Is Behavior Cloning All You Need?

Revisiting the Role of Horizon and Interaction in Imitation Learning

Dylan Foster

Microsoft Research

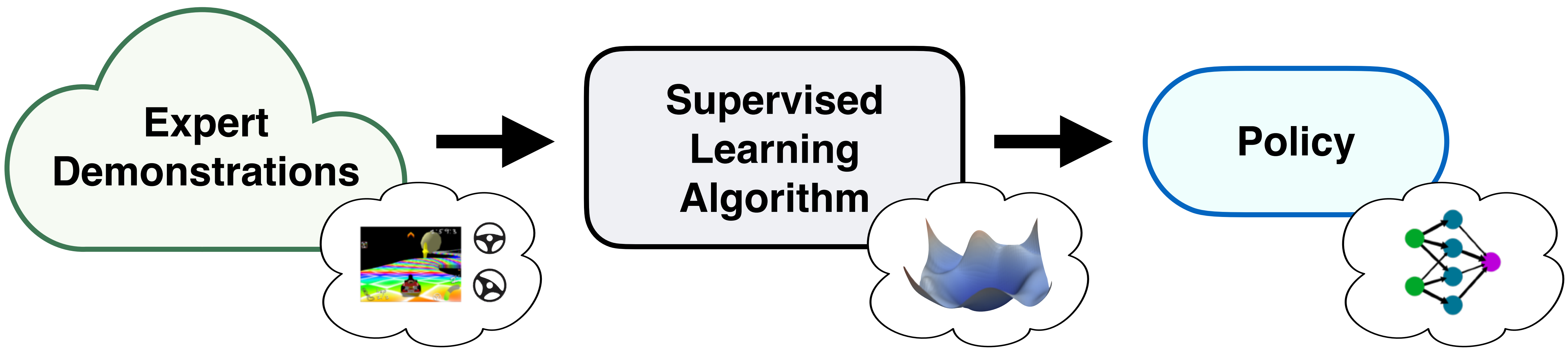
Based on work with Adam Block and Dipendra Misra

arXiv: 2407.15007

Imitation learning

Given: Expert demonstrations or access to demonstrator.

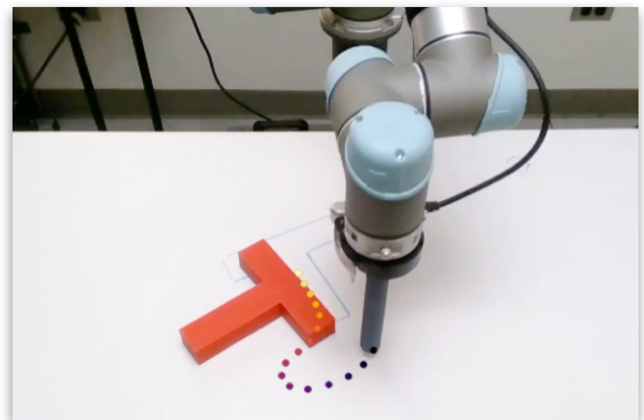
Goal: Learn policy to imitate expert behavior.



Imitation learning

Autonomous vehicles

Robotics



Language modeling

GPT-4

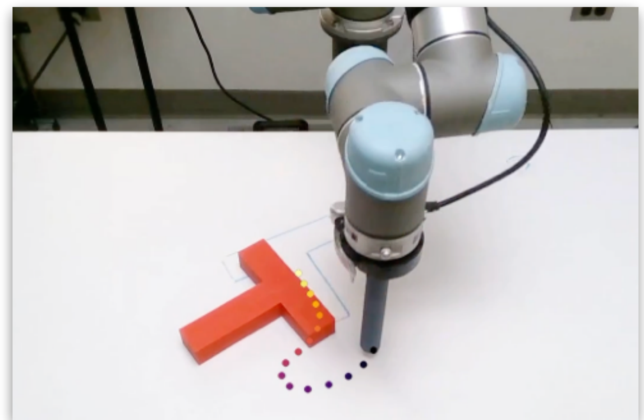
Benefits of imitation learning:

- Easier to demonstrate desired behavior versus design reward function to elicit.
- More sample-efficient and stable than RL.

Imitation learning

Autonomous vehicles

Robotics



Language modeling

Benefits of imitation learning:

- Easier to demonstrate desired behavior versus design reward function to elicit.
- More sample-efficient and stable than RL.

