Behavior Cloning, Feedback and Covariate Shift (Part 2)

Sanjiban Choudhury

Cornell Bowers CIS Computer Science

Today's class

- ☐ Feedback drives Covariate Shift
- ☐ BC has a performance gap of $O(\epsilon T^2)$
- ☐ Easy vs Hard Regimes in Imitation Learning

Feedback drives covariate shift

An old problem

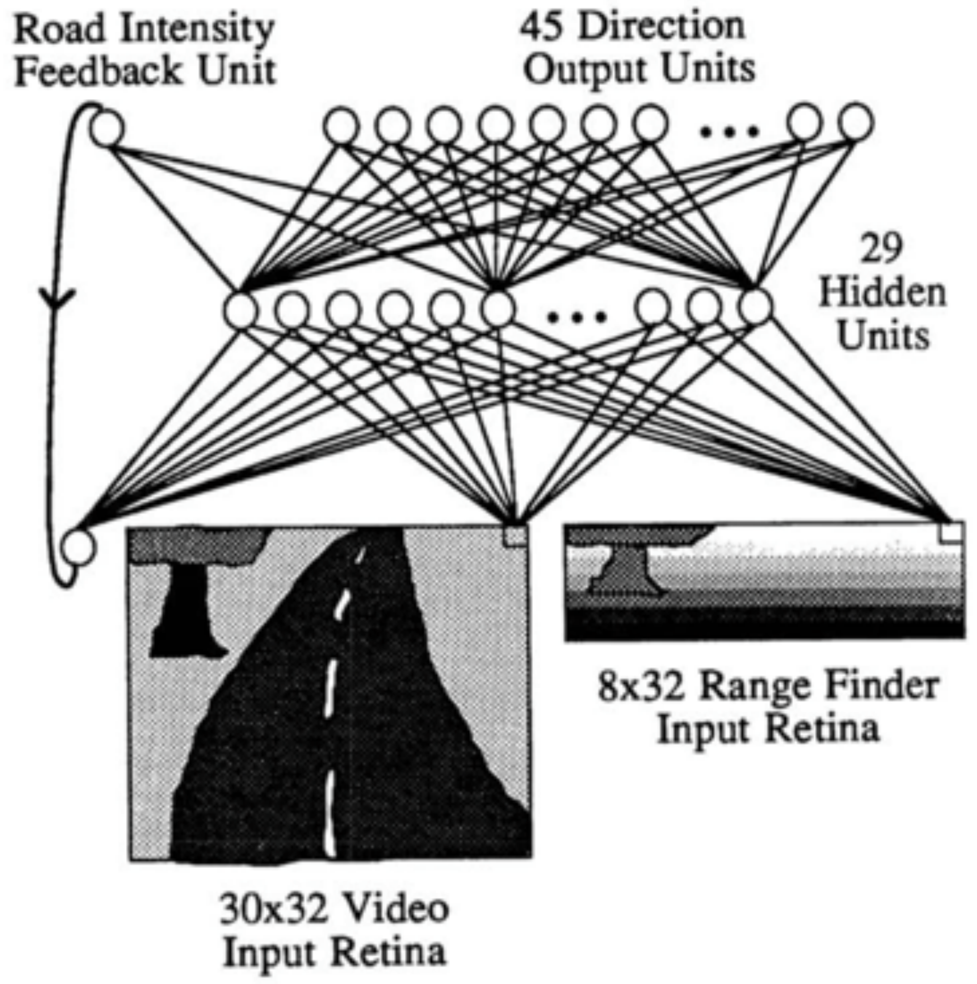


Figure 1: ALVINN Architecture

"...the network must not solely be shown examples of accurate driving, but also how to recover (i.e. return to the road center) once a mistake has been made."

D. Pomerleau
ALVINN: An Autonomous Land Vehicle In A Neural Network, NeurIPS'89

Also observed by [LeCun'05]

Feedback is a pervasive problem in self-driving

"... the inertia problem. When the ego vehicle is stopped (e.g., at a red traffic light), the probability it stays static is indeed overwhelming in the training data. This creates a spurious correlation between low speed and no acceleration, inducing excessive stopping and difficult restarting in the imitative policy ..."

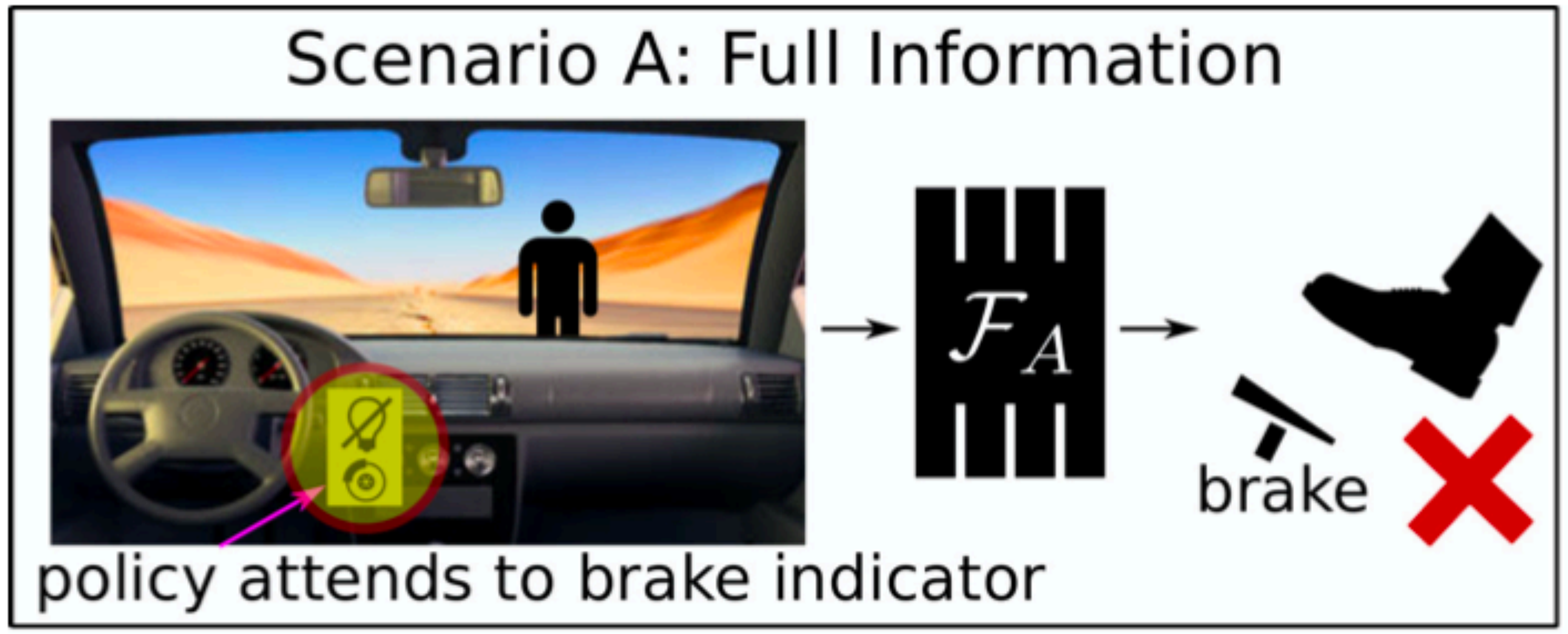
"Exploring the Limitations of Behavior Cloning for Autonomous Driving."
F. Codevilla, E. Santana, A. M. Lopez, A. Gaidon. ICCV 2019

"... small errors in action predictions to compound over time, eventually leading to states that human drivers infrequently visit and are not adequately covered by the training data. Poorer predictions can cause a feedback cycle known as cascading errors ..."

"Imitating Driver Behavior with Generative Adversarial Networks".
A. Kuefler, J. Morton, T. Wheeler, M. Kochenderfer, IV 2017

"... During closed-loop inference, this breaks down because the past history is from the net's own past predictions. For example, such a trained net may learn to only stop for a stop sign if it sees a deceleration in the past history, and will therefore never stop for a stop sign during closed-loop inference ..."

"ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst". M. Bansal, A. Krizhevsky, A. Ogale, Waymo 2018



"Causal Confusion in Imitation Learning".
P. de Haan, D. Jayaraman, S. Levine, NeurIPS '19

Feedback is a problem for LLMs

Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

"The probability of a repeated phrase increases with each repetition, creating a positive feedback loop"

The curious case of neural text de-generation
Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019).

"The main problem is that mistakes made early in the sequence generation process are fed as input to the model and can be quickly amplified because the model might be in a part of the state space it has never seen at training time."

"Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks." Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015).

Thus, the model trained with teacher forcing may over-rely on previously predicted words, which would exacerbate error propagation

"On exposure bias, hallucination and domain shift in neural machine translation." Wang, C., & Sennrich, R. (2020).

