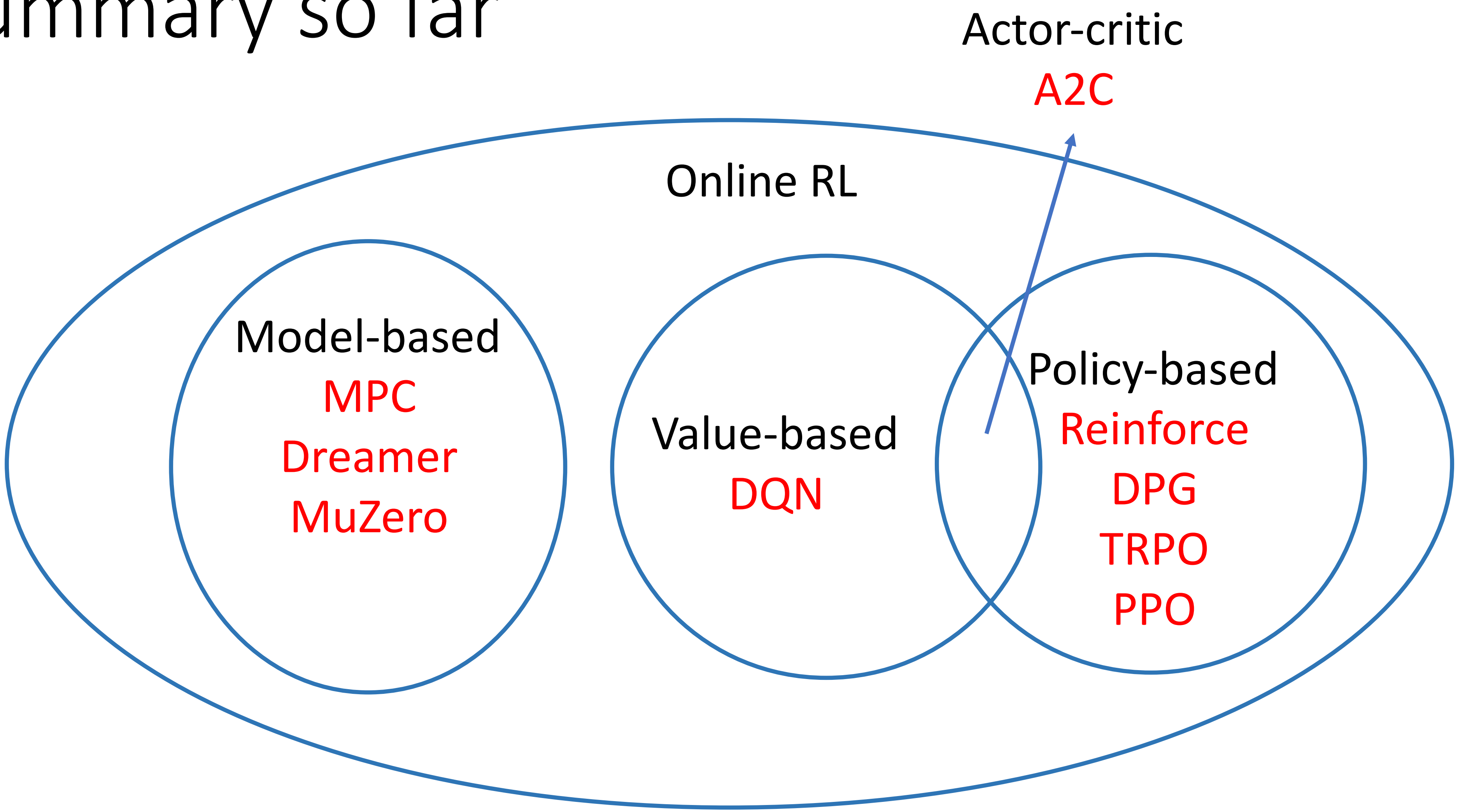
DS 598
Introduction to RL

Xuezhou Zhang

Summary so far

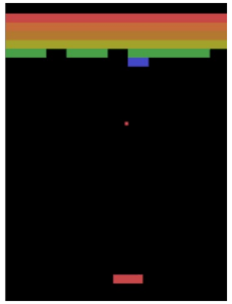


Chapter 6: Imitation Learning

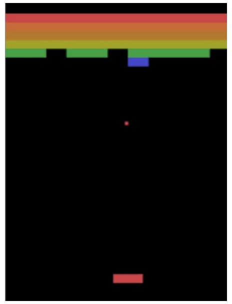
Imitation is at the heart of human/animal learning