This talk: Revisiting theoretical foundations of imitation learning.

Two frameworks for imitation learning

Offline imitation learning: Only have logged trajectories from expert.

Online/interactive imitation learning: Can interactively query expert.

Offline imitation learning

Setup:

- Finite-horizon MDP $M = (\mathcal{X}, \mathcal{A}, H, P, r)$.

Offline imitation learning

Setup:

- Finite-horizon MDP $M = (\mathcal{X}, \mathcal{A}, H, P, r)$.
- Expert policy $\pi^\star = (\pi^\star_h)_{h=1}^H$, where $\pi^\star_h : \mathcal{X} \to \Delta(\mathcal{A})$. (potentially stochastic)
- Have dataset $\mathcal{D} = \left\{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\right\}_{i=1}^n$ of n trajectories generated from $\pi^\star$.

Offline imitation learning

Setup:

- Finite-horizon MDP $M = (\mathcal{X}, \mathcal{A}, H, P, r)$.
- Expert policy $\pi^\star = (\pi^\star_h)_{h=1}^H$, where $\pi^\star_h : \mathcal{X} \to \Delta(\mathcal{A})$. (potentially stochastic)
- Have dataset $\mathcal{D} = \left\{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\right\}_{i=1}^n$ of n trajectories generated from $\pi^\star$.
- **Goal**: Learn policy $\widehat{\pi}$ such that

$$J(\pi^\star) - J(\widehat{\pi}) \leq \textsf{small},$$

where $J(\pi) = \mathbb{E}^\pi\left[\sum_{h=1}^H r_h(x_h, a_h)\right]$.

Offline imitation learning

Setup:

- Finite-horizon MDP $M = (\mathcal{X}, \mathcal{A}, H, P, r)$.
- Expert policy $\pi^\star = (\pi^\star_h)_{h=1}^H$, where $\pi^\star_h : \mathcal{X} \to \Delta(\mathcal{A})$. (potentially stochastic)
- Have dataset $\mathcal{D} = \left\{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\right\}_{i=1}^n$ of n trajectories generated from $\pi^\star$.
- **Goal**: Learn policy $\widehat{\pi}$ such that

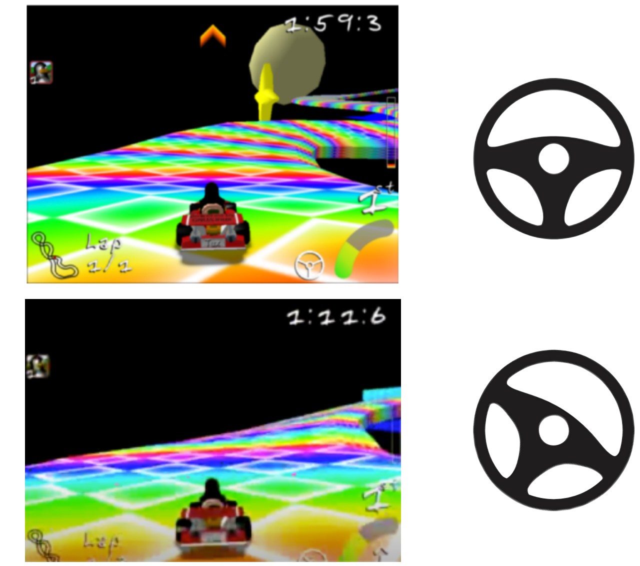
$$J(\pi^\star) - J(\widehat{\pi}) \leq \text{small},$$

where $J(\pi) = \mathbb{E}^\pi\left[\sum_{h=1}^H r_h(x_h, a_h)\right]$.

Need to imitate expert's behavior based on demonstrations alone.

(can't directly interact with M and $\pi^\star$, no reward-based feedback).

