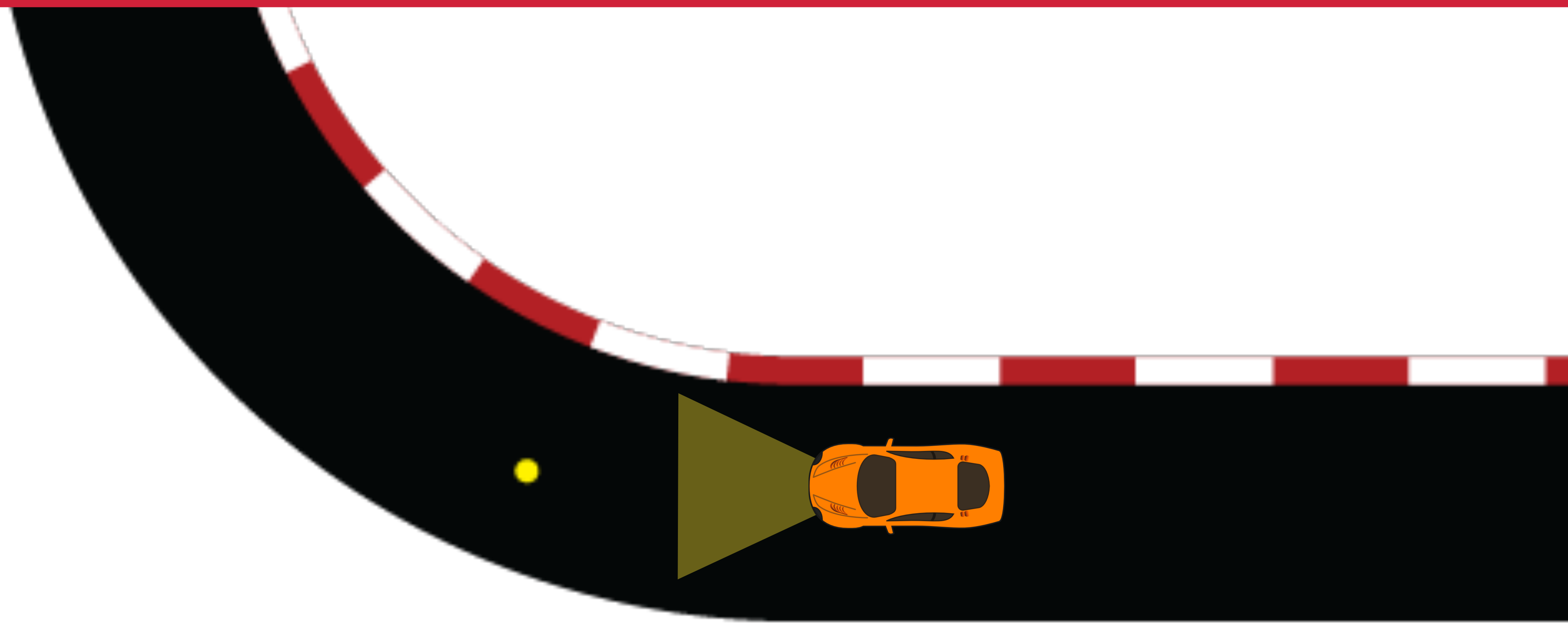
Advantage:

The agent can learn from its own state distribution.

	Supervised Learning	DAgger
Train	$(x, y) \sim D$	$(s, a) \sim d_\pi$
Test	$(x, y) \sim D$	$(s, a) \sim d_\pi$

Aggregate dataset $D \leftarrow D \cup D_i$

Train classifier $\hat{\pi}_{i+1}$ **on** D



[Stéphane Ross, Geoff J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In AI-Stats.]

Interactive Imitation Learning

with dagger

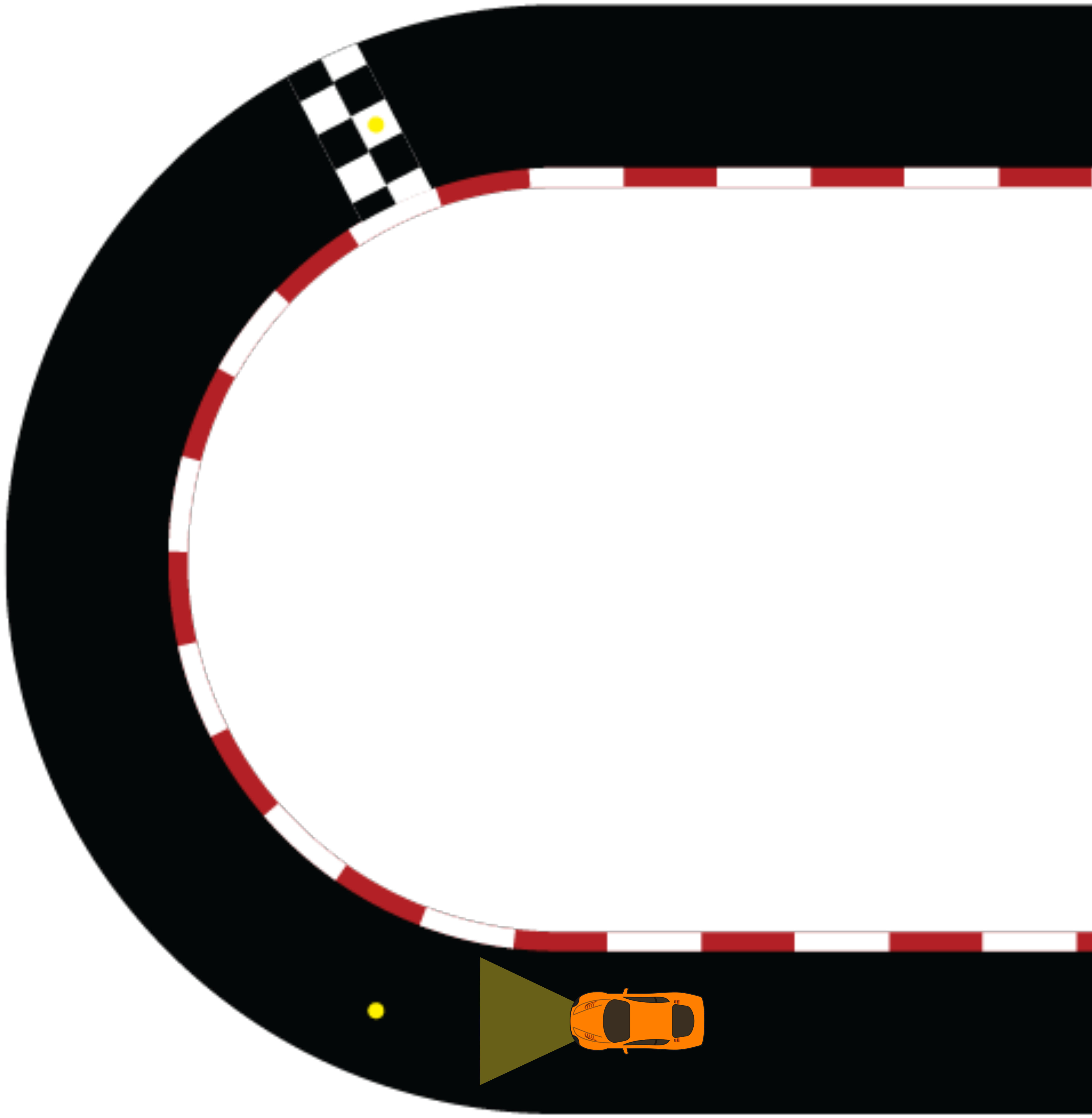
Uses an online queryable expert

Initialize Dataset D

Initialize $\hat{\pi}_1$

For $i = 1$ **to** N **do**

- $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$
- **Sample T-step trajectory from**
- **Get dataset** $D_i = \{(s, \pi^*(s))\}$
- **Aggregate dataset** $D \leftarrow D \cup D_i$
- **Train classifier** $\hat{\pi}_{i+1}$ **on** D



[Stéphane Ross, Geoff J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In AI-Stats.]

Interactive Imitation Learning

with dagger

Uses an online queryable expert

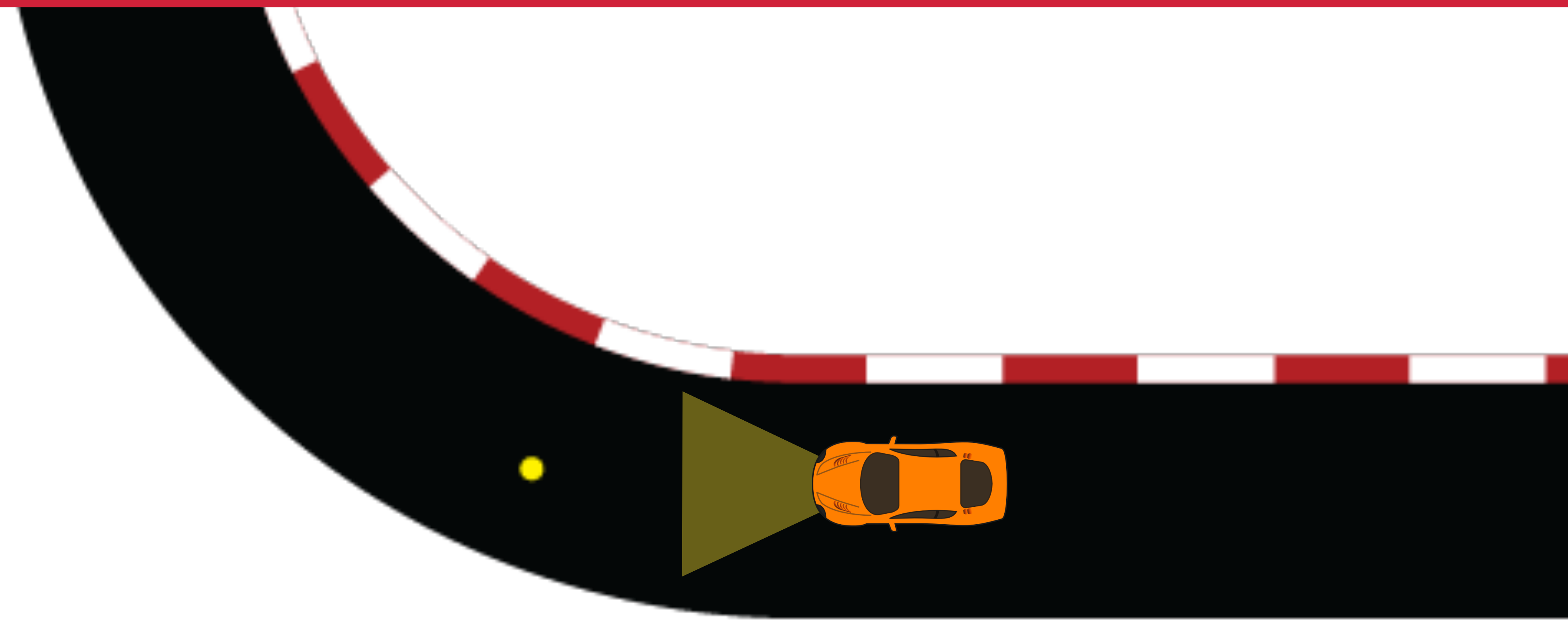
Advantage:

The agent can learn from its own state distribution.

	Supervised Learning	DAgger
Train	$(x, y) \sim D$	$(s, a) \sim d_\pi$
Test	$(x, y) \sim D$	$(s, a) \sim d_\pi$

Aggregate dataset $D \leftarrow D \cup D_i$

Train classifier $\hat{\pi}_{i+1}$ **on** D



[Stéphane Ross, Geoff J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In AI-Stats.]

Interactive Imitation Learning

with dagger

Uses an online queryable expert

Advantage:

The agent can learn from its own state distribution.

	Supervised Learning	DAgger
Train	$(x, y) \sim D$	$(s, a) \sim d_\pi$
Test	$(x, y) \sim D$	$(s, a) \sim d_\pi$

Disadvantage:

We query an online expert at every state visited to ask for a label (i.e. annotations in NLP).

[Stéphane Ross, Geoff J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In AI-Stats.]

Talk Overview

Background

- Behavior Cloning
- Interactive Imitation Learning with Dagger (Ross et al. 2011)

Modern Imi

- Uncerta

Research Question:

Can we design algorithms that deal with the covariate shift/exposure bias problem without needing an online queryable expert?