DeepMind

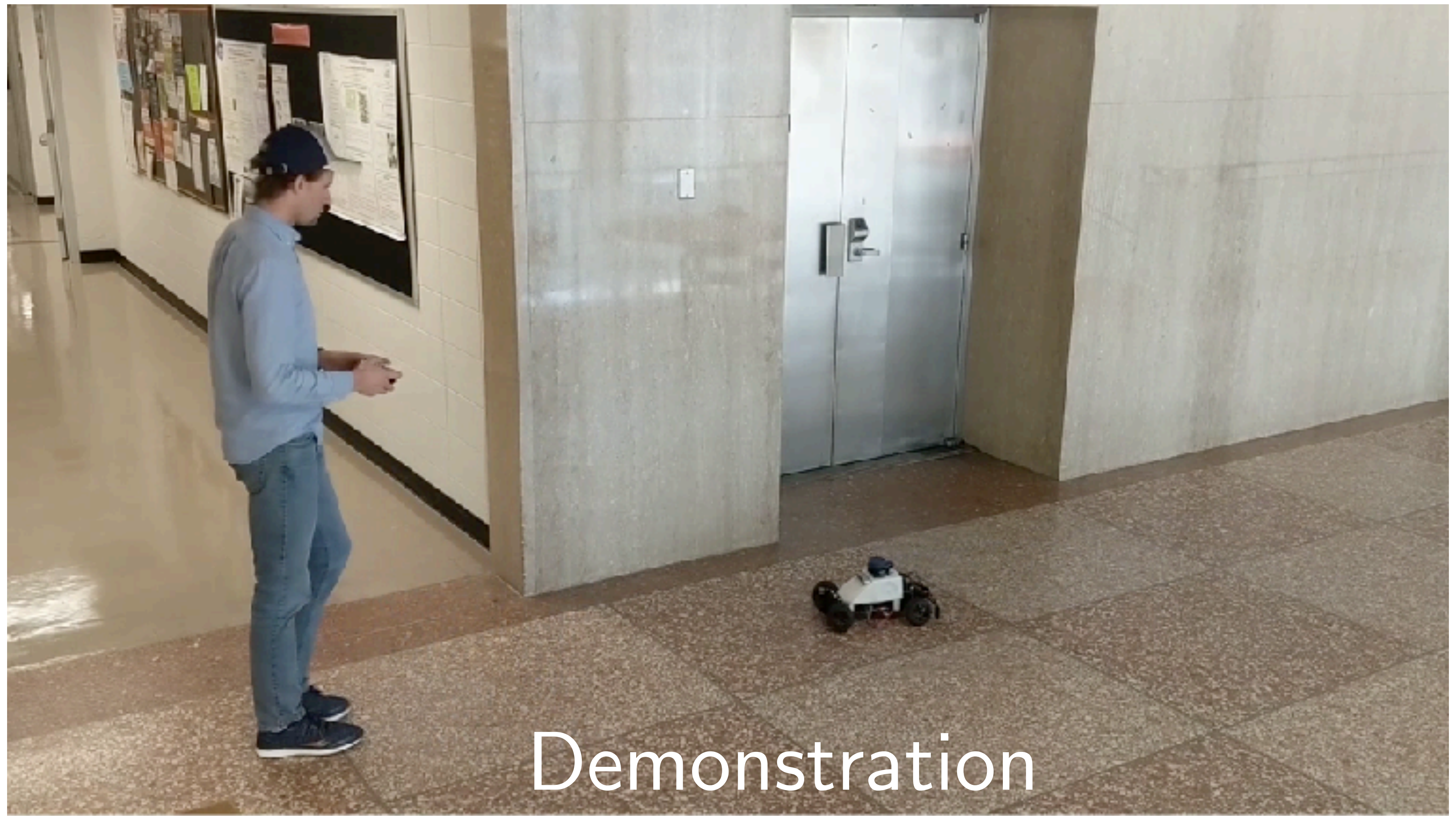
Technical Report
2021-10-22

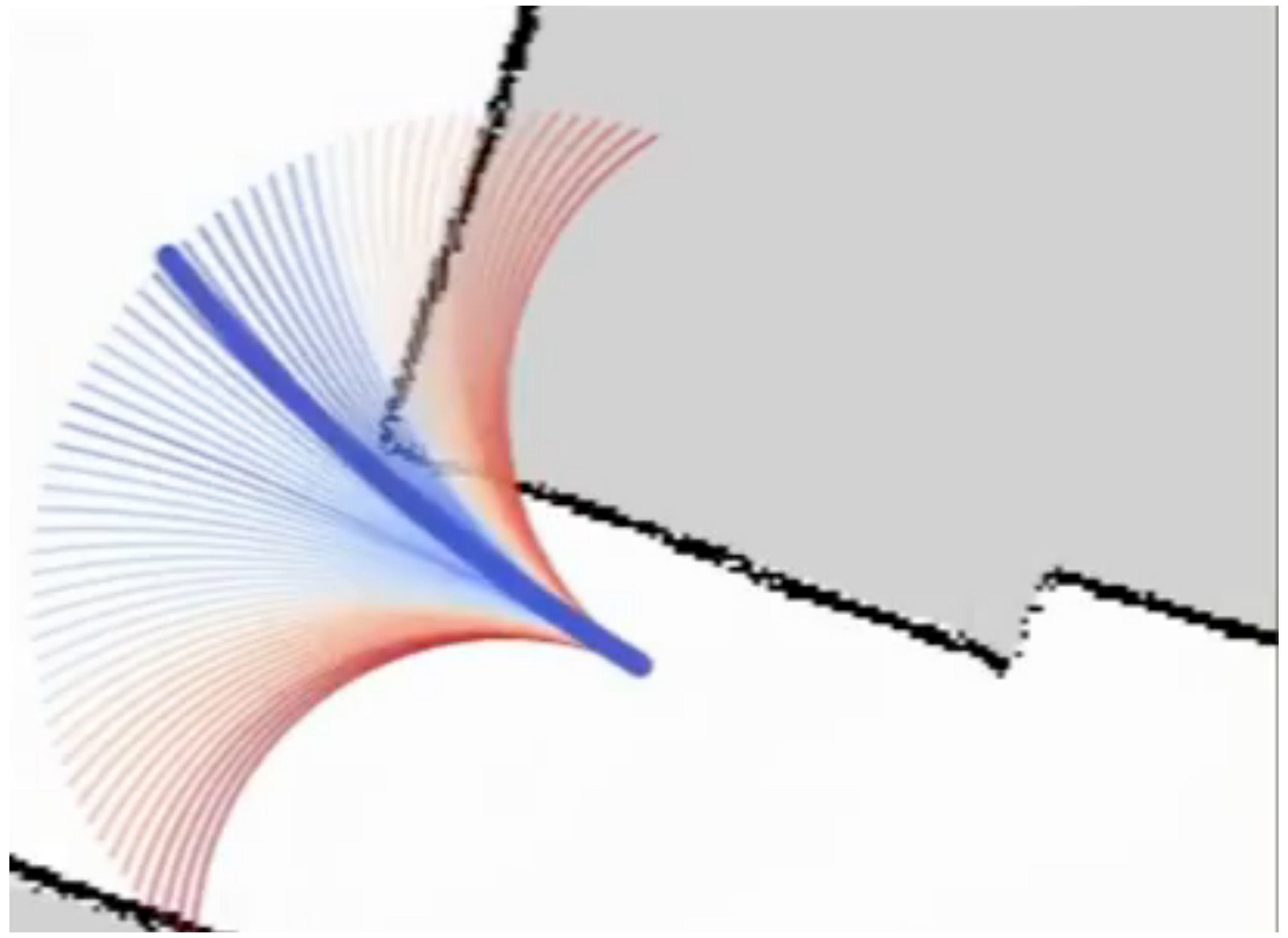
Shaking the foundations: delusions in sequence models for interaction and control

Pedro A. Ortega[*], Markus Kunesch[*], Grégoire Delétang[*], Tim Genewein[*], Jordi Grau-Moya[*], Joel Veness[1], Jonas Buchli[1], Jonas Degrave[1], Bilal Piot[1], Julien Perolat[1], Tom Everitt[1], Corentin Tallec[1], Emilio Parisotto[1], Tom Erez[1], Yutian Chen[1], Scott Reed[1], Marcus Hutter[1], Nando de Freitas[1] and Shane Legg[1]
[*]Deepmind Safety Analysis, [1]DeepMind

Feedback is an old adversary!

[SCB+ RSS'20]