Offline IL: Behavior cloning



[Ross & Bagnell '10]

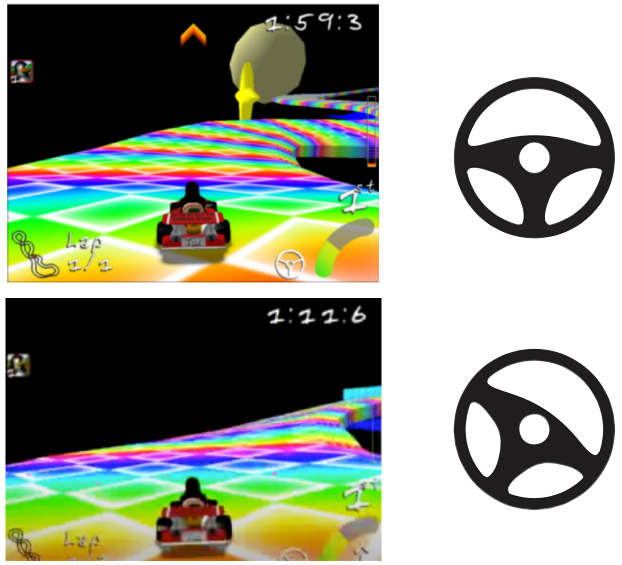
Behavior cloning: [Pomerleau '89, Ross & Bagnell '10]

- Given dataset $\mathcal{D} = \left\{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\right\}_{i=1}^n$, solve

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{n} \sum_{h=1}^{H} \mathbb{I}\{\pi_h(x_h^i) \neq a_h^i\}$$

(can use other losses, e.g., square or log)

Offline IL: Behavior cloning



[Ross & Bagnell '10]

Behavior cloning: [Pomerleau '89, Ross & Bagnell '10]

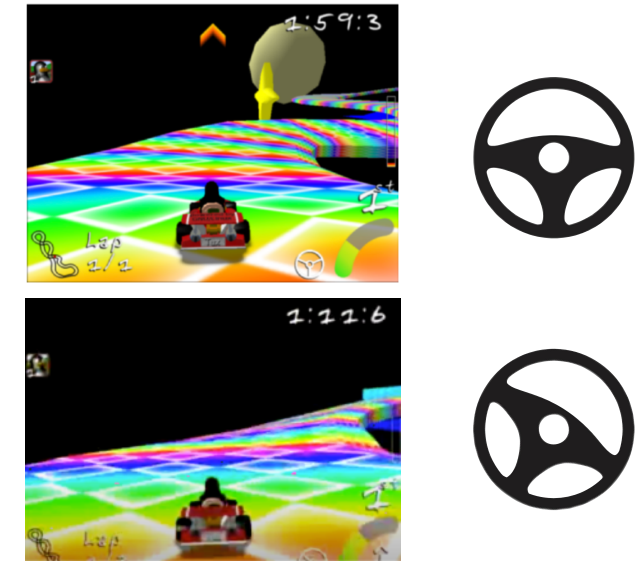
- Given dataset $\mathcal{D} = \{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\}_{i=1}^{n}$, solve

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{n} \sum_{h=1}^{H} \mathbb{I}\{\pi_h(x_h^i) \neq a_h^i\}$$

(can use other losses, e.g., square or log)

✓ Appealing simplicity, essentially a "reduction" to supervised learning.

Offline IL: Behavior cloning



[Ross & Bagnell '10]

Behavior cloning: [Pomerleau '89, Ross & Bagnell '10]

- Given dataset $\mathcal{D} = \{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\}_{i=1}^n$, solve

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{n} \sum_{h=1}^{H} \mathbb{I}\{\pi_h(x_h^i) \neq a_h^i\}$$

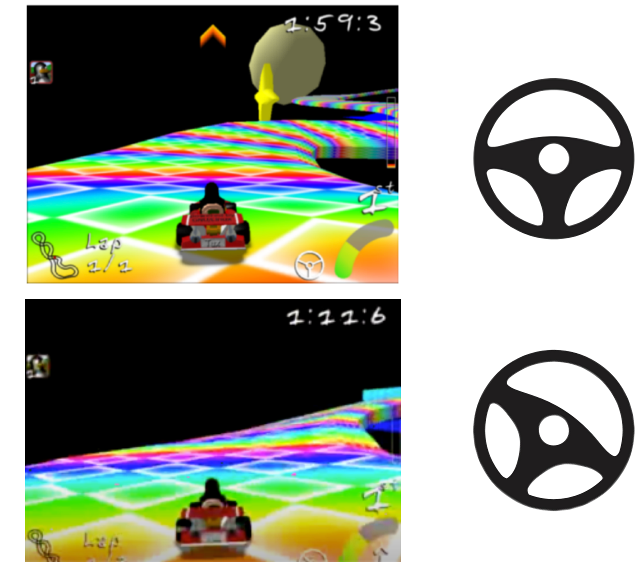
(can use other losses, e.g., square or log)

✓ Appealing simplicity, essentially a "reduction" to supervised learning.