Disagreement Regularized Imitation Learning

Kianté Brantley,[1] Wen Sun,[3] Mikael Henaff [2]

[1] University of Maryland, [2] Facebook AI Research [3] Cornell University

Imitation Learning with Behavior Cloning

Issue:

The assumptions underlying supervised learning no longer hold, resulting in a covariate shift issue.

	Supervised Learning	Behavior Cloning
Train	$(x, y) \sim D$	$(s, a) \sim d_{\pi*}$
Test	$(x, y) \sim D$	$(s, a) \sim d_{\pi}$

[ALVINN: An Autonomous Land Vehicle in a Neural Network, Dean Pomerleau Neurips 1989]
[An Invitation to Imitation - Semantic Scholar, Bagnell]

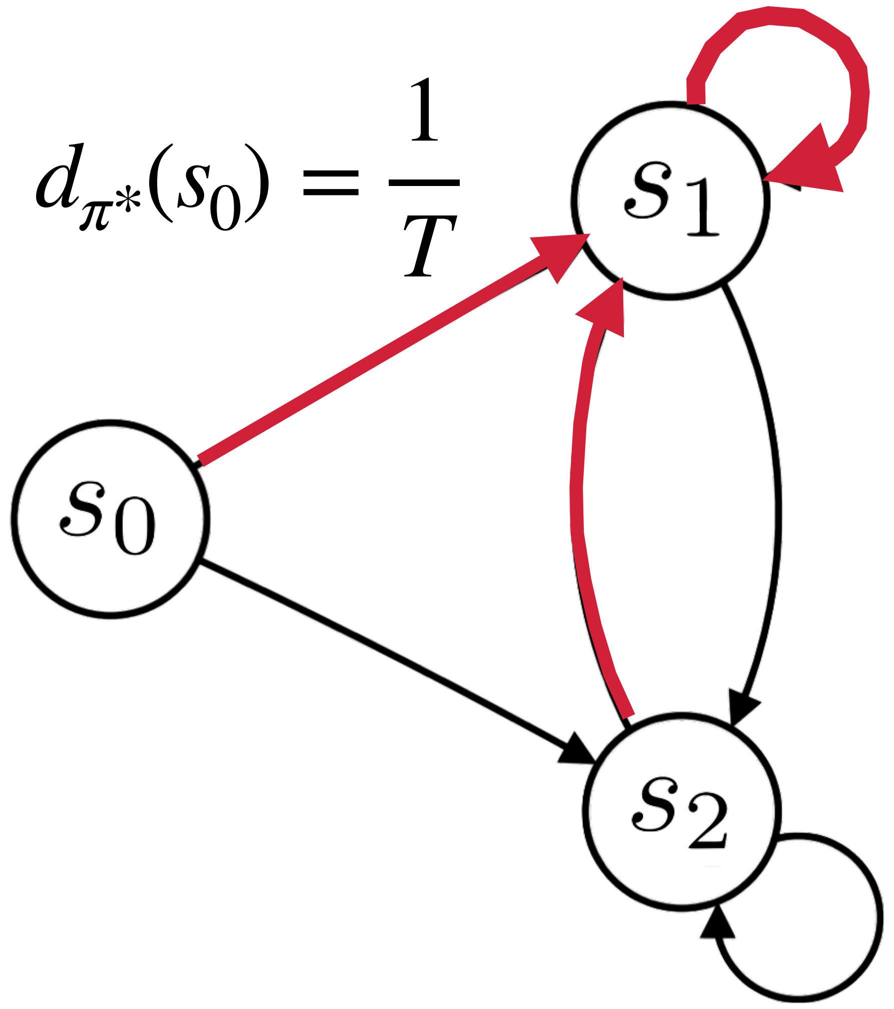
Formalizing

the covariate shift problem

Given an expert policy: π^*

$$d_{\pi^*}(s_1) = \frac{T-1}{T}$$

$$d_{\pi^*}(s_0) = \frac{1}{T}$$



Behavior Cloning Loss:

$$J_{BC}(\pi) = \epsilon$$

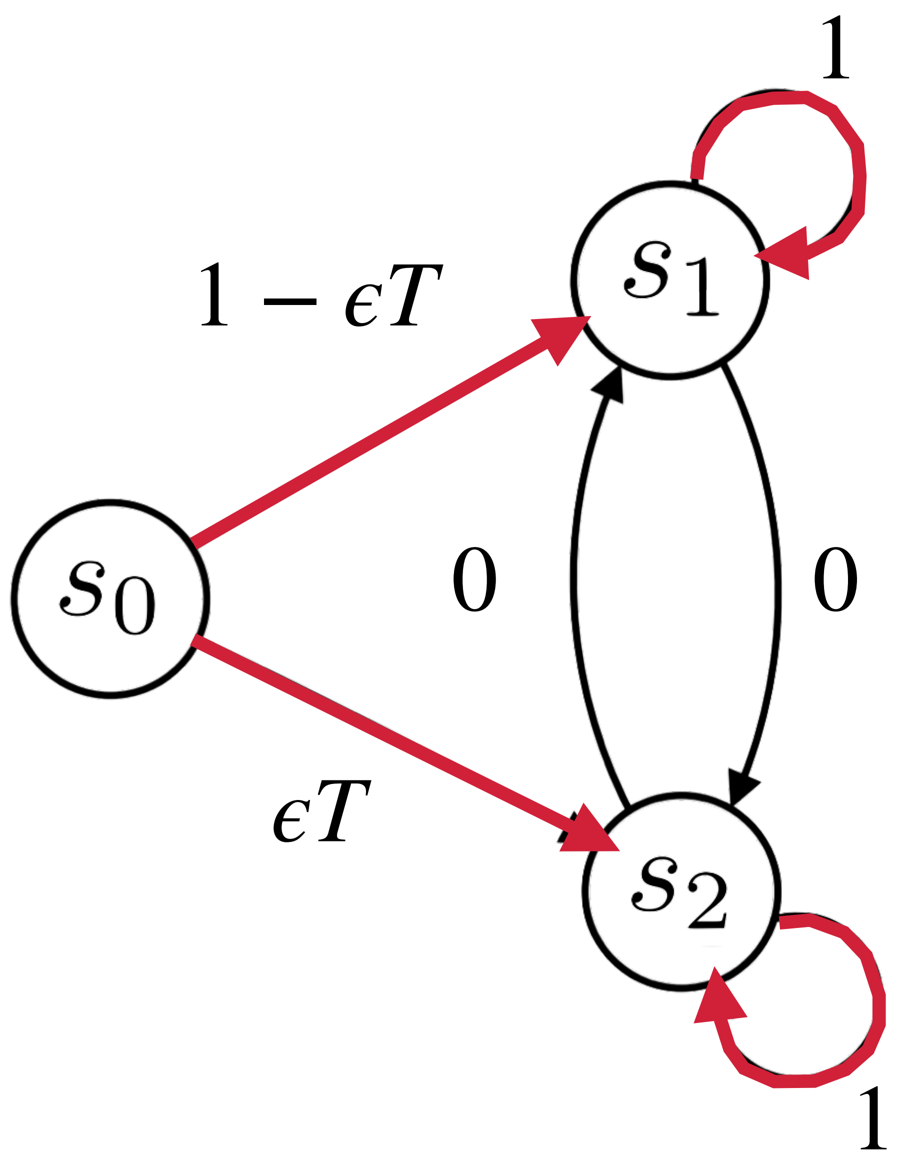
(loss is small)

Behavior Cloning Regret:

$$\text{Regret}(\hat{\pi}) = \mathcal{O}(\epsilon T^2)$$

(quadratic regret)

Consider a policy: $\hat{\pi}$



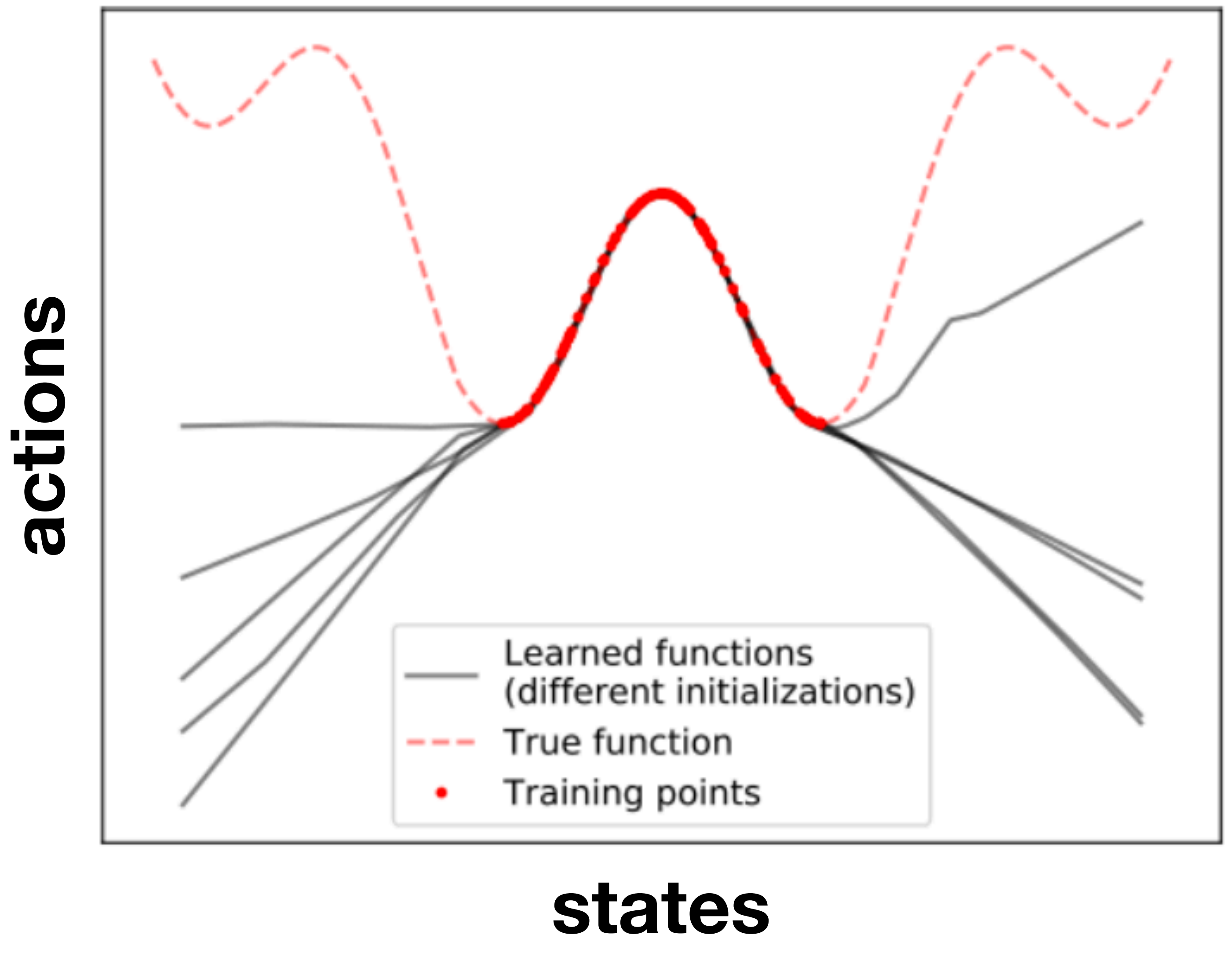
[Efficient Reductions for Imitation Learning, Ross & Bagnell, AISTATS 2010]
[Lower bounds for reductions, Matti Kääriäinen, Atomic Learning Workshop 2006]

Our Approach

dril

Motivation:
1. **Mimic expert within the expert distribution**
2. **Stay within the expert distribution**

$$J_{DRIL}(\pi) = J_{BC}(\pi) + J_U(\pi)$$



Train ensemble of polices $\Pi_E = \{\pi_1, \ldots, \pi_E\}$ **on demonstration data** D

Uncertainty Cost: $C_U(s,a) = \mathrm{Var}_{\pi \sim \Pi_E}(\pi(a \mid s))$

DRIL **cost can be optimized using any RL algorithm**

Our Approach

dril (final algorithm)