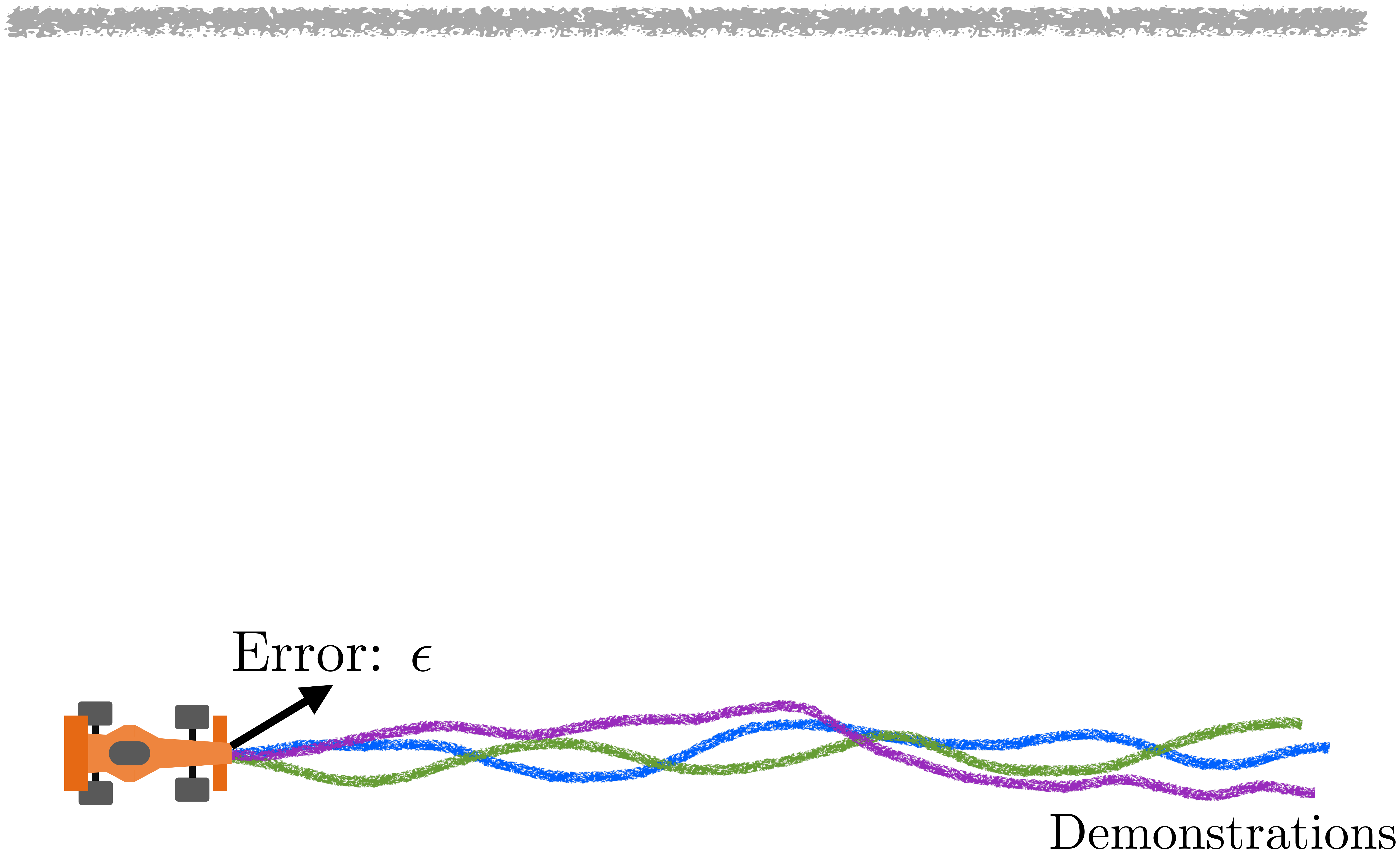




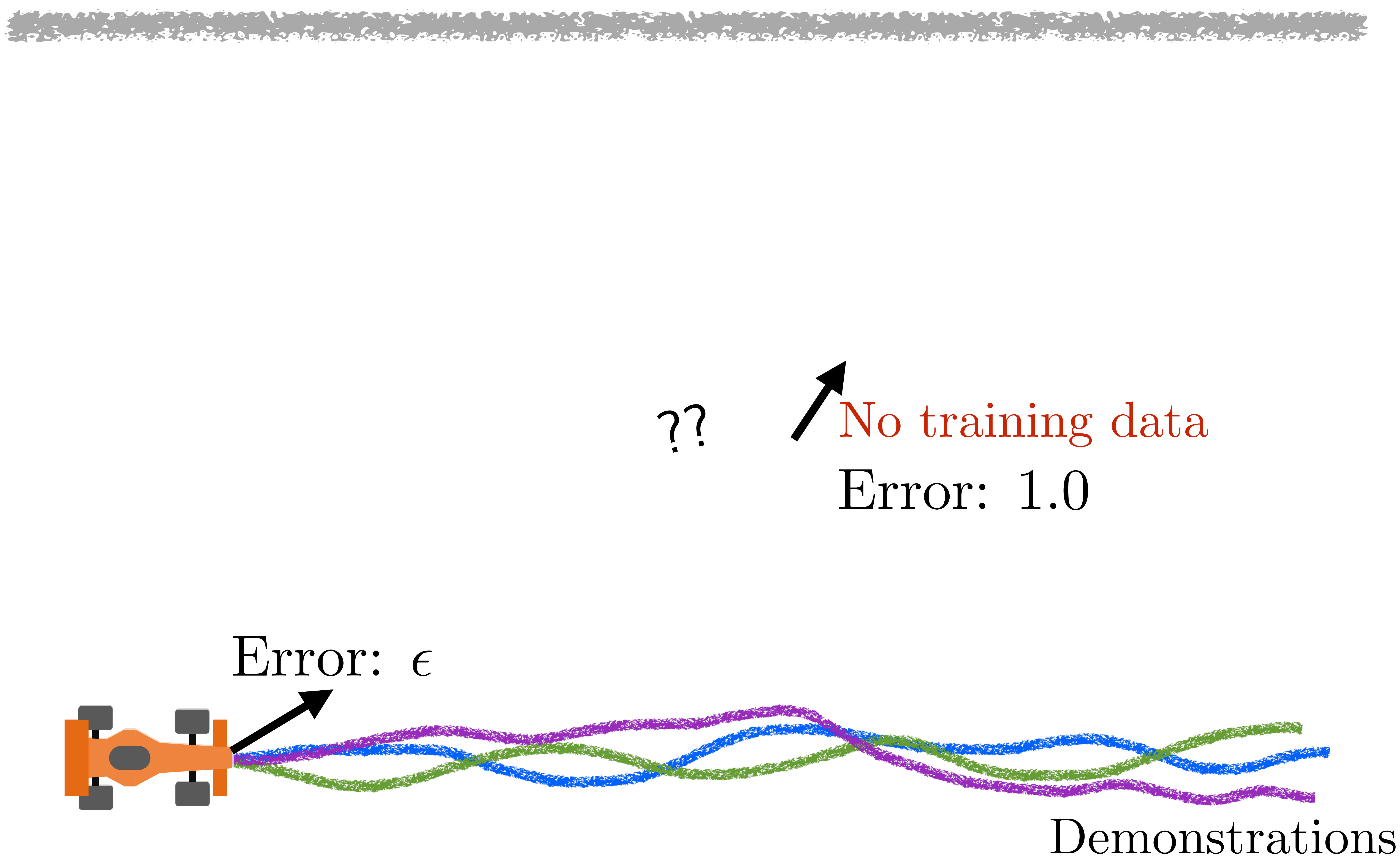
Learnt policy

Behavior Cloning crashes into a wall

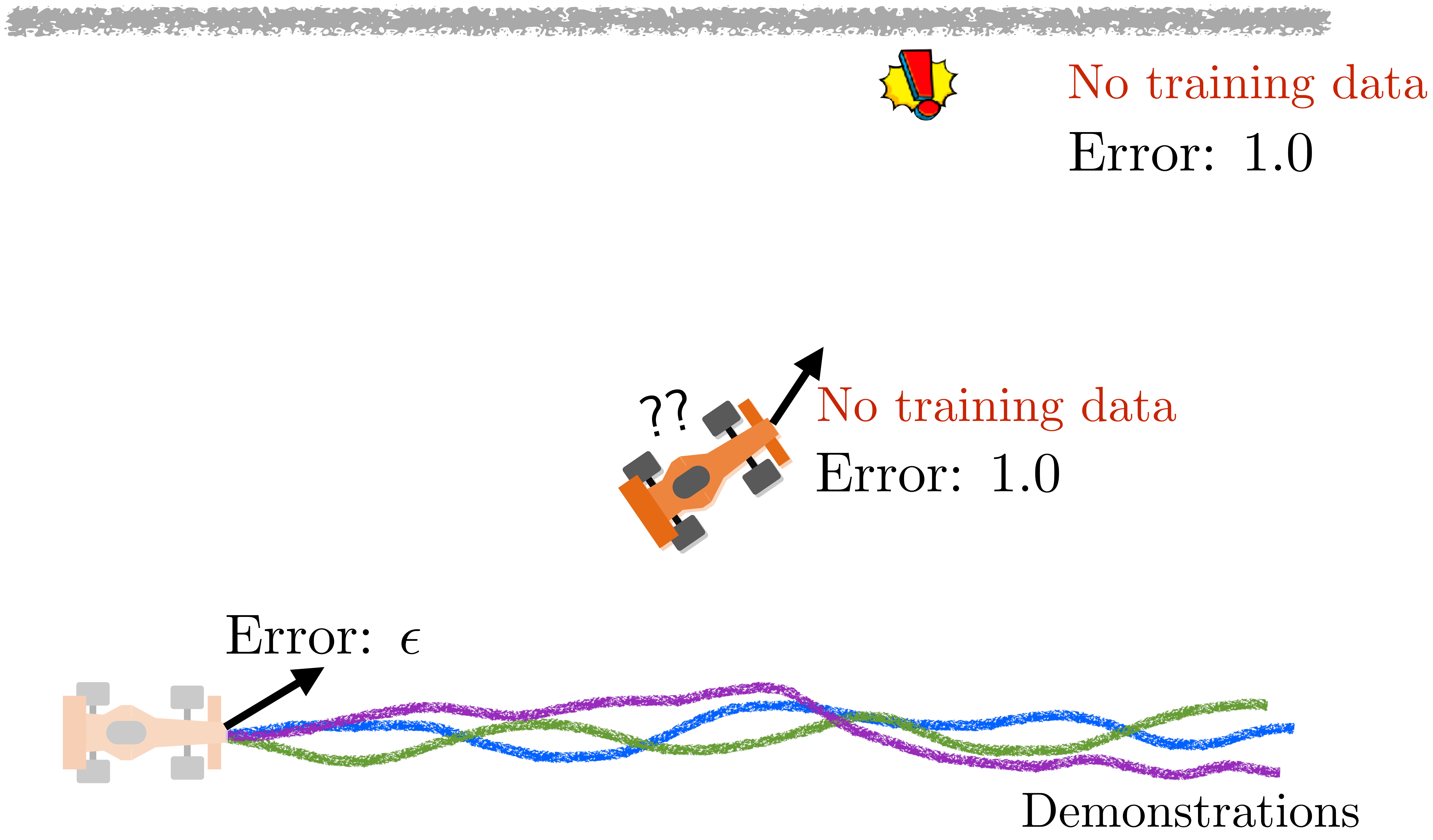
Why did the robot crash?



Why did the robot crash?



Why did the robot crash?