✗ Seemingly ignores problem of distribution shift.

- Deploying $\widehat{\pi}$ creates feedback loop; different distribution at train vs. test time.
- Supervised learning errors compound ("error amplification"); leads to instability.

Offline IL: Behavior cloning



[Ross & Bagnell '10]

Behavior cloning: [Pomerleau '89, Ross & Bagnell '10]

- Given dataset $\mathcal{D} = \{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\}_{i=1}^n$, solve

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{n} \sum_{h=1}^{H} \mathbb{I}\{\pi_h(x_h^i) \neq a_h^i\}$$

(can use other losses, e.g., square or log)

✓ Appealing simplicity, essentially a "reduction" to supervised learning.

✗ Seemingly ignores problem of distribution shift.

- Deploying $\widehat{\pi}$ creates feedback loop; different distribution at train vs. test time.
- Supervised learning errors compound ("error amplification"); leads to instability.

In theory: Need **quadratic** $\Omega(H^2)$ trajectories to learn a good policy [RB '10].

(Assuming $r_h \in [0,1]$)

Online imitation learning

Setup:

- Can directly interact with MDP M and expert $\pi^\star$ for n episodes.

  (don't observe rewards, just states and actions)

- **Goal**: Learn policy $\widehat{\pi}$ such that $J(\pi^\star) - J(\widehat{\pi}) \leq \textsf{small}$.

Online imitation learning

Setup:

- Can directly interact with MDP M and expert $\pi^\star$ for n episodes.

  (don't observe rewards, just states and actions)

- **Goal**: Learn policy $\widehat{\pi}$ such that $J(\pi^\star) - J(\widehat{\pi}) \leq \textsf{small}$.

Online IL algorithms:

- Dagger [Ross, Gordon, Bagnell '11], Aggrevate [Ross & Bagnell '14], etc.
- Roll in with $\widehat{\pi}$, ask expert $\pi^\star$ for feedback, update, ...
- Learn to correct mistakes in $\widehat{\pi}$ on-policy, avoiding distribution shift.

In theory: Can achieve **linear** $O(H)$ sample comp. for "recoverable" MDPs.

Our question

Is online IL truly more sample-efficient than offline IL?

Or can existing algorithms and analyses be improved?

Our question

Is online IL truly more sample-efficient than offline IL?

Or can existing algorithms and analyses be improved?

Focus on horizon.

See also [Rajaraman et al. '20, '21] for tabular/linear policies.

Why I care about this problem

Connection to autoregressive language modeling

- Can use other losses for behavior cloning, e.g. log loss:

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{n} \sum_{h=1}^{H} -\log(\pi_h(a_h^i \mid x_h^i)).$$

- Next-token prediction (pretraining/SFT) for autoregressive language models (LLMs) is a special case where $x_h = a_1, \ldots, a_{h-1}$ ("token-level MDP").

- Similar phenomena (error amplification, instability, ...) in both domains [Holtzmann et al. '19, Braverman et al. '20, Block-**F**-Krishnamurthy-Simchowitz-Zhang '24].

Why I care about this problem

Connection to autoregressive language modeling

- Can use other losses for behavior cloning, e.g. log loss:

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{n} \sum_{h=1}^{H} -\log(\pi_h(a_h^i \mid x_h^i)).$$

- Next-token prediction (pretraining/SFT) for autoregressive language models (LLMs) is a special case where $x_h = a_1, \ldots, a_{h-1}$ ("token-level MDP").
- Similar phenomena (error amplification, instability, ...) in both domains [Holtzmann et al. '19, Braverman et al. '20, Block-**F**-Krishnamurthy-Simchowitz-Zhang '24].

IL + RL fine-tuning as a new paradigm for decision making?

- Embodied decision making (self-driving, robotics, ...)
- Symbolic decision making (LLMs, AI agents, game playing, ...)