Input: Expert Demonstration data $D = \{(s_i, a_i)\}_{i=1}^{N}$

Train: Policy Ensemble $\Pi_E = \{\pi_1, \ldots, \pi_E\}$ using demonstration data D

Train: Policy behavior cloning π using demonstration data D

for $i = 1$ to $\ldots$ do

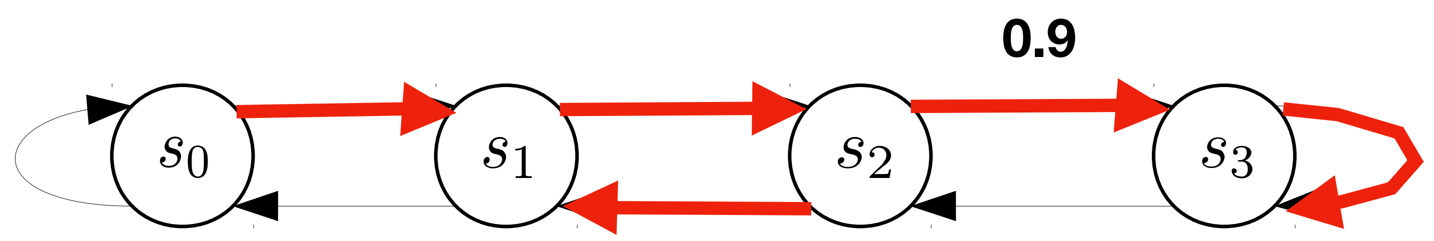
- Perform one gradient update to minimze $J_{BC}(\pi)$ using a minibatch from D
- Perform one step of policy gradient to minimize $\mathbf{E}_{s \sim d_\pi, a \sim \pi(\cdot|s)}\left[C_{\mathrm{U}}(s, a)\right]$

end for

Importance of J_{BC} update

counter example

Given an suboptimal expert policy: π^*



r(s_0) = 0 r(s_1) = 0 r(s_2) = 0 r(s_3) =1

Cost Function:

$$C_{\mathrm{U}}^{\hat{\pi}}(s,a) = \mathrm{Var}_{\pi \sim \Pi_{\mathrm{E}}}(\pi(a\,|\,s))$$

Without bootstrapping

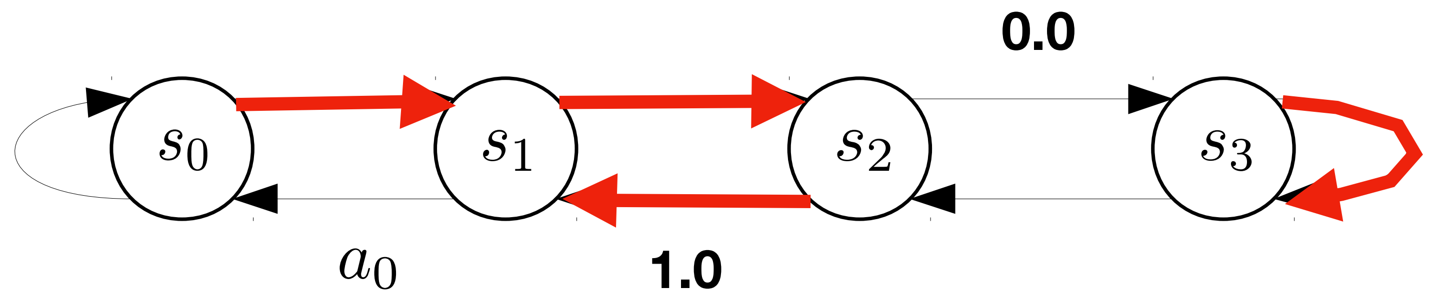
$$C_{\mathrm{U}}^{\hat{\pi}_1}(s,a) \approx C_{\mathrm{U}}^{\hat{\pi}_2}(s,a)$$

Behavior Cloning:

Smaller J_{BC} is closer to π^*

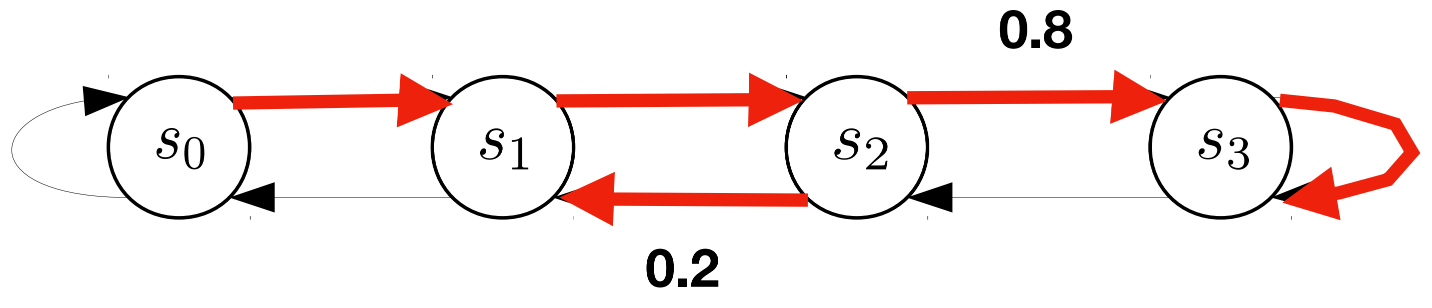
$$J_{BC}(\hat{\pi}_1) > J_{BC}(\hat{\pi}_2)$$

Consider a policy: $\hat{\pi}_1$



r(s_0) = 0 r(s_1) = 0 r(s_2) = 0 r(s_3) =1

Consider a policy: $\hat{\pi}_2$



r(s_0) = 0 r(s_1) = 0 r(s_2) = 0 r(s_3) =1

Our Approach

dril (analysis)

Assumption 1: (Realizability) $\pi^* \in \Pi$

Assumption 2: (Optimization Oracle) $J(\hat{\pi}) \leq \operatorname{argmin}_{\pi \in \Pi} J(\pi) + \epsilon$

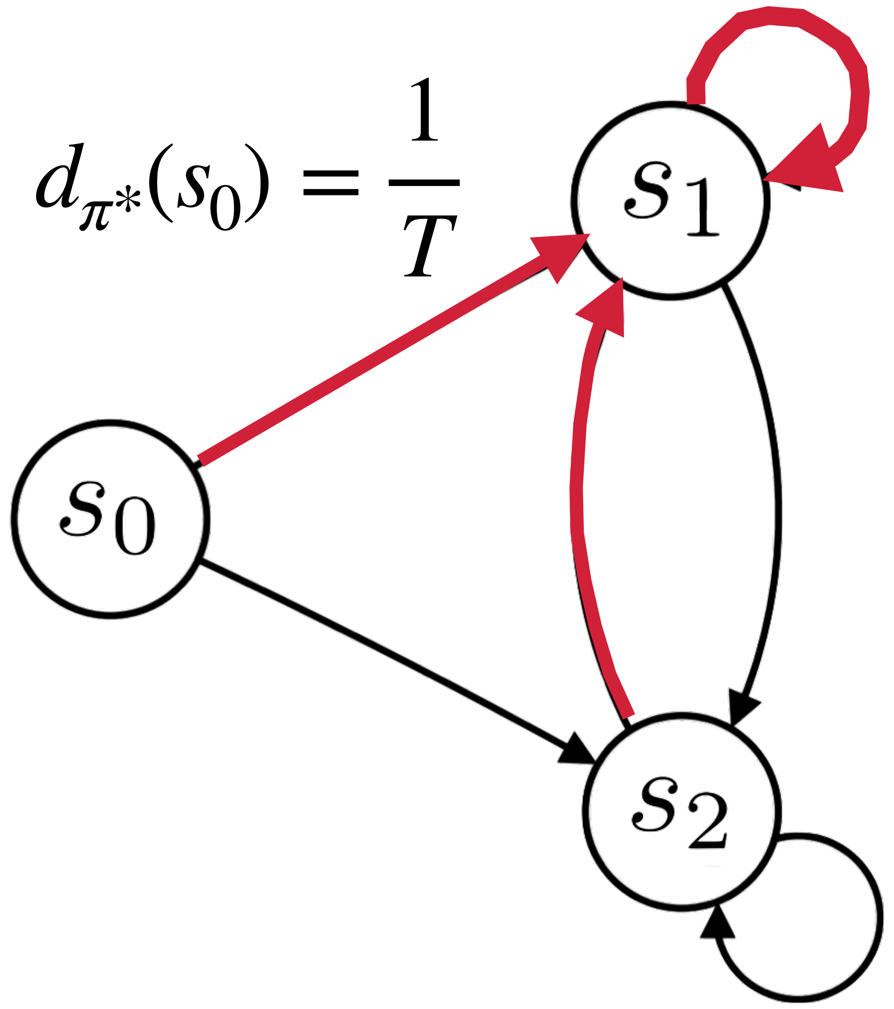
Assumption 3: (Smoothness on true Q-Function) $Q^{\pi^*}(s, a) - Q^{\pi^*}(s, \pi^*) \leq u$

Theorem (informal): $J_{DRIL}(\pi)$ **has regret** $\mathcal{O}(\epsilon \kappa T)$

Given an expert policy: π^*

$$d_{\pi^*}(s_1) = \frac{T-1}{T}$$

$$d_{\pi^*}(s_0) = \frac{1}{T}$$



Behavior Cloning Regret:

$$\text{Regret}(\hat{\pi}) = \mathcal{O}(\epsilon T^2)$$

(quadratic regret)

DRIL Regret $\mathcal{O}(\epsilon \kappa T)$

$$\kappa = 1 + \mathcal{O}\left(\frac{1}{\sqrt{|\text{ensemble}|}}\right)$$

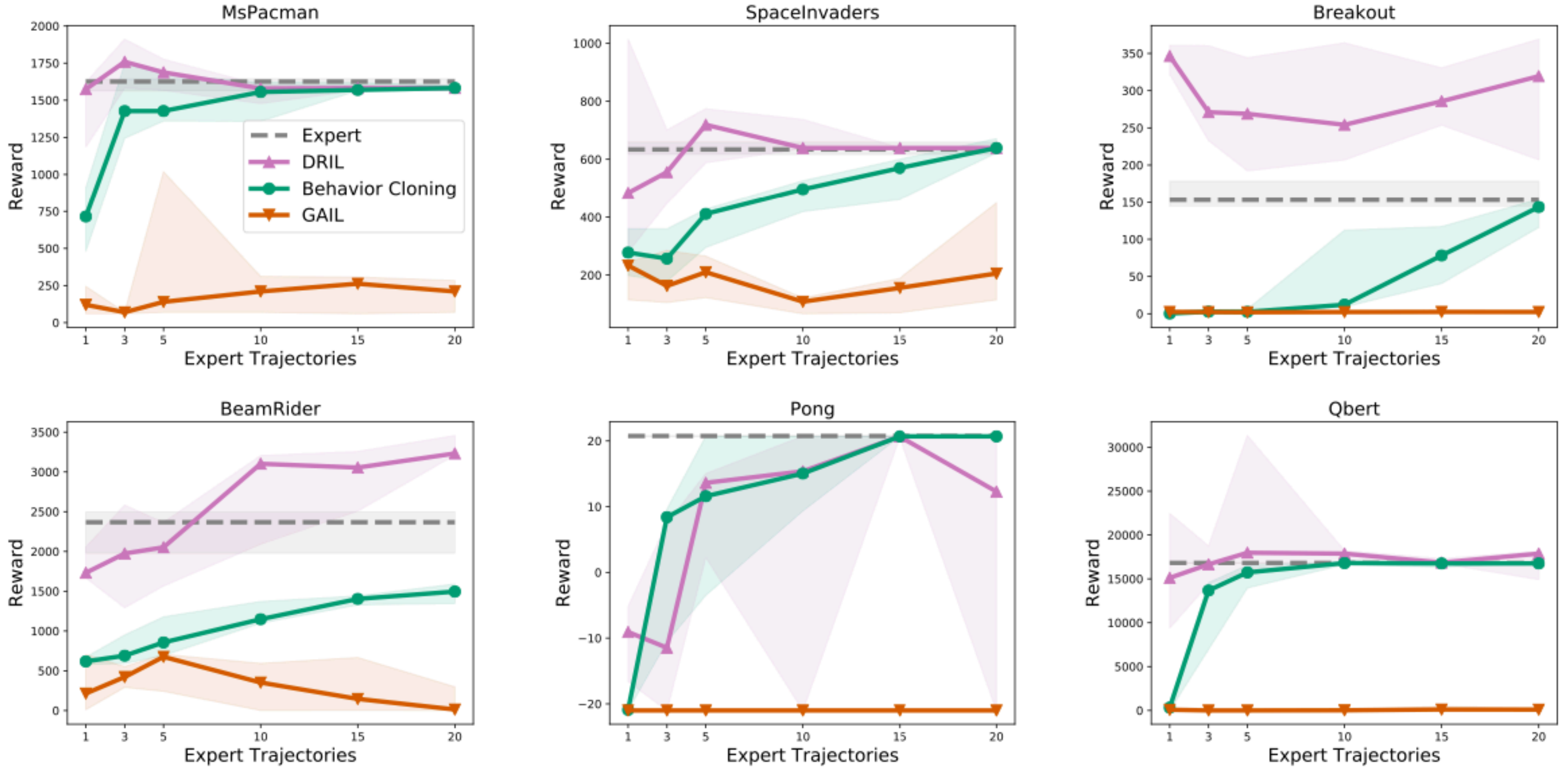
DRIL Regret:

$$\text{Regret}(\hat{\pi}) = \mathcal{O}(\epsilon T)$$

(linear regret)