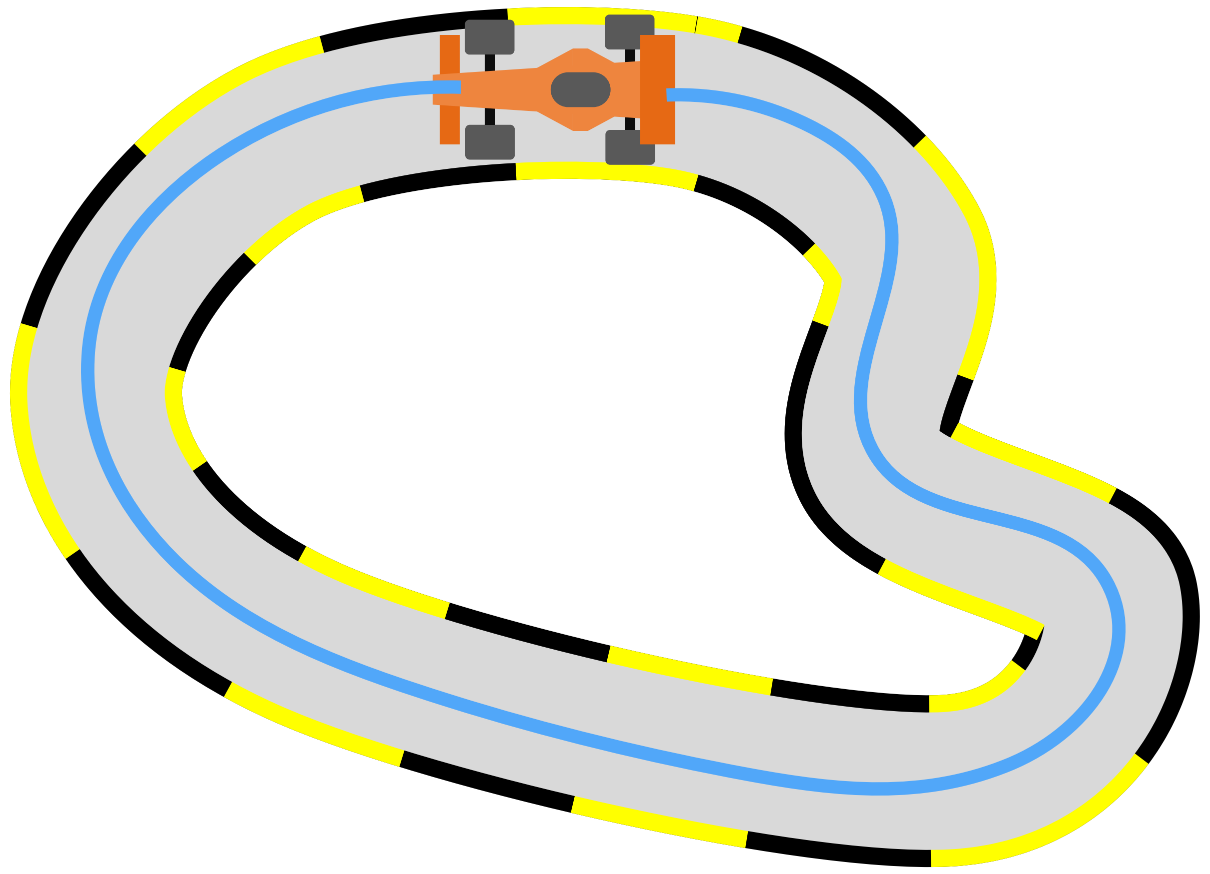
Behavior Cloning crashes into a wall

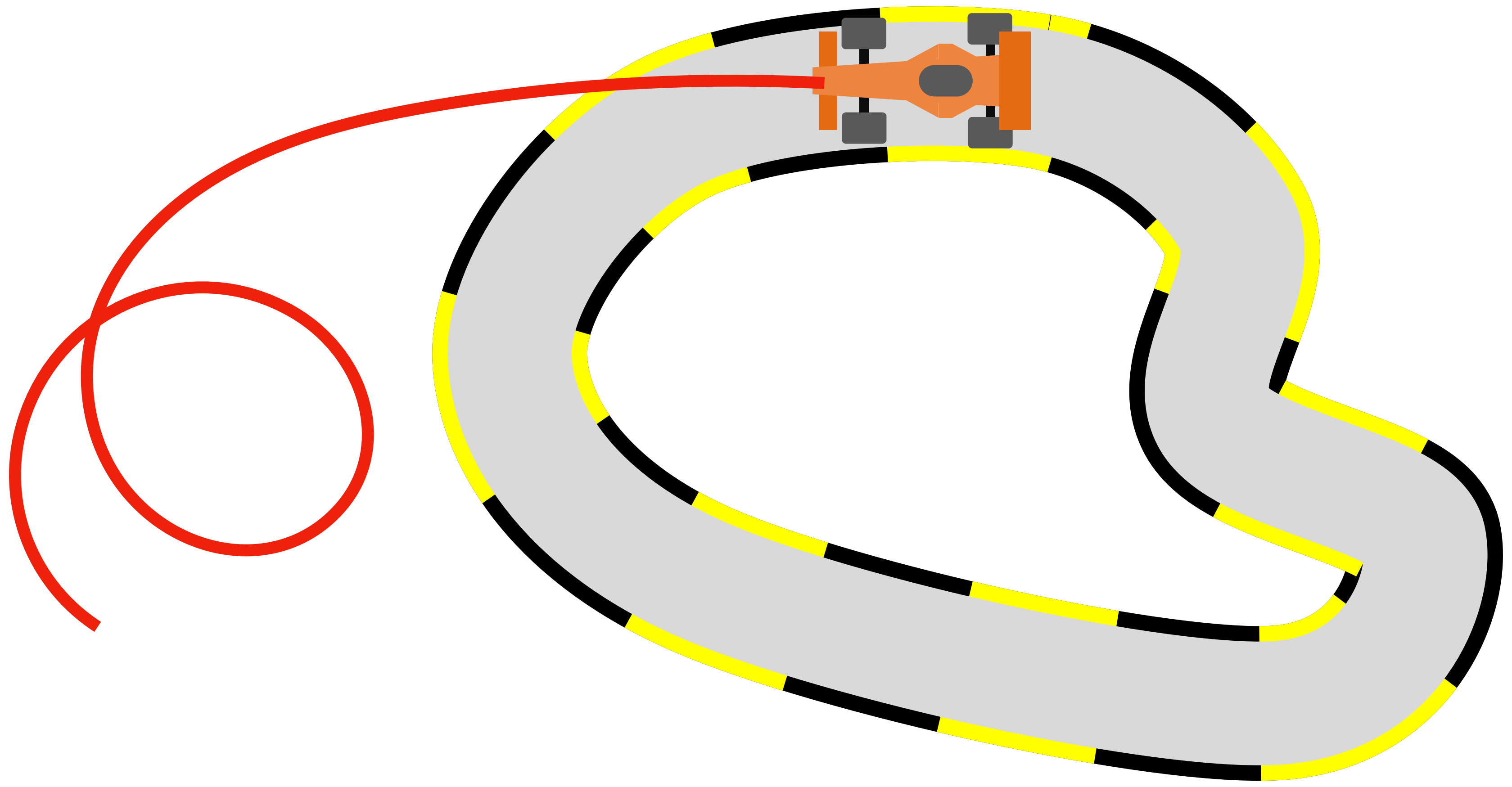
Train $\neq$ Test

Train $\neq$ Test

$$\sum_{t=0}^{T-1} \underbrace{\mathbb{E}_{s_t \sim d_t^{\pi^\star}}}_{\text{States visited by expert demonstrator}} \underbrace{[\ell(s_t, \pi(s_t))]}_{\text{Loss}}$$

$$\sum_{t=0}^{T-1} \underbrace{\mathbb{E}_{s_t \sim d_t^{\pi}}}_{\text{States visited by learner}} \underbrace{[\ell(s_t, \pi(s_t))]}_{\text{Loss}}$$





Today's class

- [x] Feedback drives Covariate Shift
- [ ] BC has a performance gap of $O(\epsilon T^2)$
- [ ] Easy vs Hard Regimes in Imitation Learning

Can we mathematically quantify how much worse BC is compared to the demonstrator?

First, let's define performance of a policy

$$J(\pi) = \mathbb{E}_{\substack{a_t \sim \pi(s_t) \\ s_{t+1} \sim \mathcal{T}(s_t, a_t)}} \left[\sum_{t=0}^{T-1} c(s_t, a_t) \right]$$

(Performance)

Second, let's define performance difference

$$J(\pi) - J(\pi^*)$$

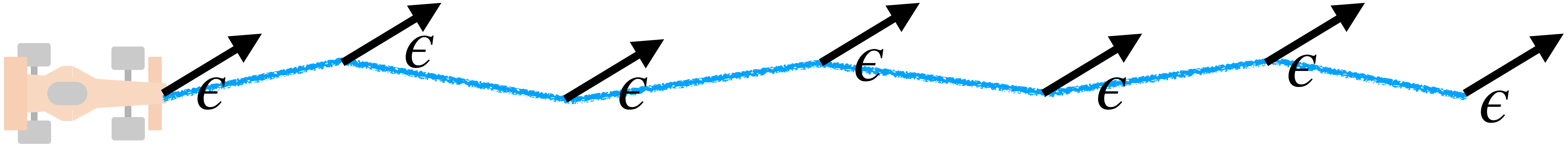
(Performance of my learner) (Performance of my demonstrator)

We want to *minimize* the performance difference

How low can we drive performance difference?

$$J(\pi) - J(\pi^*)$$

Let's say my learner is not perfect and can only drive down training / validation error to be ϵ



Cumulative error over time T $= \epsilon + \epsilon + \ldots = \epsilon T$

How low can we drive performance difference?