Outline

1. **Revisit analysis of behavior cloning (focusing on horizon)**
2. **Behavior cloning is better than you might think**
3. **Discussion and implications**

Outline

1. **Revisit analysis of behavior cloning (focusing on horizon)**
2. **Behavior cloning is better than you might think**
3. **Discussion and implications**

Analysis of behavior cloning

Behavior cloning: [Pomerleau '89, Ross & Bagnell '10]

- Given dataset $\mathcal{D} = \left\{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\right\}_{i=1}^n$, solve

$$\widehat{\pi} = \operatorname*{arg\,min}_{\pi \in \Pi} \widehat{L}_{\mathsf{bc}}(\pi) := \frac{1}{n}\sum_{i=1}^n \frac{1}{H}\sum_{h=1}^H \mathbb{I}\{\pi_h(x_h^i) \neq a_h^i\}$$

Analysis of behavior cloning

Behavior cloning: [Pomerleau '89, Ross & Bagnell '10]

- Given dataset $\mathcal{D} = \left\{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\right\}_{i=1}^n$, solve

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \widehat{L}_{\mathsf{bc}}(\pi) := \frac{1}{n}\sum_{i=1}^n \frac{1}{H}\sum_{h=1}^H \mathbb{I}\{\pi_h(x_h^i) \neq a_h^i\}$$

Assumptions for analysis:

- Expert $\pi^\star$ is deterministic.

Analysis of behavior cloning

Behavior cloning: [Pomerleau '89, Ross & Bagnell '10]

- Given dataset $\mathcal{D} = \left\{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\right\}_{i=1}^n$, solve

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \widehat{L}_{\mathsf{bc}}(\pi) := \frac{1}{n}\sum_{i=1}^n \frac{1}{H}\sum_{h=1}^H \mathbb{I}\{\pi_h(x_h^i) \neq a_h^i\}$$

Assumptions for analysis:

- Expert $\pi^\star$ is deterministic.
- Realizability: $\pi^\star \in \Pi$.

Analysis of behavior cloning

Behavior cloning: [Pomerleau '89, Ross & Bagnell '10]

- Given dataset $\mathcal{D} = \left\{(x_1^i, a_1^i, \ldots, x_H^i, a_H^i)\right\}_{i=1}^n$, solve

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \widehat{L}_{\mathsf{bc}}(\pi) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{H} \sum_{h=1}^{H} \mathbb{I}\{\pi_h(x_h^i) \neq a_h^i\}$$

Assumptions for analysis:

- Expert $\pi^\star$ is deterministic.
- Realizability: $\pi^\star \in \Pi$.
- Rewards have $\sum_{h=1}^H r_h \in [0, R]$ and $r_h \in [0, 1]$.
  - Sparse rewards (e.g., single reward at goal): $R = O(1)$.
  - Dense rewards ($r_h \in [0, 1]$ for all h): $R = H$.

[Jiang-Agarwal '18, Wang et al. '20]

Analysis of behavior cloning

Assumptions for analysis:

- Expert $\pi^\star$ is deterministic.
- Realizability: $\pi^\star \in \Pi$.
- Rewards have $\sum_{h=1}^{H} r_h \in [0, R]$.

Analysis of behavior cloning

Assumptions for analysis:

- Expert $\pi^\star$ is deterministic.
- Realizability: $\pi^\star \in \Pi$.
- Rewards have $\sum_{h=1}^{H} r_h \in [0, R]$.

Standard analysis of BC: Define $L_{\mathsf{bc}}(\pi) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{P}^{\pi^\star}\left[\pi(x_h) \neq \pi^\star(x_h)\right]$.

Analysis of behavior cloning

Assumptions for analysis:

- Expert $\pi^\star$ is deterministic.
- Realizability: $\pi^\star \in \Pi$.
- Rewards have $\sum_{h=1}^{H} r_h \in [0, R]$.

Standard analysis of BC: Define $L_{\mathsf{bc}}(\pi) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{P}^{\pi^\star}\left[\pi(x_h) \neq \pi^\star(x_h)\right]$.

1. By uniform convergence, $L_{\mathsf{bc}}(\widehat{\pi}) \lesssim \frac{\log|\Pi|}{n}$ whp. (**supervised learning guarantee**)

Analysis of behavior cloning

Assumptions for analysis:

- Expert $\pi^\star$ is deterministic.
- Realizability: $\pi^\star \in \Pi$.
- Rewards have $\sum_{h=1}^{H} r_h \in [0, R]$.

Standard analysis of BC: Define $L_{\mathsf{bc}}(\pi) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{P}^{\pi^\star}\left[\pi(x_h) \neq \pi^\star(x_h)\right]$.