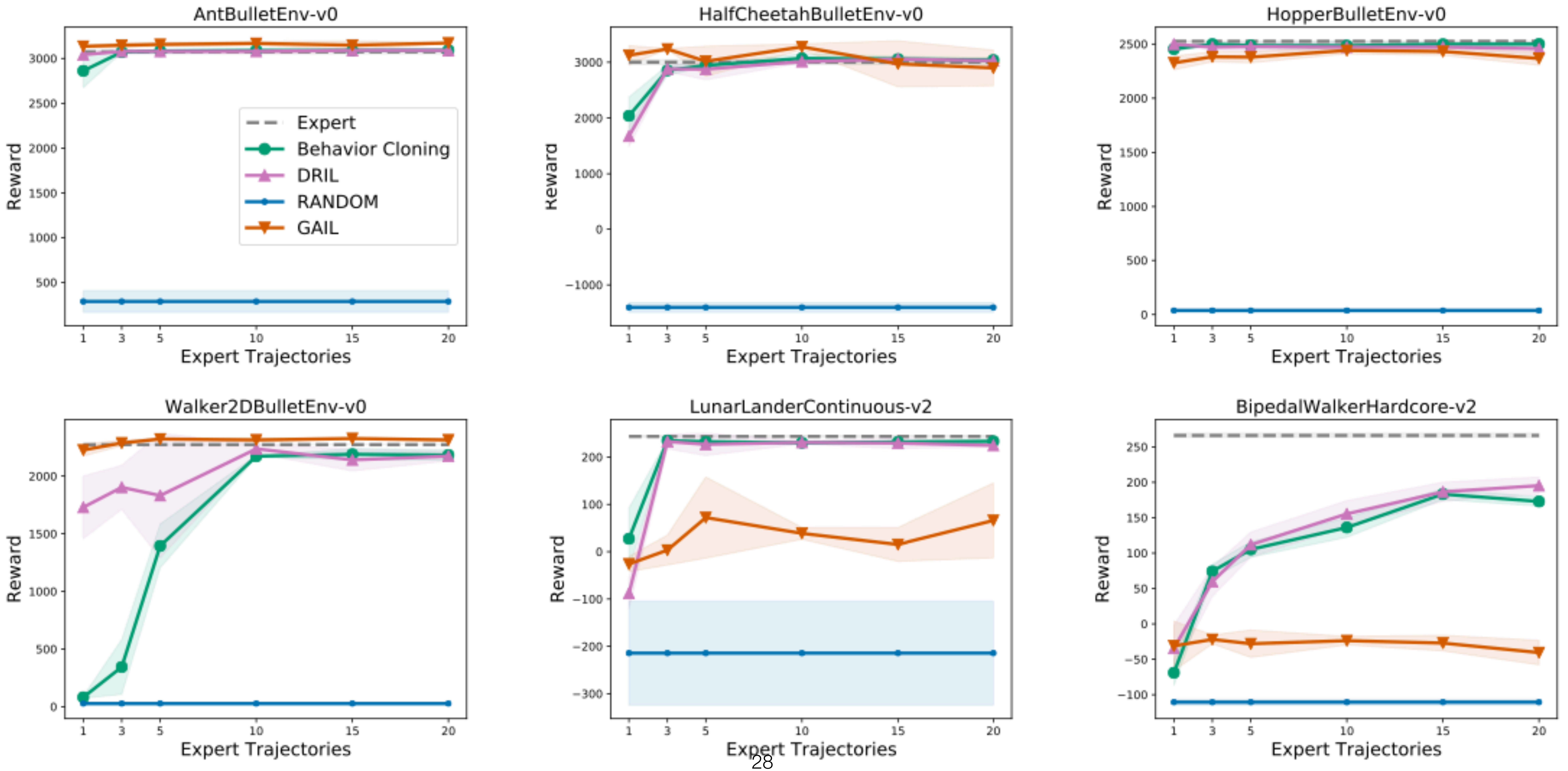
Experiments

atari



Experiments

continuous control



Summary

- The covariate shift problem has been a fundamental issue in imitation learning
- We provide a new algorithm to address the covariate shift problem
- Our algorithm has good empirical results and theoretical guarantees in some settings
- We provide a counter example showing the importance of interleaving J_{BC} updates

Talk Overview

Background

- Behavior Cloning
- Interactive Imitation Learning with Dagger (Ross et al. 2011)

Modern Imitatio

- Uncertainty
- An Empirica

Research Question:

Does performing behavior cloning updates help in similar style algorithms as DRIL?

An Empirical Study of Imitation Learning

Kianté Brantley[1]

[1] University of Maryland

Motivation

large-scale structured-prediction for nlp

A Deep Reinforced Model For Abstractive Summarization, Paulus et al. 2017 Cited by 1100

$$L_{mixed} = \gamma L_{rl} + (1-\gamma)L_{ml}$$

" …. optimizing ROUGE does not guarantee an increase in quality …. output. It is possible to …. increase their score without an actual increase in readability or relevance" Paulus et al. 2017

r(s,a) = ROUGE-L

Googles's Neural Machine Translation System: Bridging, the Gap between Human and Machine Translation Wu et al. 2018 Cited by 4842

$$\mathcal{O}_{mixed}(\theta) = \alpha * \mathcal{O}_{ML} + \mathcal{O}_{RL}$$

stabilize training

r(s,a) = GLEU

∑, Li et al. 2017 Cited by 807

"…. final training alternately update the … using the adversarial objective and the MLE objective " Li et al. 2017

stabilize training

r(s,a) = D(s,a) (similar to GAIL)

Deep reinforcement learning for dialogue generation, Li et al. 2016 Cited by 1039

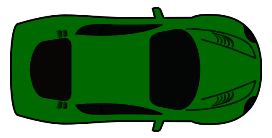
"for every sequence of length T we use the MLE loss for the first L tokens and the reinforcement algorithm for the remaining T – L tokens" Li et al. 2016

following previous work

r(s,a) = Fixed pertained models

Modern Imitation Learning

basics

Expert/Oracle Demonstrator 

- state
- actions

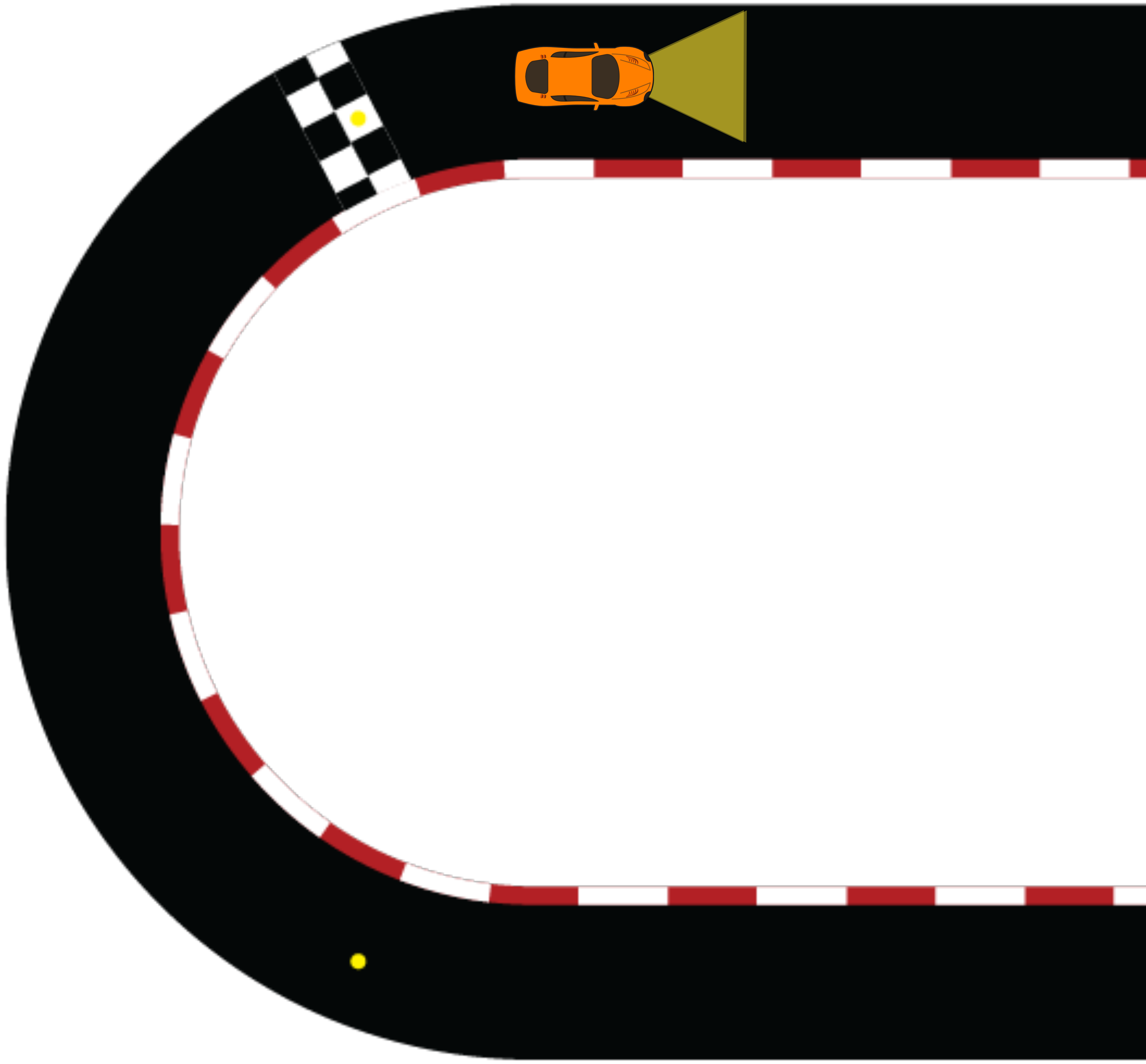
Training set:

D = {(state, actions)} from expert π^*

Goal:

learn reward function $r(s, a)$ using D

learn an agent π_θ by maximizing $r(s, a)$ with RL



Note: These objectives studied in this paper are the dual of inverse reinforcement learning objectives

Modern Imitation Learning

baselines

Behavior Cloning (bc)

Modern Imitation Learning

baselines

Behavior Cloning (bc)

Lazy Learners

- k-nearest neighbor (knn)

Modern Imitation Learning

baselines

Behavior Cloning (bc)

Lazy Learners

- k-nearest neighbor (knn)
- Locally weighted Learning (lwl)

Modern Imitation Learning

baselines

Behavior Cloning (bc)