Let's say my learner is not perfect and can only drive down training / validation error to be ϵ

😄 The best we can hope for is that error grows **linearly** in time

$$J(\pi) - J(\pi^*) \leq O(\epsilon T)$$

😱 The worst case is if error **compounds quadratically** in time

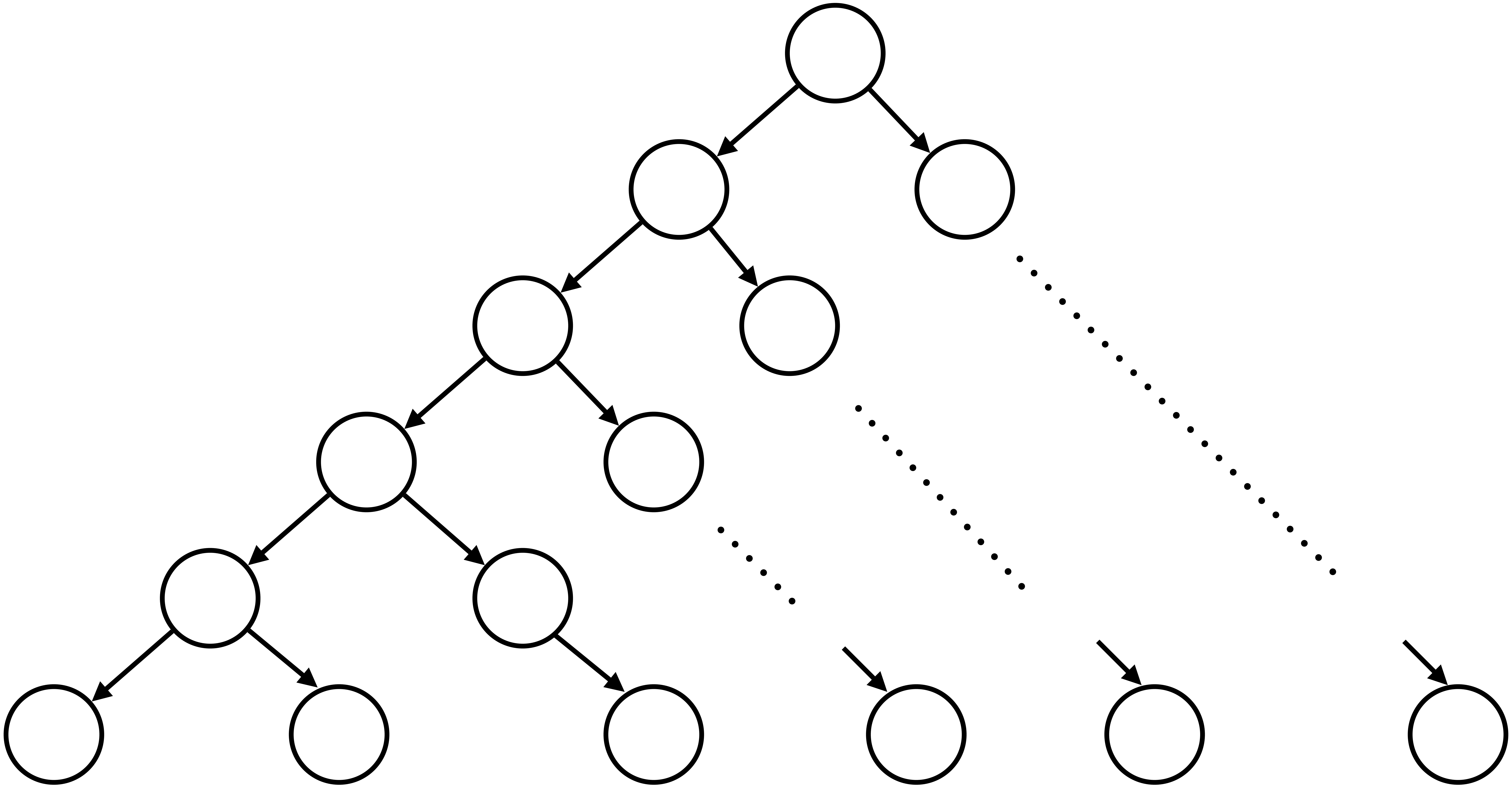
$$J(\pi) - J(\pi^*) \leq O(\epsilon T^2)$$

Behavior cloning hits the worst case!

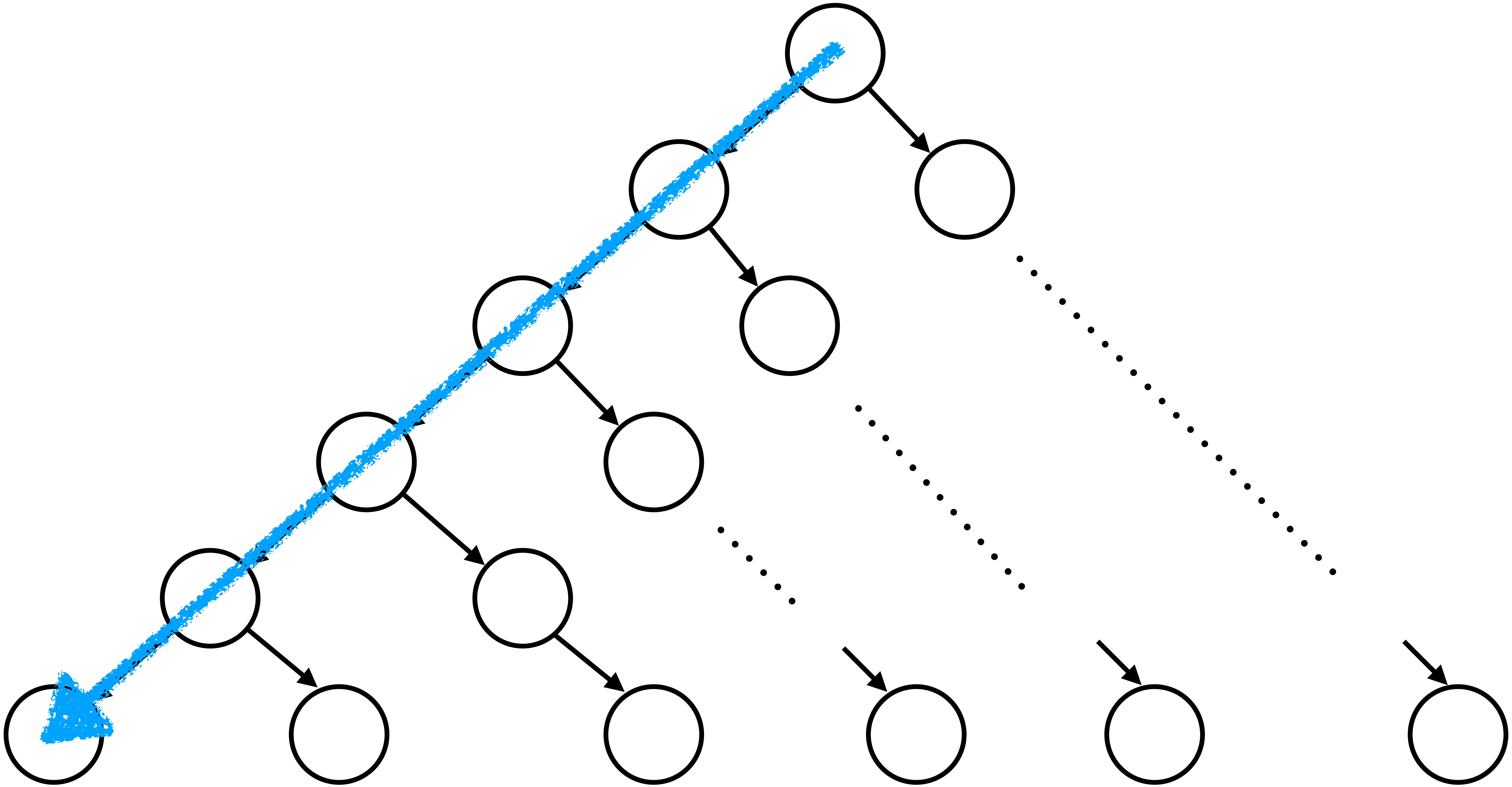
There exists an MDP where BC
has a performance difference of $O(\epsilon T^2)$

We are going to such a MDP right now,
and you will see more in A1!