1. By uniform convergence, $L_{\mathsf{bc}}(\widehat{\pi}) \lesssim \frac{\log|\Pi|}{n}$ whp. (**supervised learning guarantee**)
2. Performance difference lemma: $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH \cdot L_{\mathsf{bc}}(\widehat{\pi})$.

Analysis of behavior cloning

Assumptions for analysis:

- Expert $\pi^\star$ is deterministic.
- Realizability: $\pi^\star \in \Pi$.
- Rewards have $\sum_{h=1}^{H} r_h \in [0, R]$.

Standard analysis of BC: Define $L_{\mathsf{bc}}(\pi) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{P}^{\pi^\star}\left[\pi(x_h) \neq \pi^\star(x_h)\right]$.

1. By uniform convergence, $L_{\mathsf{bc}}(\widehat{\pi}) \lesssim \frac{\log|\Pi|}{n}$ whp. (**supervised learning guarantee**)
2. Performance difference lemma: $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH \cdot L_{\mathsf{bc}}(\widehat{\pi})$.
3. Combining gives $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH\frac{\log|\Pi|}{n}$.

[Ross & Bagnell ’10]

✗ Under dense rewards $(R = H)$, need $\Omega(H^2)$ trajectories for constant accuracy.

Terminology note: We will refer to $R \cdot H$ as “quadratic” sample complexity.

Analysis of behavior cloning

Standard analysis of BC: Define $L_{\mathsf{bc}}(\pi) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{P}^{\pi^\star}\left[\pi(x_h) \neq \pi^\star(x_h)\right]$.

1. By uniform convergence, $L_{\mathsf{bc}}(\widehat{\pi}) \lesssim \frac{\log|\Pi|}{n}$ whp. (**supervised learning guarantee**)
2. Performance difference lemma: $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH \cdot L_{\mathsf{bc}}(\widehat{\pi})$.
3. Combining gives $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH\frac{\log|\Pi|}{n}$.

Analysis of behavior cloning

Standard analysis of BC: Define $L_{\mathsf{bc}}(\pi) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{P}^{\pi^\star}\left[\pi(x_h) \neq \pi^\star(x_h)\right]$.

1. By uniform convergence, $L_{\mathsf{bc}}(\widehat{\pi}) \lesssim \frac{\log|\Pi|}{n}$ whp. (**supervised learning guarantee**)
2. Performance difference lemma: $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH \cdot L_{\mathsf{bc}}(\widehat{\pi})$.
3. Combining gives $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH\frac{\log|\Pi|}{n}$.

Discussion:

- Step **(1)** is tight even when $|\Pi| = 2$ (dependent nature of trajectories).
  - Even though $\mathcal{D}$ contains $n \cdot H$ examples $(x_h, \pi^\star(x_h))$, get $1/n$ rate instead of $1/(nH)$ due to dependence.

Analysis of behavior cloning

Standard analysis of BC: Define $L_{\mathsf{bc}}(\pi) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{P}^{\pi^\star}\left[\pi(x_h) \neq \pi^\star(x_h)\right]$.

1. By uniform convergence, $L_{\mathsf{bc}}(\widehat{\pi}) \lesssim \frac{\log|\Pi|}{n}$ whp. (**supervised learning guarantee**)
2. Performance difference lemma: $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH \cdot L_{\mathsf{bc}}(\widehat{\pi})$.
3. Combining gives $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH\frac{\log|\Pi|}{n}$.

Discussion:

- Step **(1)** is tight even when $|\Pi| = 2$ (dependent nature of trajectories).
  - Even though $\mathcal{D}$ contains $n \cdot H$ examples $(x_h, \pi^\star(x_h))$, get $1/n$ rate instead of $1/(nH)$ due to dependence.
- Step **(2)** is tight for MDPs with 3 states [Ross & Bagnell '10]
  - Distribution shift?

Analysis of behavior cloning

Standard analysis of BC: Define $L_{\mathsf{bc}}(\pi) = \frac{1}{H}\sum_{h=1}^{H}\mathbb{P}^{\pi^\star}\left[\pi(x_h) \neq \pi^\star(x_h)\right]$.