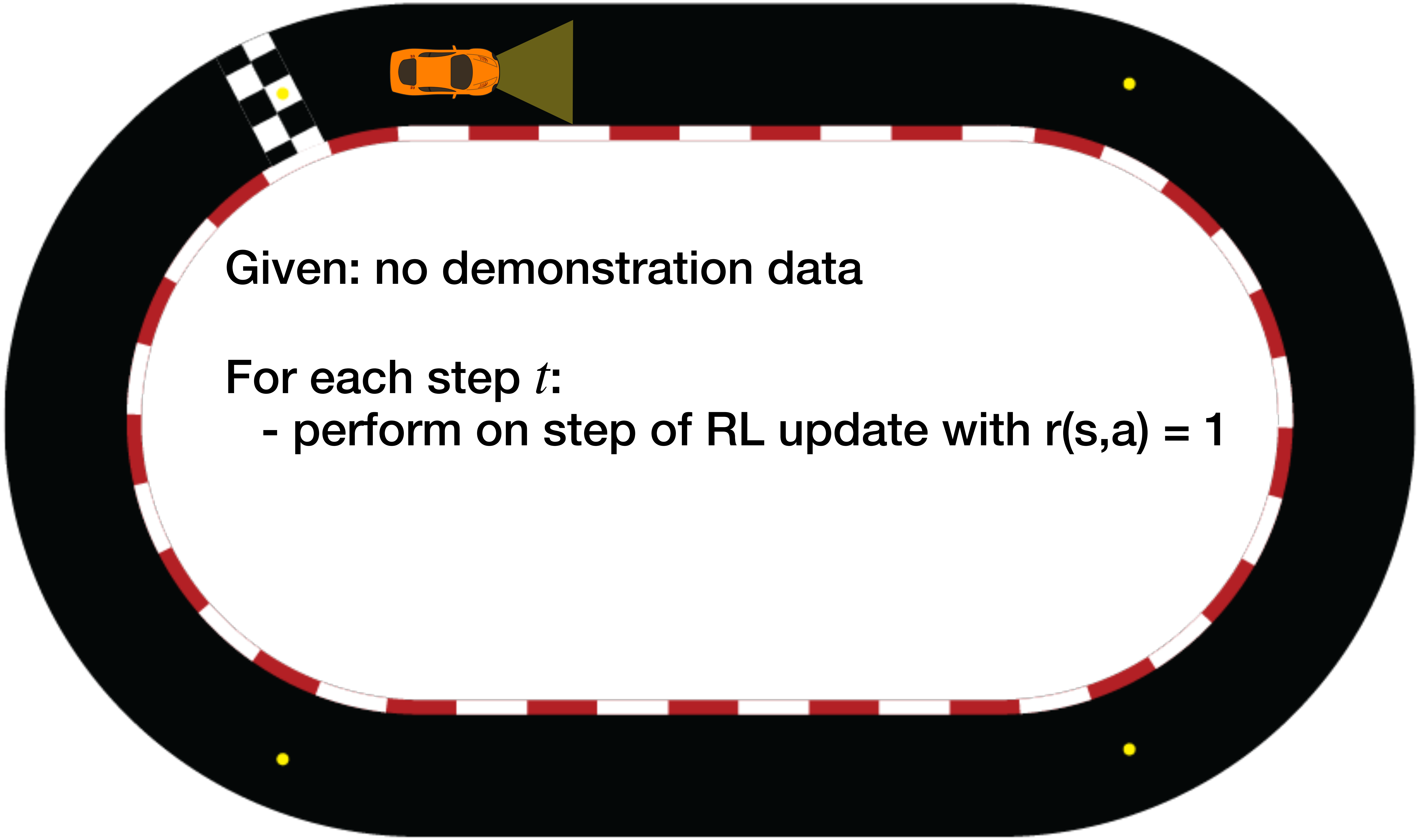
Lazy Learners

- k-nearest neighbor (knn)
- Locally weighted Learning (lwl)

Constant Reward (cr)

Modern Imitation Learning

constant reward (cr)



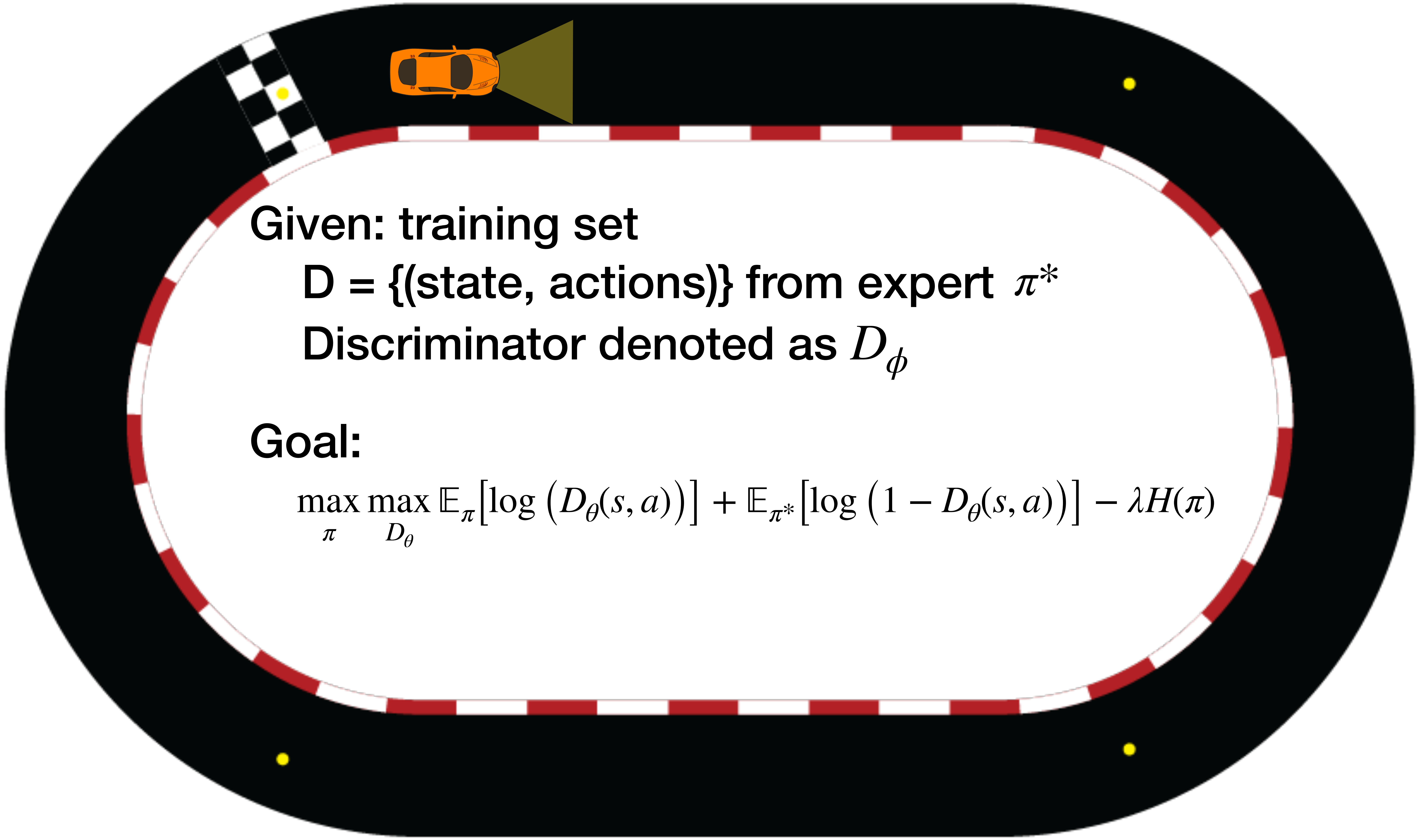
Given: no demonstration data

For each step t:
- perform on step of RL update with r(s,a) = 1

Modern Imitation Learning

algorithms

General Adversarial Imitation Learning (gail)



Given: training set

D = {(state, actions)} from expert $\pi*$

Discriminator denoted as D_ϕ

Goal:

$$\max_{\pi} \max_{D_\theta} \mathbb{E}_\pi\left[\log\left(D_\theta(s,a)\right)\right] + \mathbb{E}_{\pi*}\left[\log\left(1 - D_\theta(s,a)\right)\right] - \lambda H(\pi)$$

Modern Imitation Learning

general adversarial imitation learning (gail)

Given: training set

D = {(state, actions)} from expert π^*

Discriminator denoted as D_ϕ

For each step t:

- update D_θ with binary classifier
- perform one RL update with

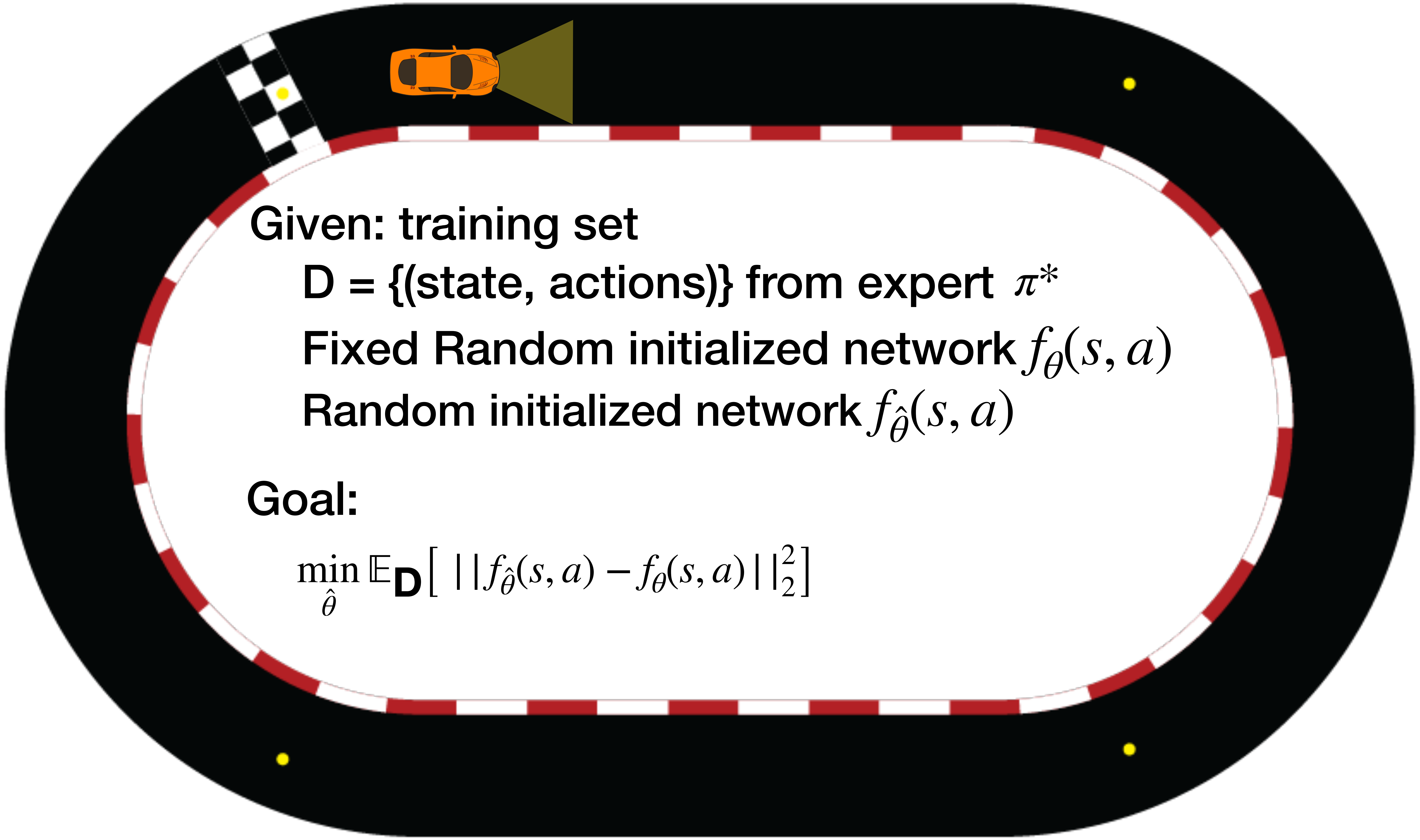
$$r(s, a) = -\log\left(D_\theta(s, a)\right)$$

Modern Imitation Learning

- algorithms
- General Adversarial Imitation Learning (gail)
- Adversarial Imitation Learning (airl)
- Random Expert Distillation (red)

Modern Imitation Learning

random expert distillation (red)