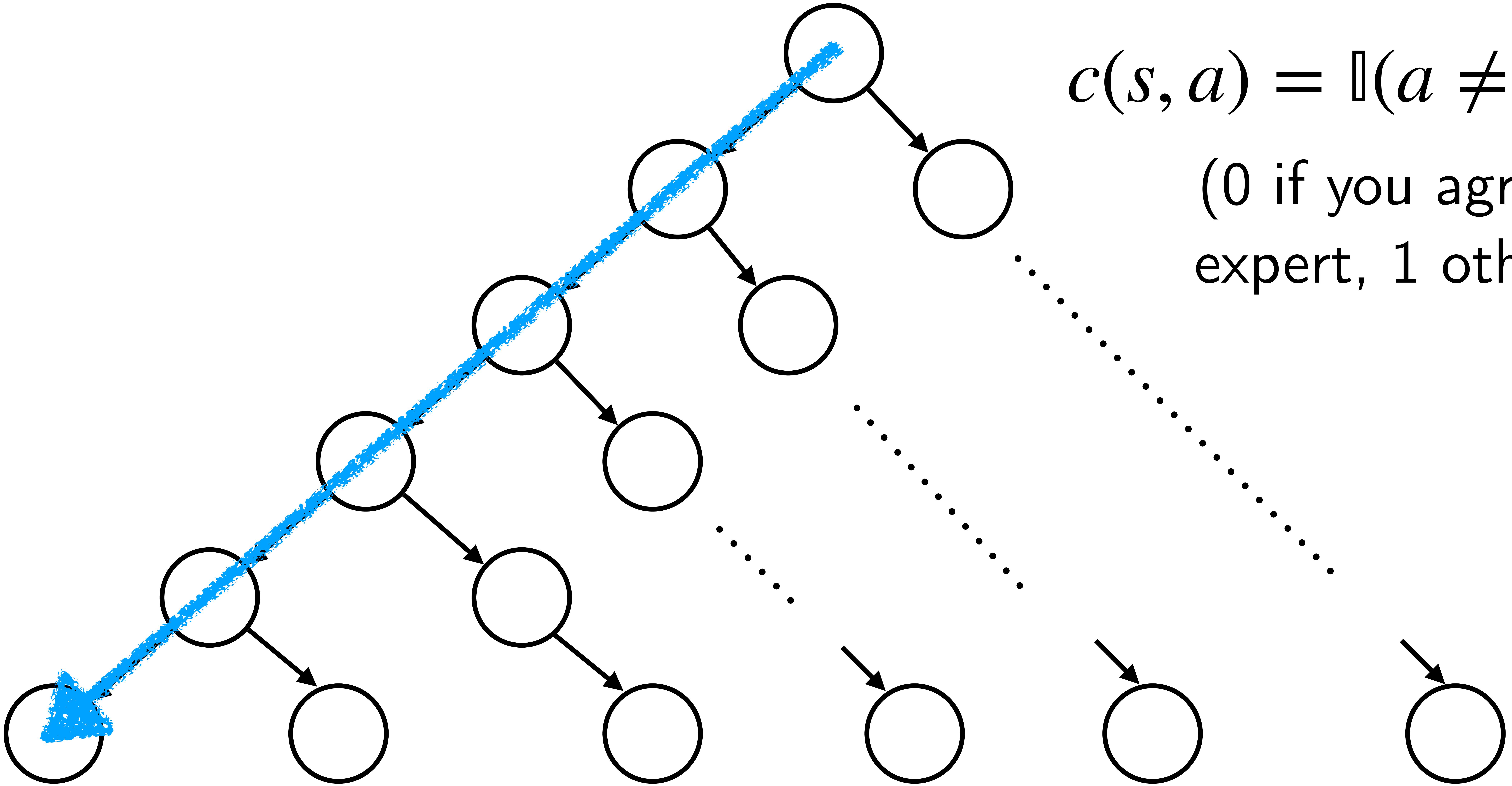
A Tree MDP



The demonstrator always takes a left



Assume the following cost function

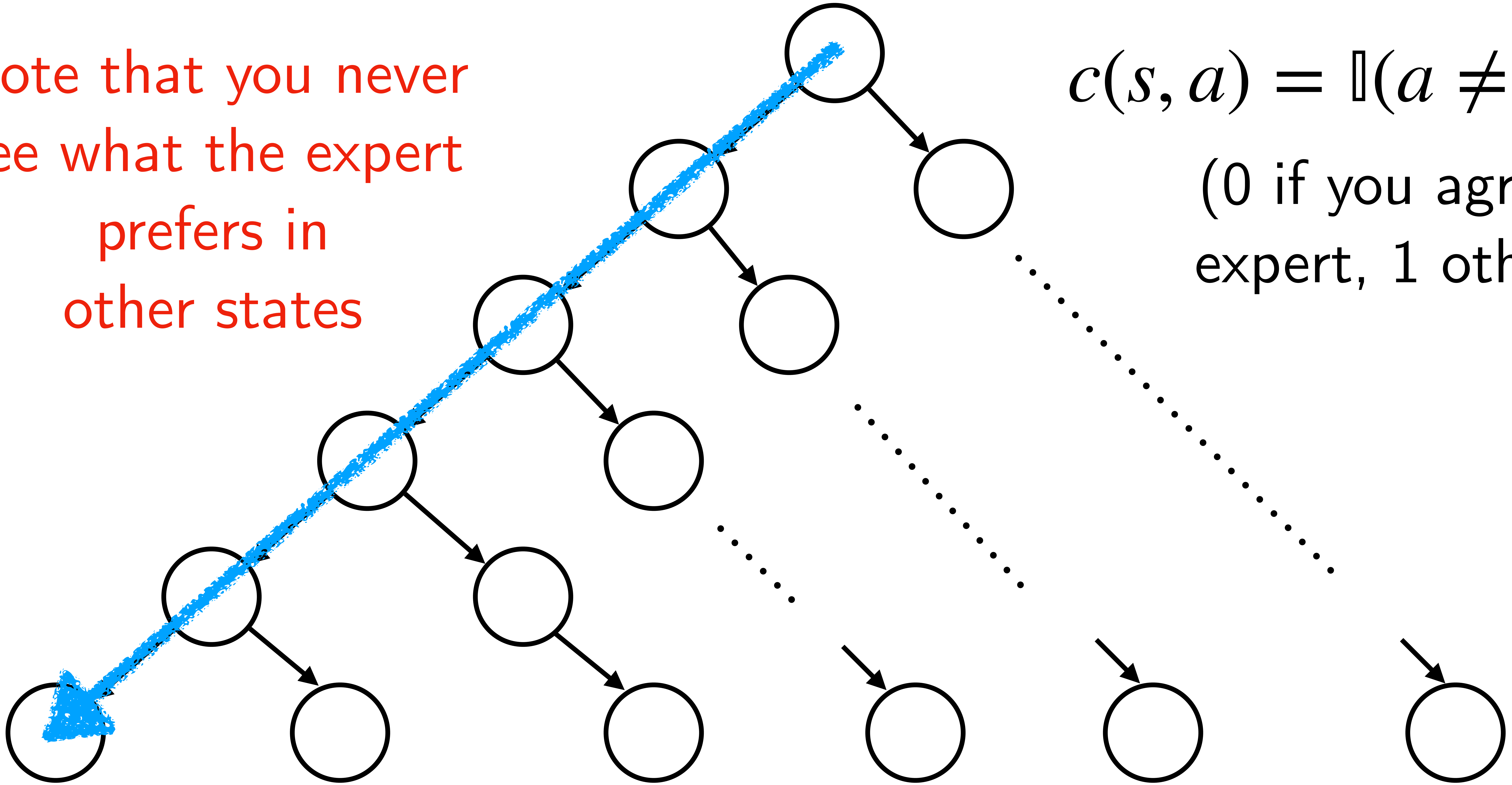


$$c(s, a) = \mathbb{I}(a \neq \pi^*(s))$$

(0 if you agree with expert, 1 otherwise)

Assume the following cost function

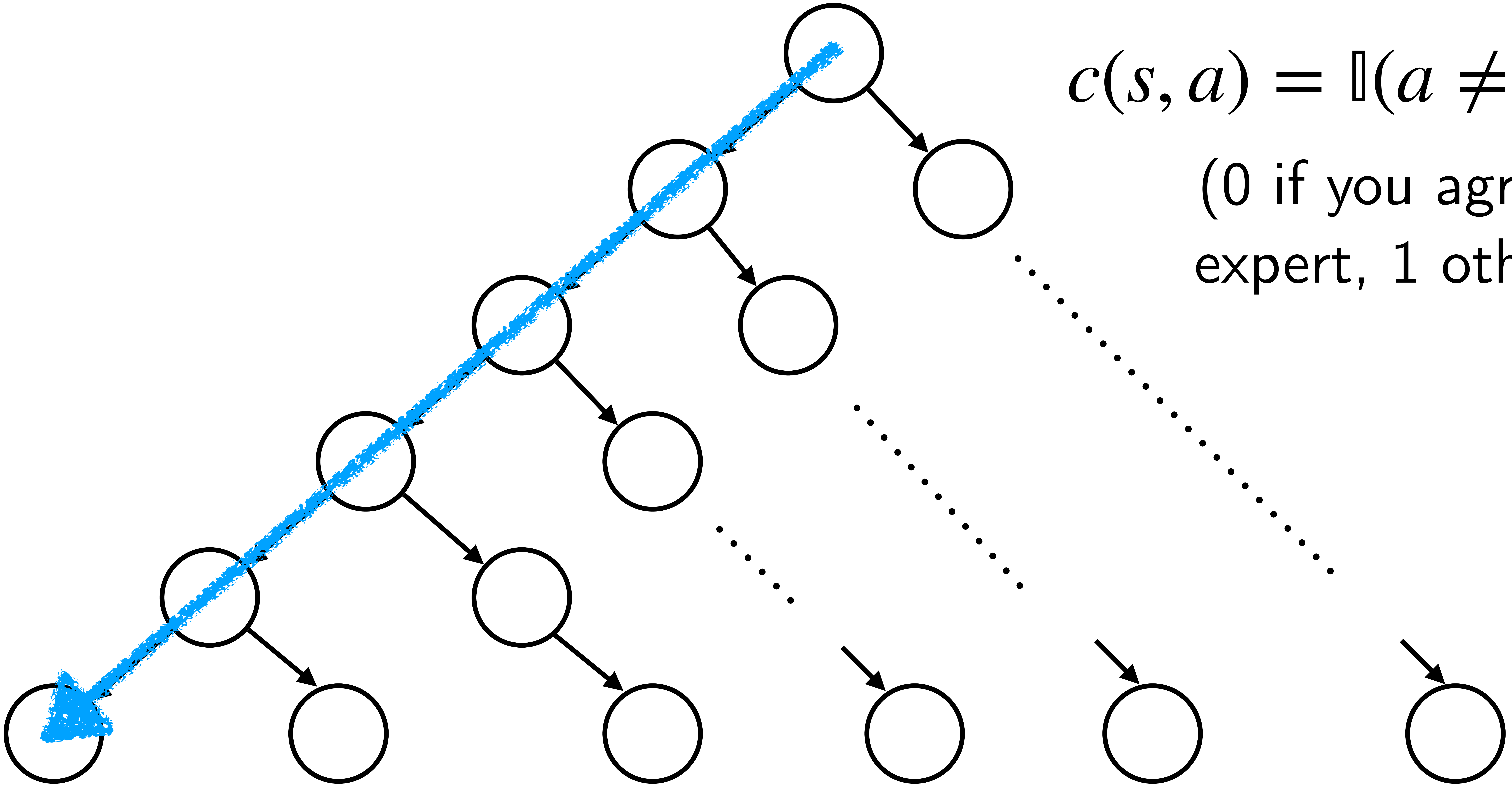
Note that you never see what the expert prefers in other states



$$c(s, a) = \mathbb{I}(a \neq \pi^*(s))$$

(0 if you agree with expert, 1 otherwise)