Imitation is at the heart of human/animal learning

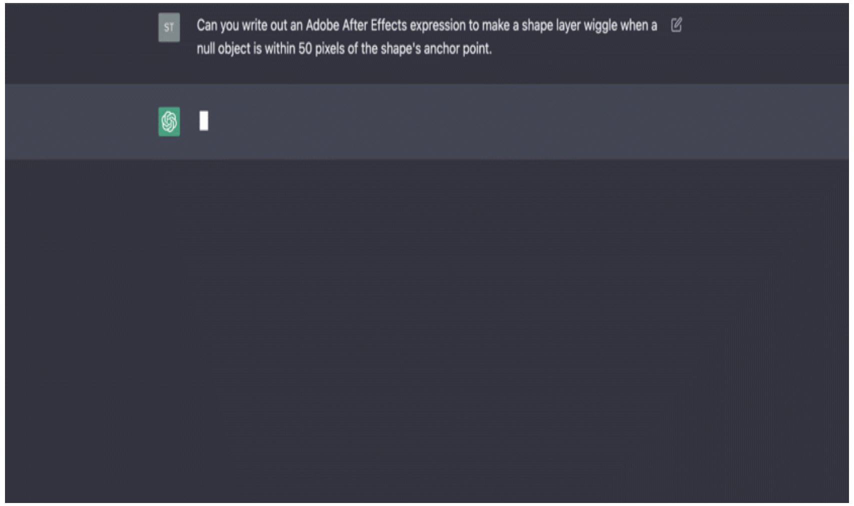
What is imitation learning?

( , )$_{1:N}$

Goal: find f such that

f() $\approx$ 

Examples of Imitation Learning

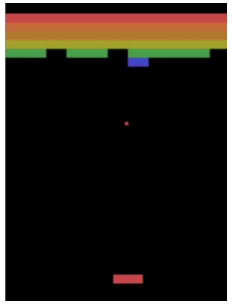


Why do imitation learning

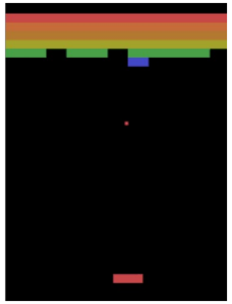
1. There are expert data available, why not make use of it? e.g. LLM.

2. It's hard to define a reward function for the desired behavior, give me a demo. e.g. autonomous driving.

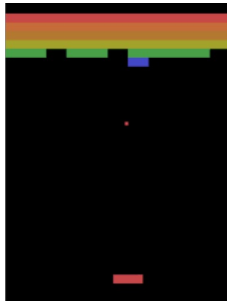
How to perform imitation learning?

$($  $,$  $)_{1:M} \sim \pi^\star$

Goal: find f such that

$f($  $) \approx$ 

Approach 1: Behavior Cloning (BC)

( , )$_{1:M}$

Given a data set of (X, Y) pairs, predict Y as a function of X.

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

- Classification (finite discrete actions)

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

- Regression (continuous actions)

  Square loss: $\ell(\pi, s, a^\star) = \|\pi(s) - a^\star\|_2^2$

Approach 1: Behavior Cloning (BC)

How well does this work?

Let's assume supervised learning succeeded:

$$\mathbb{E}_{s\sim d_{\pi^\star}}[\hat{\pi}(s) \neq \pi^\star(s)] \leq \epsilon \approx O(\sqrt{1/N})$$

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

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

Quadratic error amplification

Approach 1: Behavior Cloning (BC)

SL guarantee: $\mathbb{E}_{s\sim d_{\pi^\star}}[\hat{\pi}(s) \neq \pi^\star(s)] \leq \epsilon \approx O(\sqrt{1/N})$

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

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

Proof: Performance Difference Lemma: $(1-\gamma)\left(f(\pi) - f(\pi')\right) = \mathbb{E}_{s,a\sim d^{\pi}}\left[A^{\pi'}(s,a)\right]$