Given: training set

D = {(state, actions)} from expert π^*

Fixed Random initialized network $f_\theta(s, a)$

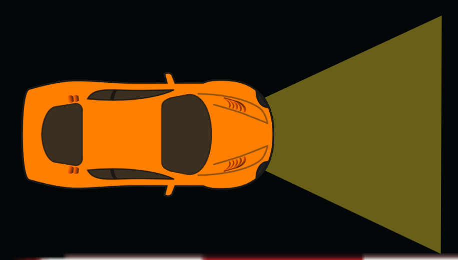
Random initialized network $f_{\hat{\theta}}(s, a)$

Goal:

$$\min_{\hat{\theta}} \mathbb{E}_{\mathbf{D}}\left[\ ||f_{\hat{\theta}}(s,a) - f_\theta(s,a)||_2^2\right]$$

Modern Imitation Learning

random expert distillation (red)



Given: training set

D = {(state, actions)} from expert π^*

Fixed Random initialized network $f_\theta(s, a)$

Random initialized network $f_{\hat{\theta}}(s, a)$

For each step t:

- perform one RL update with

$$r(s,a) = -\left(|| f_{\hat{\theta}}(s,a) - f_\theta(s,a) ||_2^2 \right)$$

Modern Imitation Learning

algorithms

General Adversarial Imitation Learning (gail)

- RL updates

Adversarial Imitation Learning (airl)

- RL updates

Random Expert Distillation (red)

- RL updates

Disagreement-regularized imitation learning (dril)

- Interleave RL updates with BC updates
- Importance of interleaving BC updates

Modern Imitation Learning

baselines

Behavior Cloning (bc)

Lazy Learners

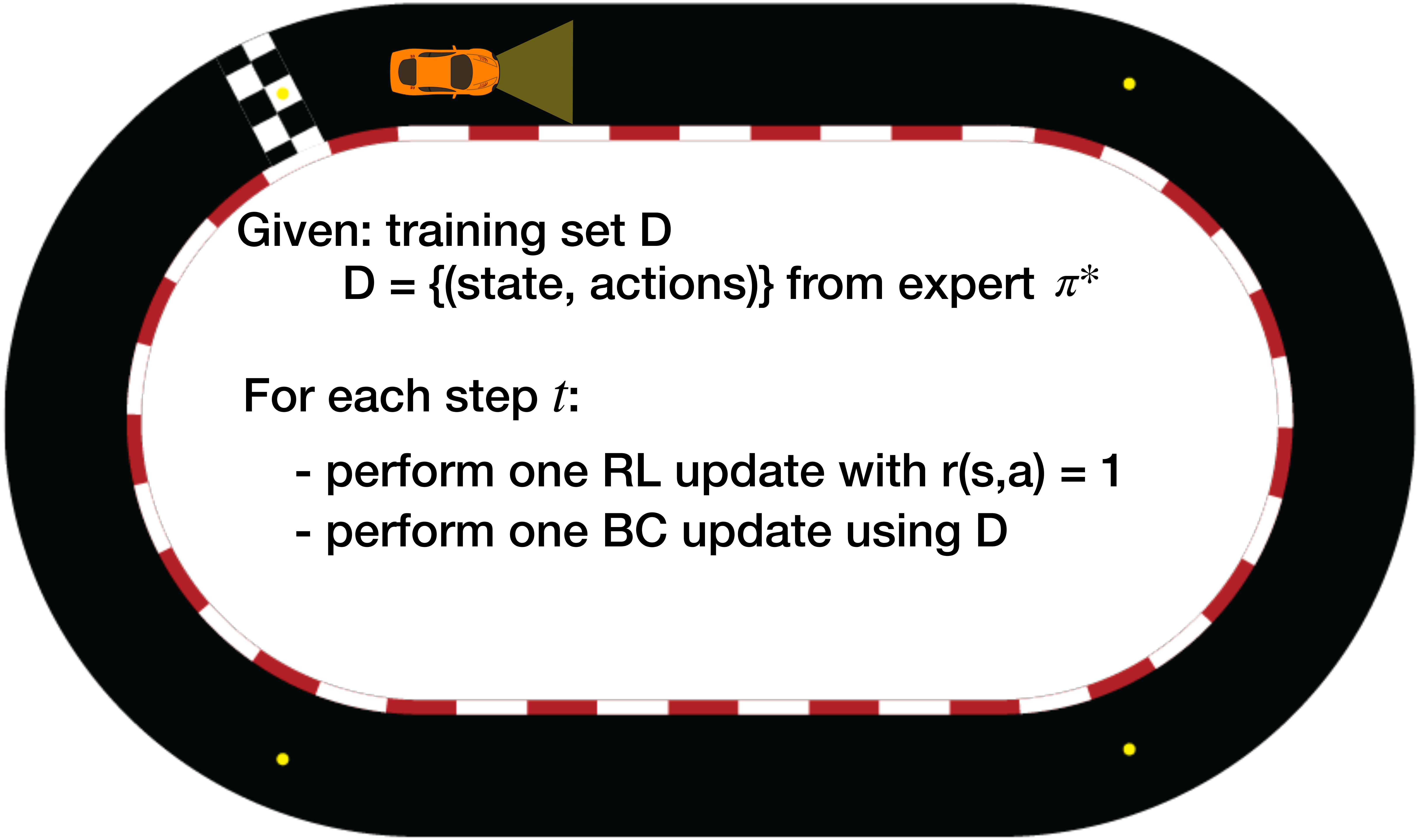
- k-nearest neighbor (knn)
- Locally weighted Learning (lwl)

Constant Reward

- constant Reward (cr)
- behavior cloning -regularized constant Reward (bc-cr)

Modern Imitation Learning

behavior cloning -regularized constant reward (bc-cr)



Given: training set D

D = {(state, actions)} from expert π^*

For each step t:

- perform one RL update with r(s,a) = 1
- perform one BC update using D

Modern Imitation Learning

- algorithms
- General Adversarial Imitation Learning (gail)
- Adversarial Imitation Learning (airl)
- Random Expert Distillation (red)
- Disagreement-regularized imitation learning (dril)
- Behavior Cloning -regularized General Adversarial Imitation Learning (bc-gail)
- Behavior Cloning -regularized Adversarial Imitation Learning (bc-airl)
- Behavior Cloning -regularized Random Expert Distillation (bc-red)

Modern Imitation Learning

- algorithms
- Behavior Cloning (bc)
- k-nearest neighbor (knn)
- Locally weighted Learning (lwl)
- Constant reward (cr)
- behavior cloning -regularized constant Reward (bc-cr)
- General Adversarial Imitation Learning (gail)
- Adversarial Imitation Learning (airl)
- Random Expert Distillation (red)
- Disagreement-regularized imitation learning (dril)
- Behavior Cloning -regularized General Adversarial Imitation Learning (bc-gail)
- Behavior Cloning -regularized Adversarial (bc-airl)
- Behavior Cloning -regularized Random Expert Distillation (bc-red)

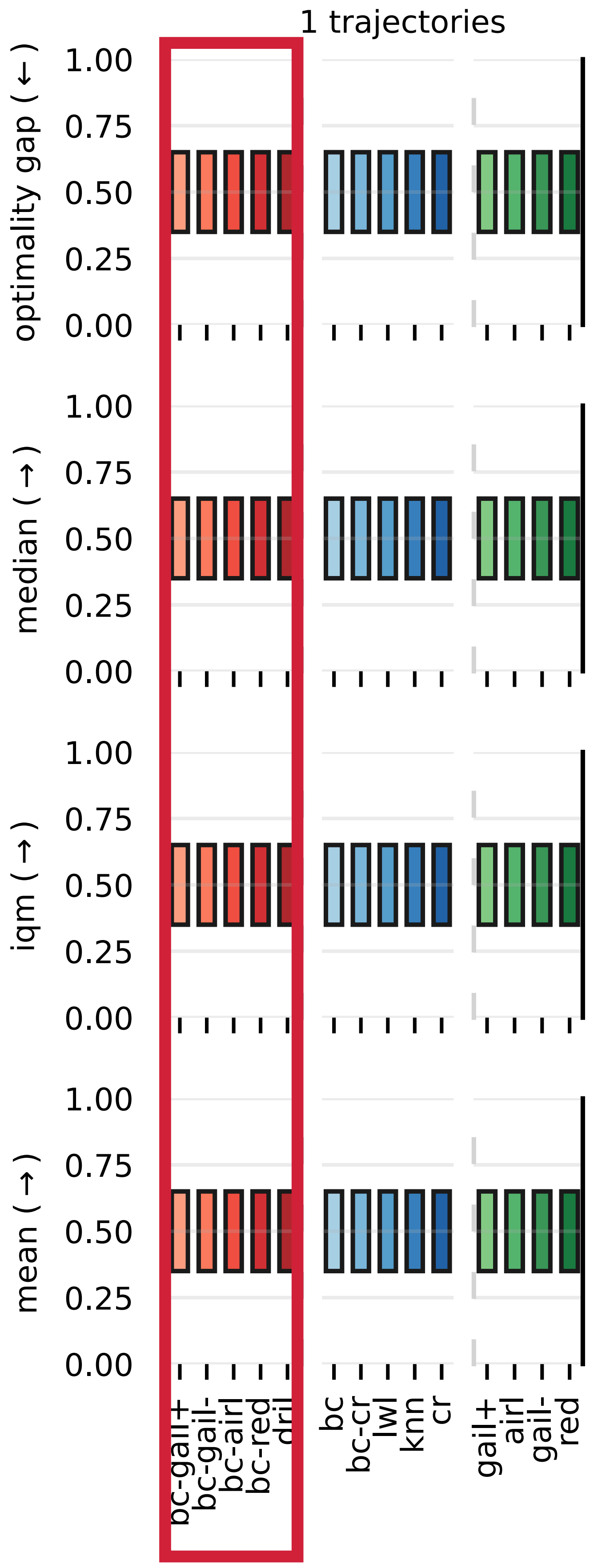
Modern Imitation Learning

tasks

Task	Feature-Based	Pixel-Based	Structure-Prediction
Toolkits	Mujoco, Pybullet	DMC, Box2D	NLPGYM
Trajectories	[1,2,3,5,10]	[1,3,5,10]	entire dataset
# Environments	25	6	5
# Seeds	5	5	5
# Experiments per task	(5*25*5) = 625	(4*6*5)=120	(1*5*5)=25
# Action Space	continuous	continuous	discrete
# Observation Space	state features	pixels	word embeddings
# Dyamics	deterministic	deterministic	deterministic

Experiments

setup



- Optimality gap is an alternative to mean which measures the amount an algorithm fails to meet a minimum score of γ
- Median is the middle score of a order list fo task scores, but is a poor indicator of overall performance
- IQM is an alternative to median, discarding the bottom and top 25% runs and using the remaining 50% runs to calculate the mean score
- Mean is average score of a task across 5 runs, but is often dominated by performance of outlier tasks
- Red is imitation Learning algorithms that interleave behavior cloning updates
- Blue is baseline algorithms
- Green is imitation Learning algorithms that ***do not*** interleave behavior cloning updates