Show that BC has a performance difference of $O(\epsilon T^2)$



$$c(s, a) = \mathbb{I}(a \neq \pi^*(s))$$

(0 if you agree with expert, 1 otherwise)

Proof

Today's class

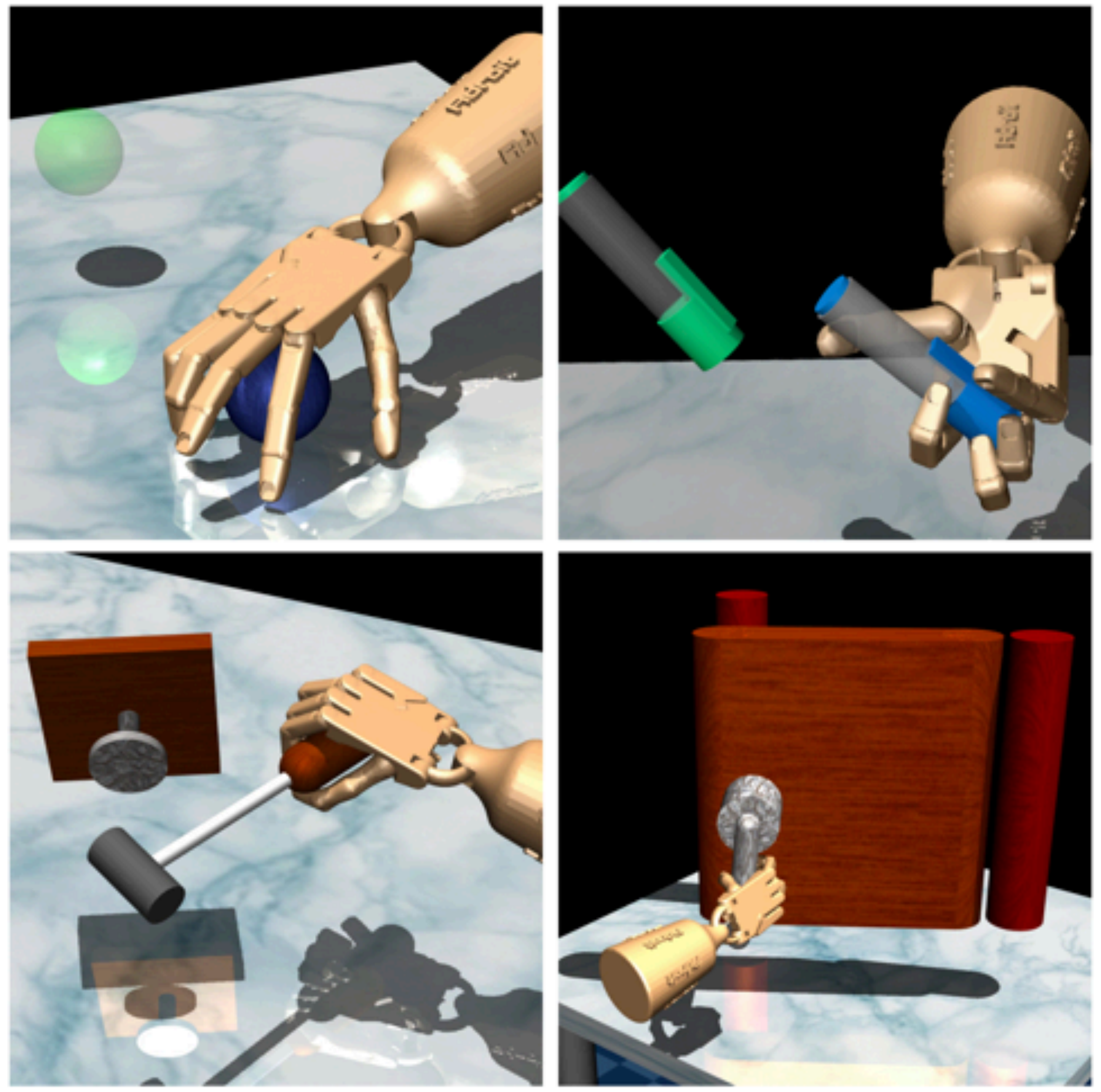
- ☑ Feedback drives Covariate Shift
- ☑ BC has a performance gap of $O(\epsilon T^2)$
- ☐ Easy vs Hard Regimes in Imitation Learning

So, it seems BC is totally doomed ...

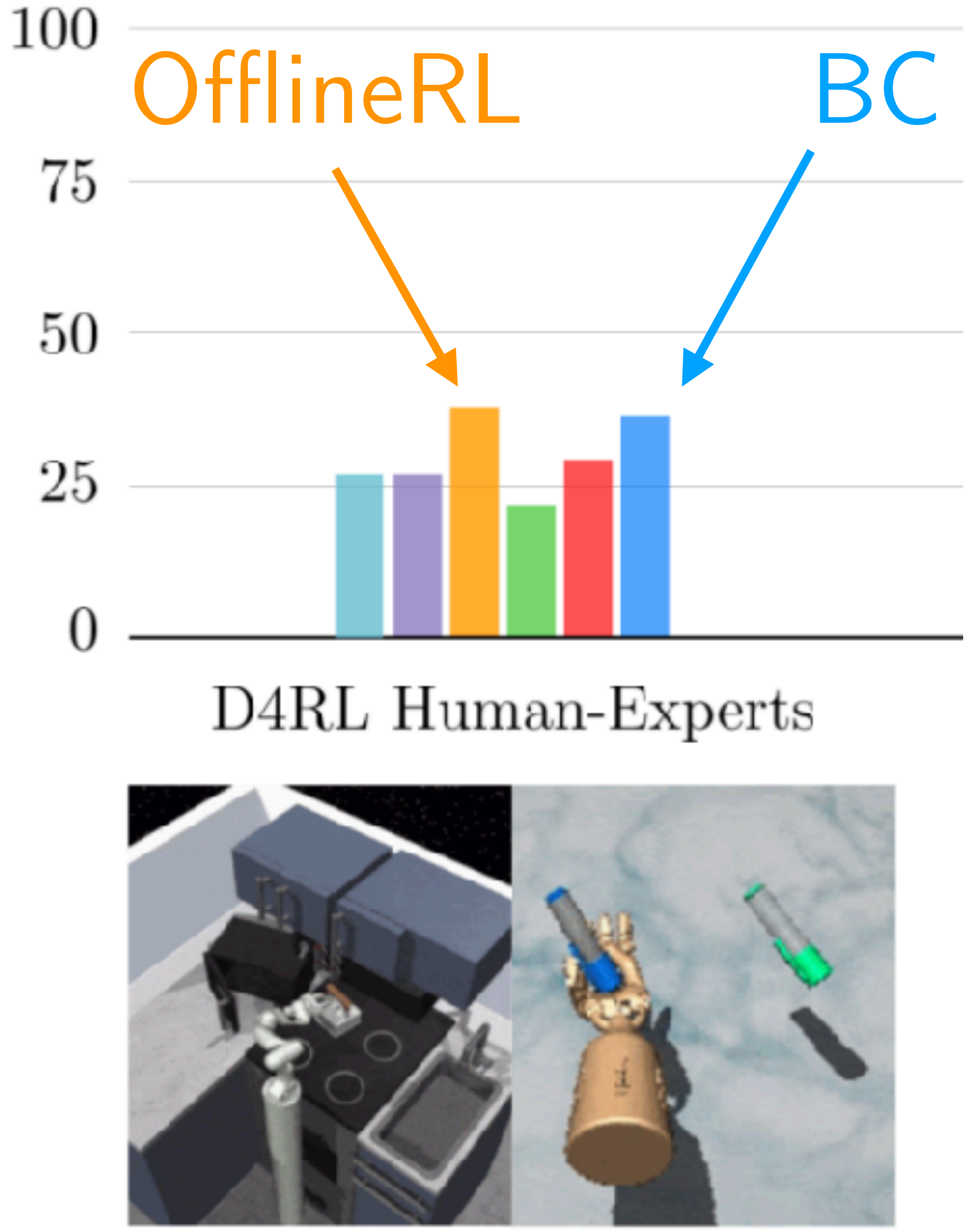
But, BC works surprisingly often!!