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

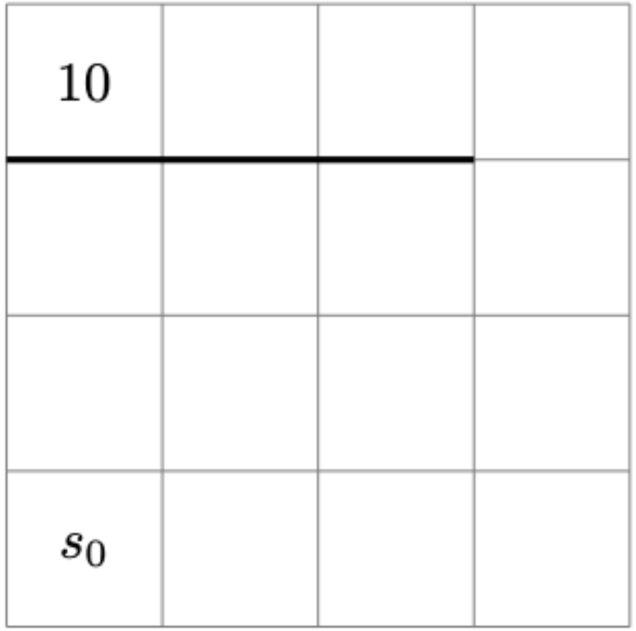
$$= \mathbb{E}_{s\sim d^{\pi^\star}} A^{\hat{\pi}}(s, \pi^\star(s)) - \mathbb{E}_{s\sim d^{\pi^\star}} A^{\hat{\pi}}(s, \hat{\pi}(s))$$