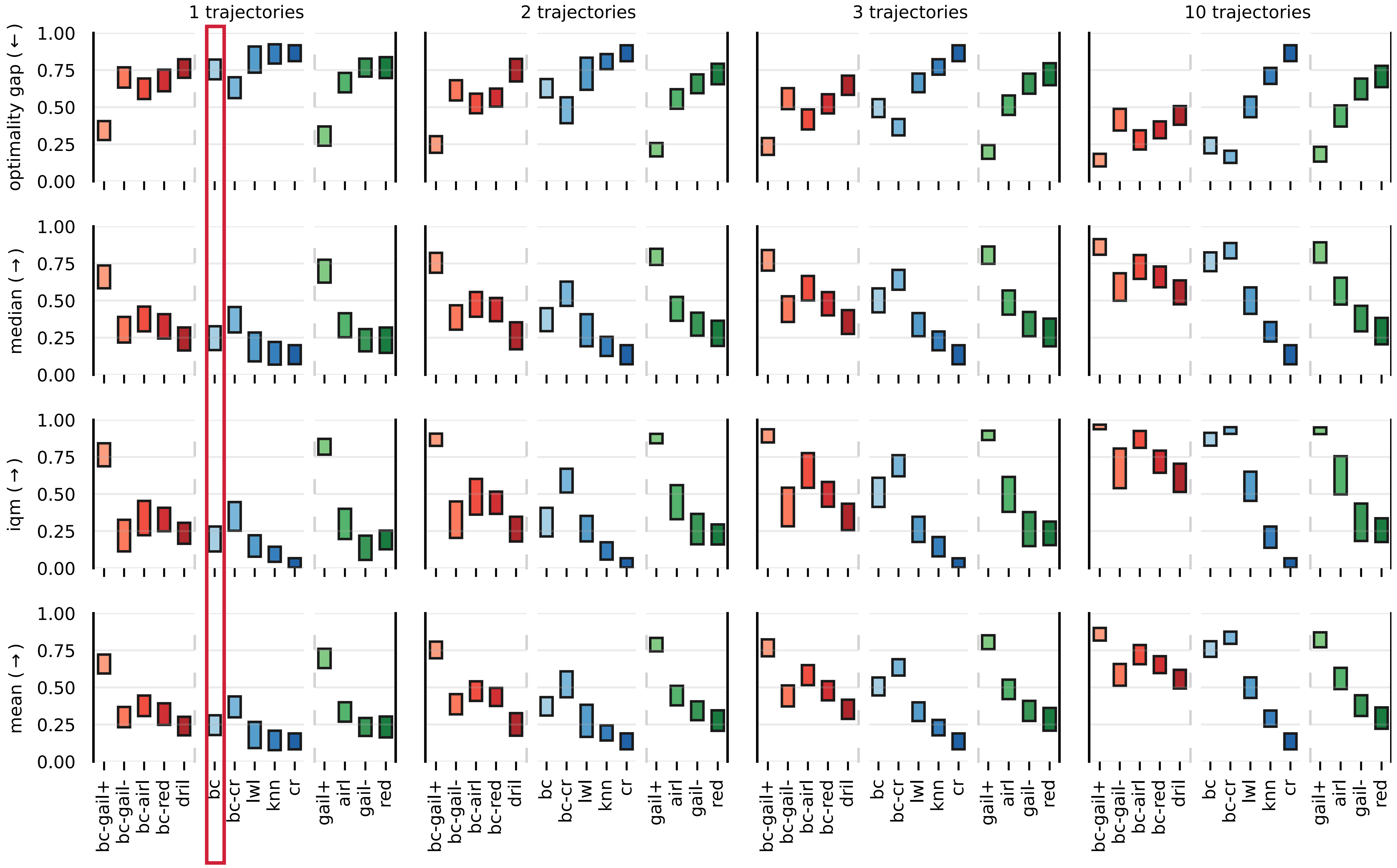
[Deep Reinforcement Learning at the Edge of the Statistical Precipice, Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, Marc G. Bellemare]

Note:

Practitioners artificially subsample states in trajectories to make behavior cloning perform worse, to create a gap between the performance of expert and behavior cloning.

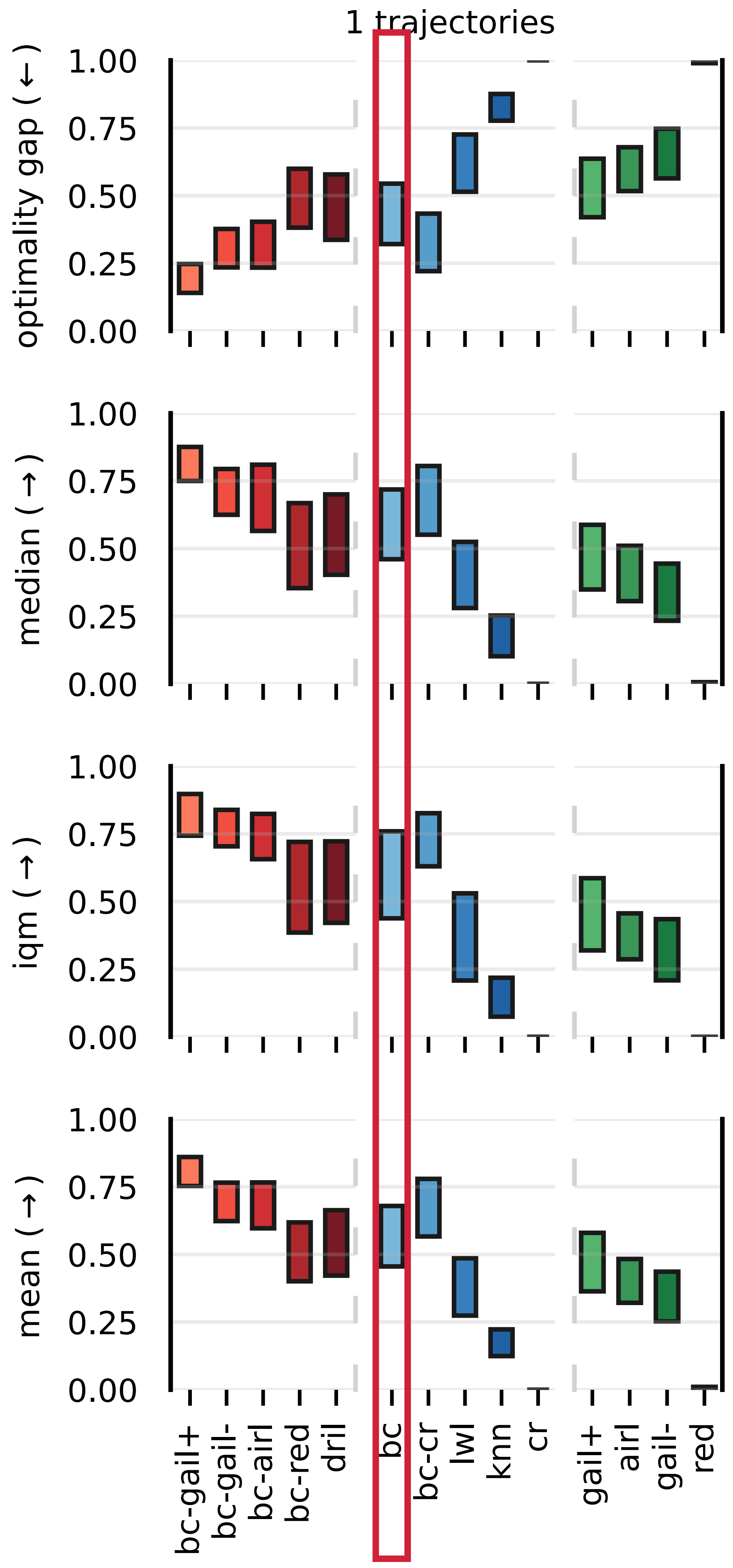
Experiments

featured-based **subsampled** tasks



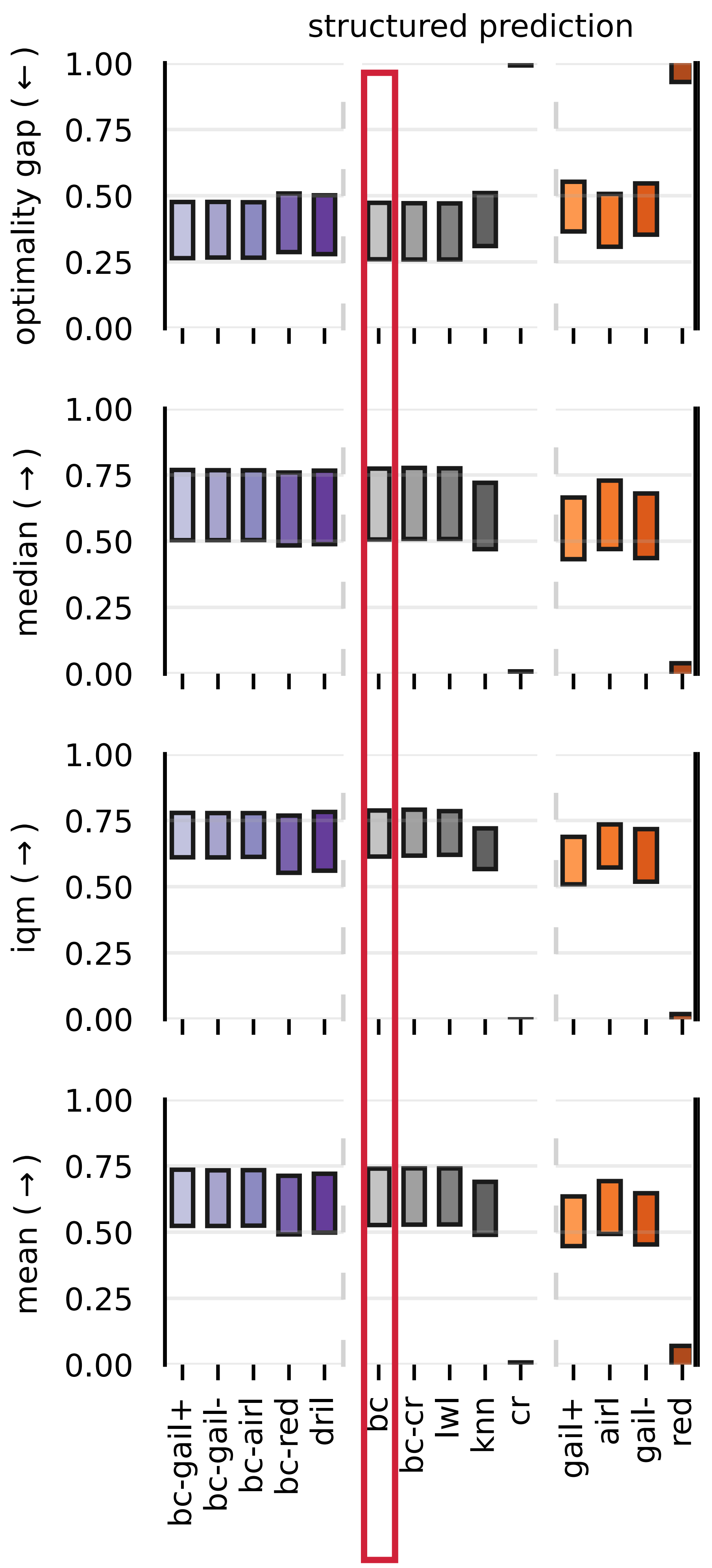
Experiments

pixel-based tasks



Experiments

structured-prediction tasks



Modern Imitation Learning

takeaways

Behavior cloning is a very strong baseline

Interleaving behavior cloning updates improve performance agnostic of any task and any algorithm

Interleaving behavior cloning updates improve performance in modern nlp structured prediction task

Summary

Studying issues that arise when solving sequential decision-making problems with expert demonstration data is important

We performed a thorough empirical comparison of all algorithms

We relate modern imitation learning algorithms to modern large-scale nlp structured prediction algorithms

Future Work

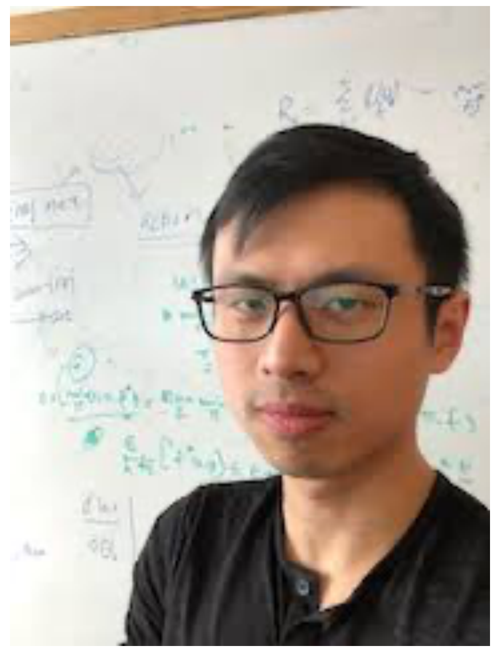
Active Reward-Learning Imitation Learning

Modern Imitation Learning for large-scale NLP structure prediction problems

Thank You

Questions?