Environment	Expert	BC
CartPole	500 ± 0	500 ± 0
Acrobot	-71.7 ± 11.5	-78.4 ± 14.2
MountainCar	-99.6 ± 10.9	-107.8 ± 16.4
Hopper	3554 ± 216	3258 ± 396
Walker2d	5496 ± 89	5349 ± 634
HalfCheetah	4487 ± 164	4605 ± 143
Ant	4186 ± 1081	3353 ± 1801

[SCV+ arXiv '21]



[Rajeswaran et al. '17]



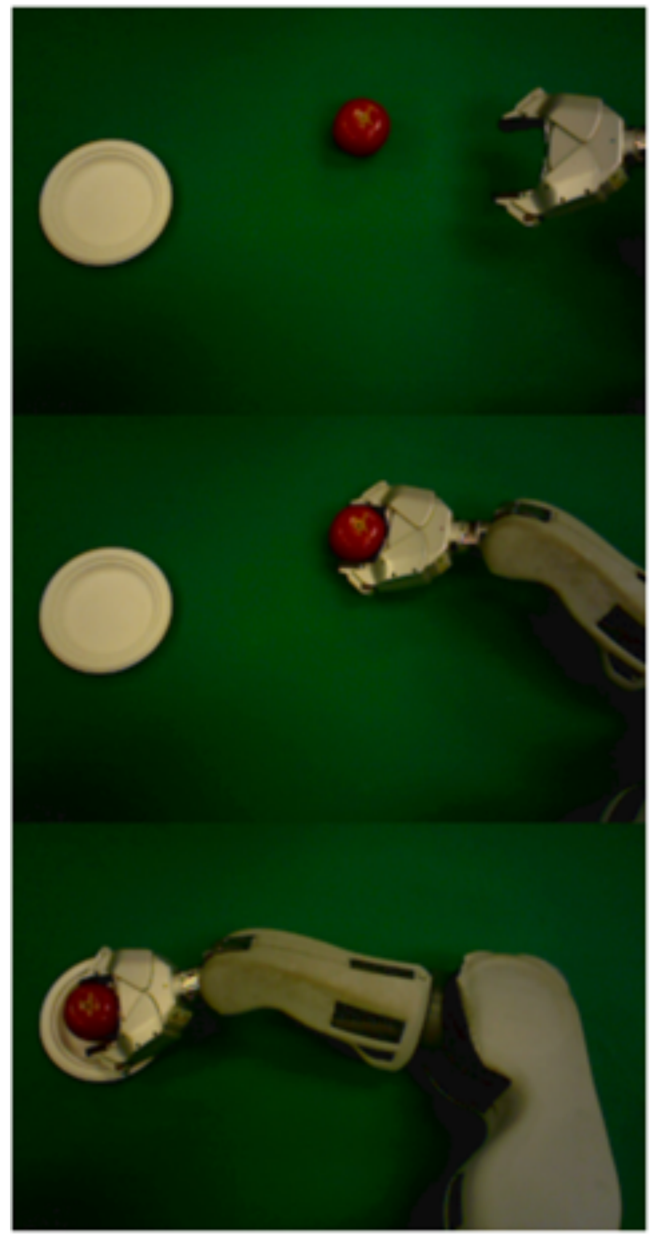
[Florence et al. '21]

But, BC works surprisingly often!!

Deep Imitation Learning for Complex Manipulation Tasks from Virtual Reality Teleoperation

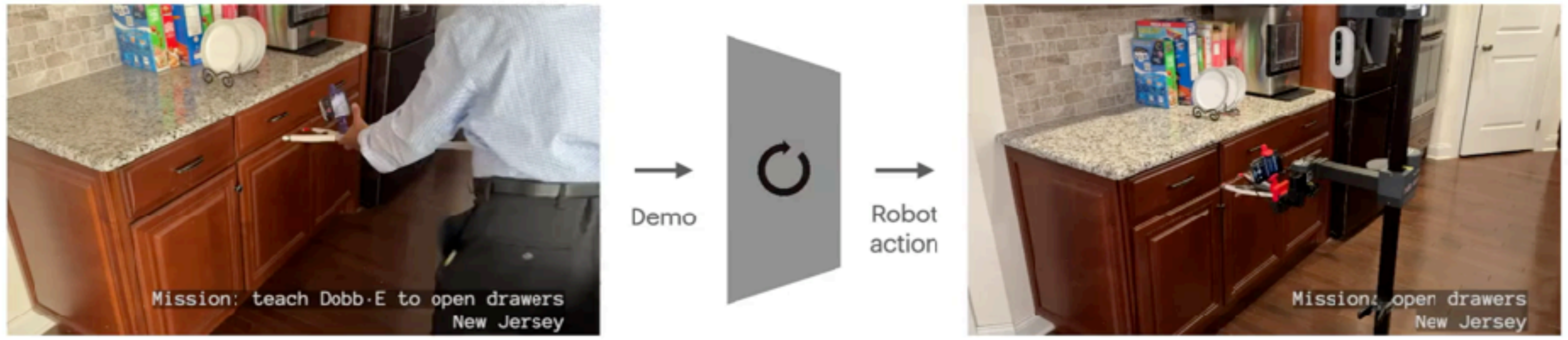
Tianhao Zhang*[12], Zoe McCarthy*[1], Owen Jow[1], Dennis Lee[1], Xi Chen[12], Ken Goldberg[1], Pieter Abbeel[1-4]

Fig. 1: Virtual Reality teleoperation in action



On Bringing Robots Home

Nur Muhammad (Mahi) Shafiullah*† NYU
Anant Rai* NYU
Haritheja Etukuru NYU
Yiqian Liu NYU
Ishan Misra Meta
Soumith Chintala Meta
Lerrel Pinto NYU



🤔 Why does BC work in these cases?

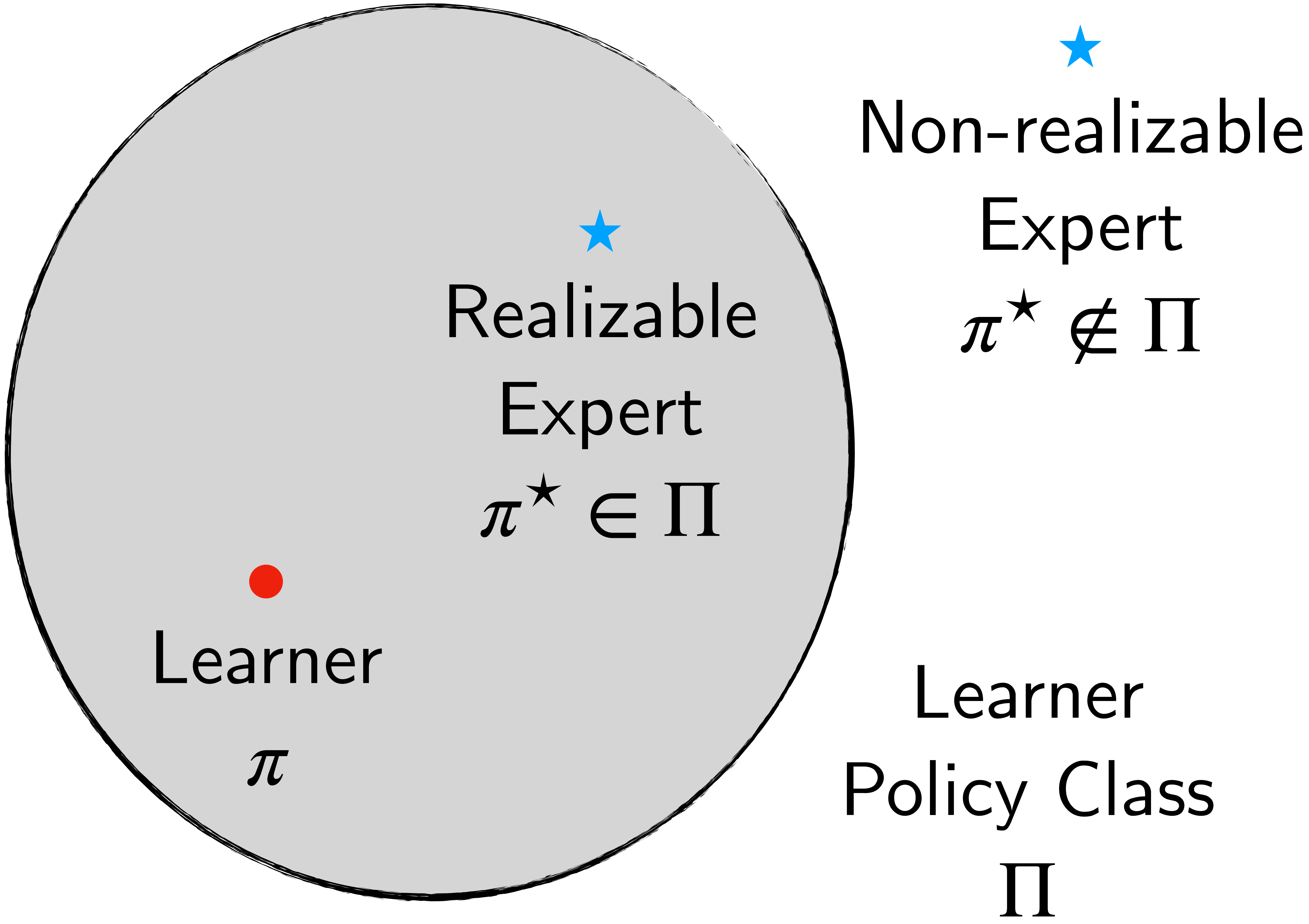
$$O(\epsilon T^2)$$

Drive ϵ to 0!

When can we actually do this?

The Realizable Setting

With infinite data and a realizable expert, can drive $\epsilon \to 0$



Realizable settings are EASY!

With infinite data, can drive $\epsilon \to 0$

BC works just fine!

Non-realizable Expert is HARD

Even with infinite data, $\epsilon > 0$

Learner will make an error, go to a state that expert has not visited, $O(\epsilon T^2)$

What is the hard case where $\epsilon > 0$?

Non-realizable Expert!

Survey

Realizable vs Non-Realizable Expert

When poll is active respond at **PollEv.com/sc2582**

Today's class

- ☑ Feedback drives Covariate Shift
- ☑ BC has a performance gap of $O(\epsilon T^2)$
- ☑ Easy vs Hard Regimes in Imitation Learning

What can we do in the HARD setting?

Query the expert on states the learner visits

DAgger (Dataset Aggregation)

Initialize with a random policy π_1 # Can be BC

Initialize empty data buffer $\mathcal{D} \leftarrow \{\}$

For $i = 1, \ldots, N$

Execute policy π_i in the real world and collect data

$$\mathcal{D}_i = \{s_0, a_0, s_1, a_1, \ldots\}$$ # Also called a rollout

Query the expert for the optimal action on learner states

$$\mathcal{D}_i = \{s_0, \pi^\star(s_0), s_1, \pi^\star(s_1), \ldots\}$$

Aggregate data $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_i$

Train a new learner on this dataset $\pi_{i+1} \leftarrow \text{Train}(\mathcal{D})$

Select the best policy in $\pi_{1:N+1}$

Why does DAgger work?

Theory of Online Learning
explains why
(Next Lecture!)