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

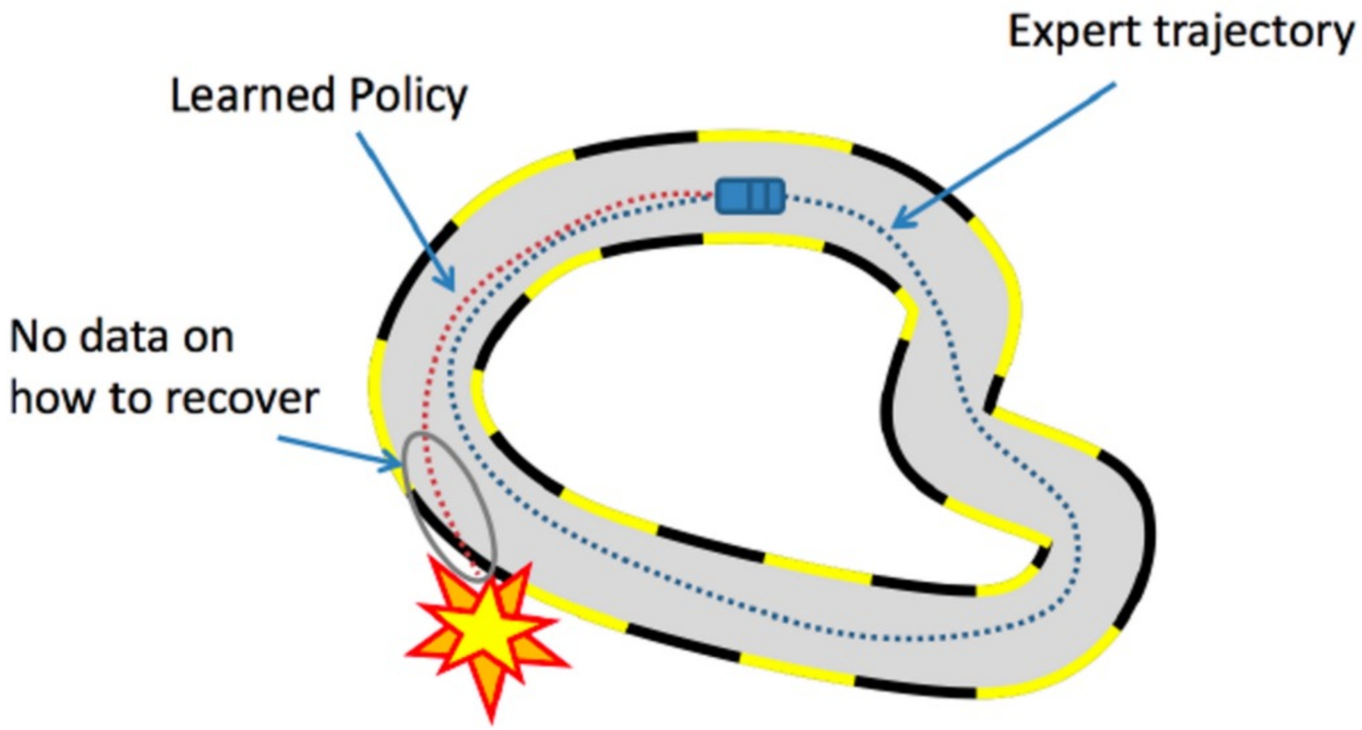
$$\leq \frac{2}{1-\gamma}\epsilon$$

The Distribution Shift problem in BC

- Let’s see an example



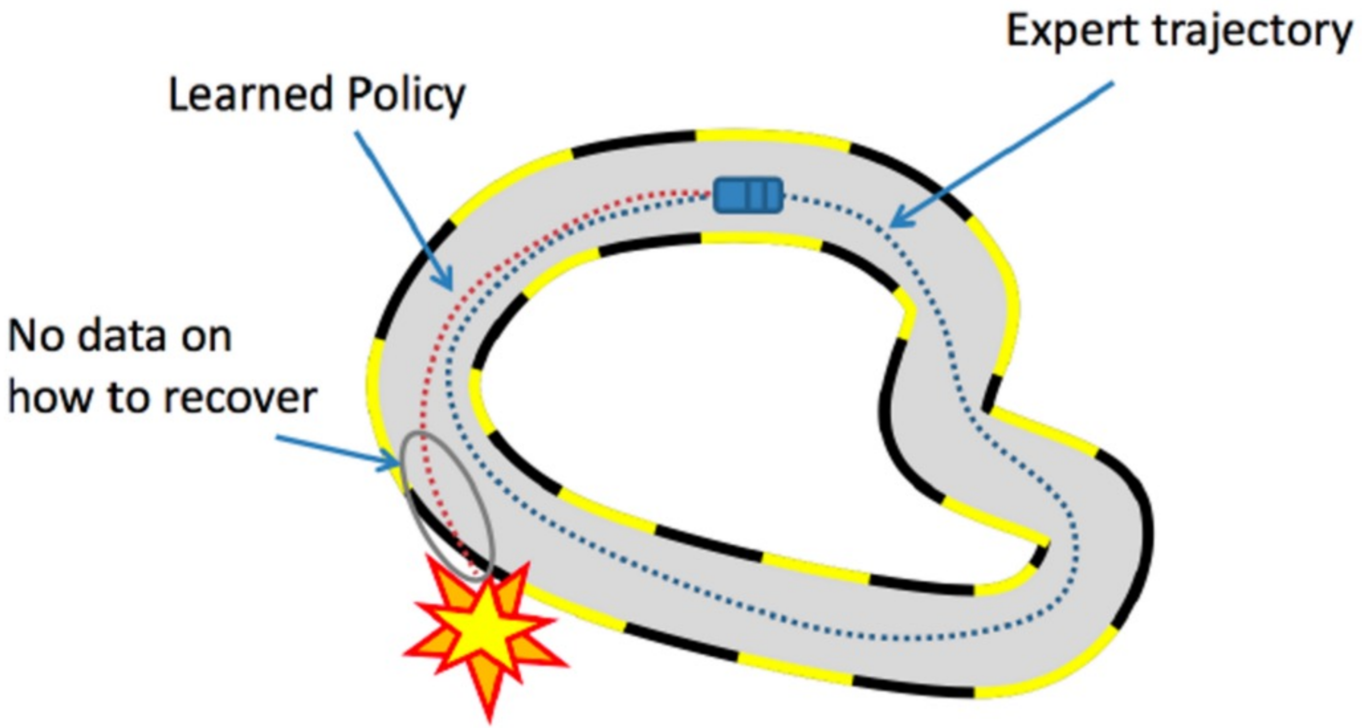
The Distribution Shift problem in BC



- This is fundamental to offline RL/IL.

How to prevent it?

- Naïve approach: expert demonstrations from all possible starting states.



Analysis

SL guarantee: $\forall p, \mathbb{E}_{s\sim p}[\hat{\pi}(s) \neq \pi^\star(s)] \leq \epsilon \approx O(\sqrt{1/N})$

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

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

Proof: Performance Difference Lemma: $(1-\gamma)\big(f(\pi) - f(\pi')\big) = \mathbb{E}_{s,a\sim d^\pi}\left[A^{\pi'}(s,a)\right]$

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

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

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

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

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

$$\leq -\max_{s,a} A^{\pi^\star}(s,a) \mathbb{E}_{s\sim d^{\hat{\pi}}} \mathbf{1}\{\hat{\pi}(s) \neq \pi^\star(s)\}$$

$$\leq \epsilon \max_{s,a} |A^{\pi^\star}(s,a)|$$