Collaborators:



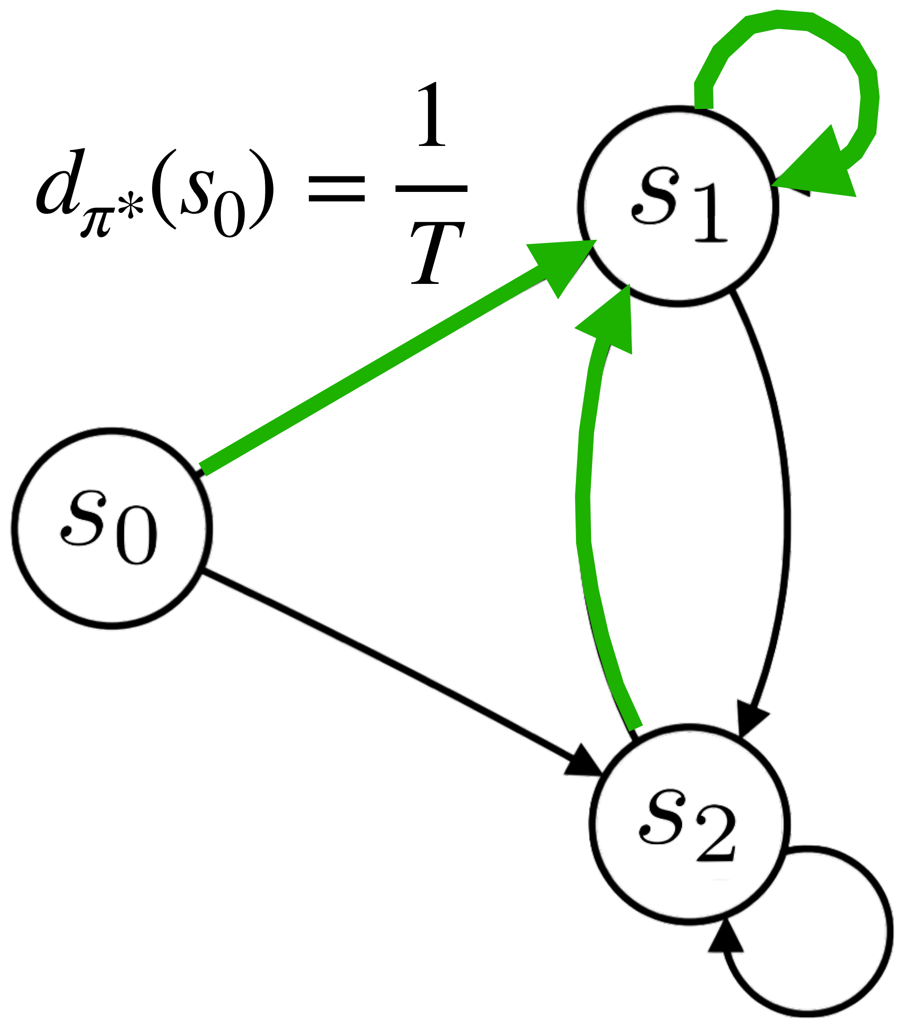
Wen Sun

Mikael Henaff

Formalizing the Behavior Cloning Issue

Given an expert policy: π^*

$$d_{\pi^*}(s_1) = \frac{T-1}{T}$$



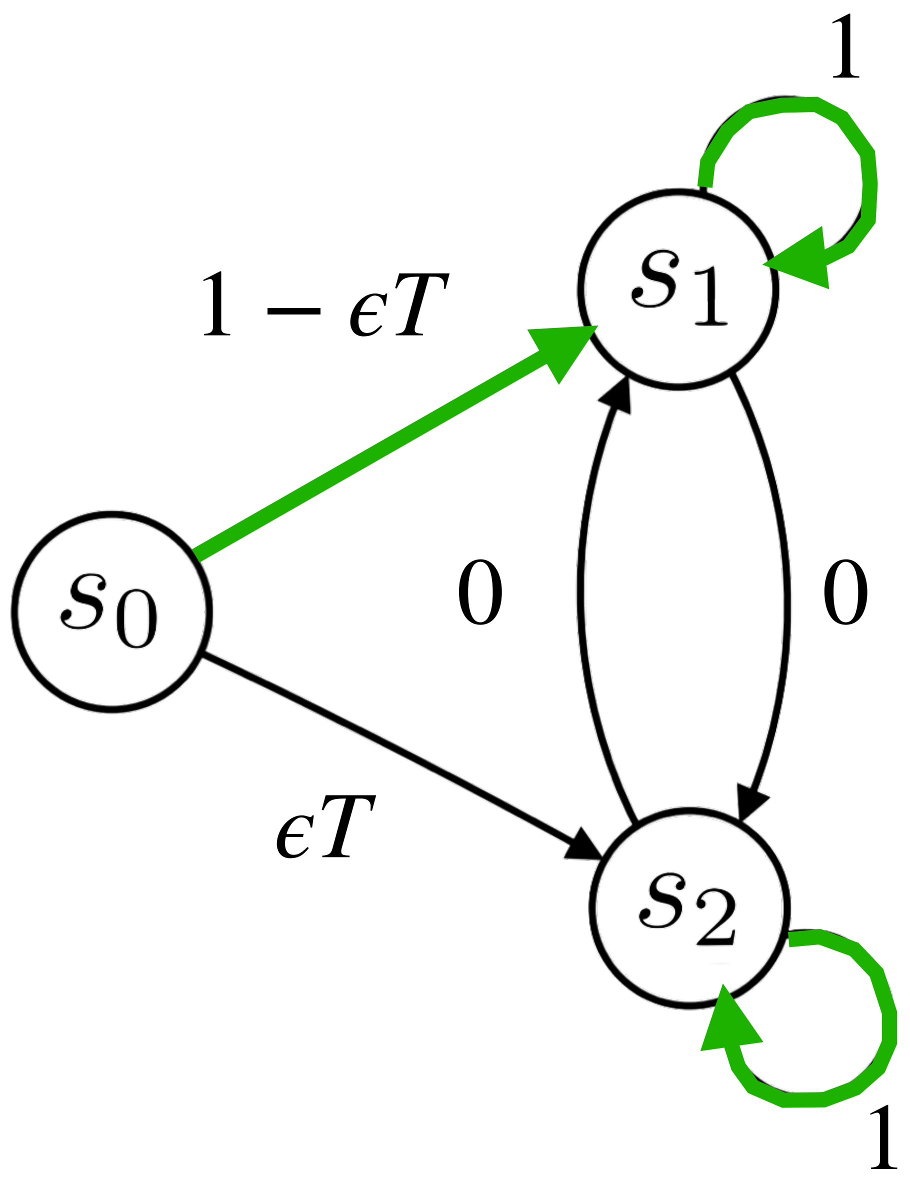
Behavior Cloning Loss:

$$J_{BC}(\pi) = \mathbf{E}_{s \sim d_{\pi^*}}\left[\ell(\pi(s), \pi^*(s))\right]$$

$$\begin{aligned} J_{BC}(\pi) = &\, d_{\pi^*}(s_0)\ell(\hat{\pi}(s_0), \pi^*(s_0)) \\ &+ d_{\pi^*}(s_1)\ell(\hat{\pi}(s_1), \pi^*(s_1)) \\ &+ d_{\pi^*}(s_2)\ell(\hat{\pi}(s_2), \pi^*(s_2)) \end{aligned}$$

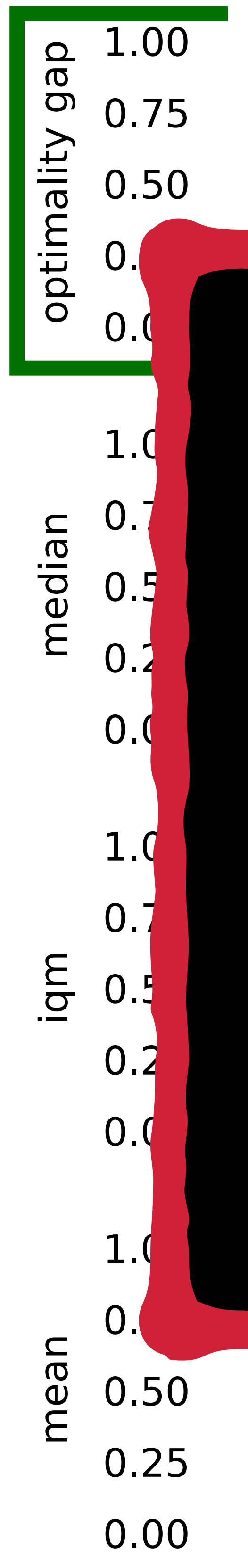
$$J_{BC}(\pi) = \frac{1}{T} * \epsilon T = \epsilon$$

Consider a policy: $\hat{\pi}$



Experiments

featured-based tasks

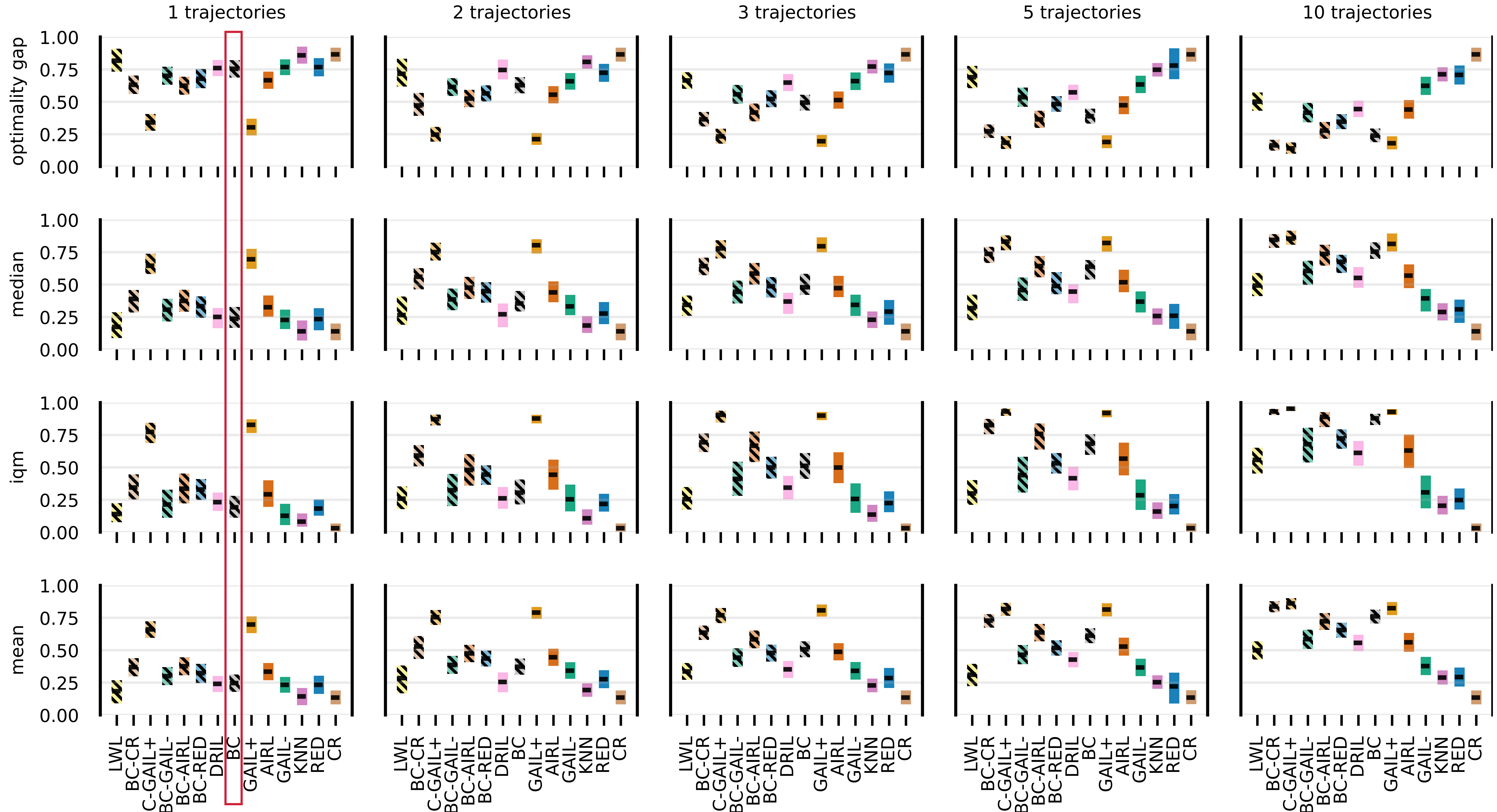


Note:

Practitioners artificially subsample states in trajectories to make behavior cloning perform worse, to create a gap between the performance of expert and behavior cloning.

Experiments

featured-based **subsampled** tasks



Experiments

pixel-based tasks