1. By uniform convergence, $L_{\mathsf{bc}}(\widehat{\pi}) \lesssim \frac{\log|\Pi|}{n}$ whp. (**supervised learning guarantee**)
2. Performance difference lemma: $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH \cdot L_{\mathsf{bc}}(\widehat{\pi})$.
3. Combining gives $J(\pi^\star) - J(\widehat{\pi}) \lesssim RH\frac{\log|\Pi|}{n}$.

Discussion:

- Step **(1)** is tight even when $|\Pi| = 2$ (dependent nature of trajectories).
  - Even though $\mathcal{D}$ contains $n \cdot H$ examples $(x_h, \pi^\star(x_h))$, get $1/n$ rate instead of $1/(nH)$ due to dependence.
- Step **(2)** is tight for MDPs with 3 states [Ross & Bagnell '10]
  - Distribution shift?

Out of luck?

Online imitation learning: Dagger

$$Q_h^{\pi^\star}(x,a) = \mathbb{E}^{\pi^\star}\left[\sum_{h'=h}^{H} r_{h'} \mid x_h = x, a_h = a\right]$$

- Define *recoverability constant*

$$\mu_{\mathsf{rec}} = \max_{x\in\mathcal{X}, a\in\mathcal{A}, h\in[H]} \left\{ (Q_h^{\pi^\star}(x,\pi_h^\star(x)) - Q_h^{\pi^\star}(x,a))_+ \right\} \in [0, R].$$

- **Intuition:** $\mu_{\mathsf{rec}} = O(1)$ if we can quickly recover from taking a bad action.

  Ex: H independent contextual bandits problems. $R = H$ but $\mu_{\mathsf{rec}} = 1$.

Online imitation learning: Dagger

$$Q_h^{\pi^\star}(x,a) = \mathbb{E}^{\pi^\star}\left[\sum_{h'=h}^{H} r_{h'} \mid x_h = x, a_h = a\right]$$

- Define *recoverability constant*

$$\mu_{\mathsf{rec}} = \max_{x \in \mathcal{X}, a \in \mathcal{A}, h \in [H]} \left\{ (Q_h^{\pi^\star}(x, \pi_h^\star(x)) - Q_h^{\pi^\star}(x,a))_+ \right\} \in [0, R].$$

- **Intuition:** $\mu_{\mathsf{rec}} = O(1)$ if we can quickly recover from taking a bad action.

  Ex: H independent contextual bandits problems. $R = H$ but $\mu_{\mathsf{rec}} = 1$.

- Dagger [Ross, Gordon, & Bagnell '11] uses online expert access to achieve

$$J(\pi^\star) - J(\widehat{\pi}) \lesssim \mu_{\mathsf{rec}} H \cdot \frac{\log|\Pi|}{n}.$$

  Other online IL methods (e.g., Aggrevate) have similar guarantees.

- Improves over BC whenever $\mu_{\mathsf{rec}} \ll R$ (recall: $\mu_{\mathsf{rec}} H$ vs. RH).

Our result

Behavior cloning is horizon-independent

(if you use the **log loss**)

Outline

1. Revisit analysis of behavior cloning (focusing on horizon)
2. **Behavior cloning is better than you might think**
3. Discussion and implications

Log loss behavior cloning

Behavior cloning with logarithmic loss:

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \widehat{L}_{\mathsf{log}}(\pi) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{H} \sum_{h=1}^{H} \log\left(\frac{1}{\pi(a_h^i \mid x_h^i)} \right)$$

Log loss behavior cloning

Behavior cloning with logarithmic loss:

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \widehat{L}_{\mathsf{log}}(\pi) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{H}\sum_{h=1}^{H} \log\left(\frac{1}{\pi(a_h^i \mid x_h^i)}\right)$$

Intuition:

- Suppose $\pi^\star$ and π are both deterministic.
- Then

$$\widehat{L}_{\mathsf{log}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{H}\sum_{h=1}^{H} \log\left(\frac{1}{\pi(a_h^i \mid x_h^i)}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_\infty\{\exists h : \pi(x_h) \neq \pi^\star(x_h)\}.$$

Log loss behavior cloning

Behavior cloning with logarithmic loss:

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \widehat{L}_{\mathsf{log}}(\pi) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H}\sum_{h=1}^{H}\log\left(\frac{1}{\pi(a_h^i \mid x_h^i)}\right)$$

Intuition:

- Suppose $\pi^\star$ and π are both deterministic.
- Then

$$\widehat{L}_{\mathsf{log}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H}\sum_{h=1}^{H}\log\left(\frac{1}{\pi(a_h^i \mid x_h^i)}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_\infty\{\exists h : \pi(x_h) \neq \pi^\star(x_h)\}.$$

Penalizes deviations more aggressively than indicator loss BC, since

$$\mathbb{I}\{\exists h : \pi(x_h) \neq \pi^\star(x_h)\} \geq \frac{1}{H}\sum_{h=1}^{H}\mathbb{I}\{\pi(x_h) \neq \pi^\star(x_h)\}.$$

Get loss 1 if we deviate *anywhere* along trajectory (vs avg. loss along trajectory).

Log loss behavior cloning

Behavior cloning with logarithmic loss:

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \widehat{L}_{\mathsf{log}}(\pi) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{H}\sum_{h=1}^{H} \log\left(\frac{1}{\pi(a_h^i \mid x_h^i)}\right)$$

Supervised learning guarantee:

- Hellinger distance: $D^2_{\mathsf{Hel}}(P,Q) = \sum_z (\sqrt{P(z)} - \sqrt{Q(z)})^2$.

Log loss behavior cloning

Behavior cloning with logarithmic loss:

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \widehat{L}_{\mathsf{log}}(\pi) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{H}\sum_{h=1}^{H} \log\left(\frac{1}{\pi(a_h^i \mid x_h^i)}\right)$$

Supervised learning guarantee:

- Hellinger distance: $D^2_{\mathsf{Hel}}(P, Q) = \sum_z (\sqrt{P(z)} - \sqrt{Q(z)})^2$.
- Define $\mathbb{P}^{\pi} = \mathbb{P}^{\pi}\left[(x_1, a_1), \ldots, (x_H, a_H) = \cdot\right]$. Equivalent to MLE over set $\{\mathbb{P}^{\pi}\}_{\pi \in \Pi}$.

Log loss behavior cloning

Behavior cloning with logarithmic loss:

$$\widehat{\pi} = \underset{\pi \in \Pi}{\arg\min}\, \widehat{L}_{\mathsf{log}}(\pi) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H}\sum_{h=1}^{H}\log\left(\frac{1}{\pi(a_h^i \mid x_h^i)}\right)$$

Supervised learning guarantee:

- Hellinger distance: $D^2_{\mathsf{Hel}}(P, Q) = \sum_z (\sqrt{P(z)} - \sqrt{Q(z)})^2$.
- Define $\mathbb{P}^{\pi} = \mathbb{P}^{\pi}\left[(x_1, a_1), \ldots, (x_H, a_H) = \cdot\right]$. Equivalent to MLE over set $\{\mathbb{P}^{\pi}\}_{\pi \in \Pi}$.

Supervised learning guarantee (Wong & Shen ’95, van de Geer ’00, Zhang ’06)

As long as $\pi^\star \in \Pi$, with probability at least $1 - \delta$, Log-Loss BC satisfies

$$D^2_{\mathsf{Hel}}\left(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\right) \leq \frac{2\log(|\Pi|\delta^{-1})}{n}.$$

Holds regardless of whether $\pi^\star$ is deterministic or stochastic.

Main result (deterministic case)

Theorem (Horizon-independent regret; deterministic case)

For any deterministic expert policy $\pi^\star$ and potentially stochastic imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 4R \cdot D^2_{\mathsf{Hel}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big).$$

No explicit dependence on horizon!

Main result (deterministic case)

Theorem (Horizon-independent regret; deterministic case)

For any deterministic expert policy $\pi^\star$ and potentially stochastic imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 4R \cdot D^2_{\mathsf{Hel}}\left(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\right).$$

No explicit dependence on horizon!

Corollary (Horizon-independent regret for Log-Loss BC)

For any deterministic expert $\pi^\star \in \Pi$, Log-Loss BC ensures with prob at least $1-\delta$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 8R \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Features:

- Tightest known guarantee for IL with general policy classes.
- Improves rate for vanilla BC ($RH \cdot \frac{\log(|\Pi|\delta^{-1})}{n}$) and Dagger ($\mu_{\mathsf{rec}} H \cdot \frac{\log(|\Pi|\delta^{-1})}{n}$).

Interpreting the main theorem

Corollary (Horizon-independent regret for Log Loss BC)

For any deterministic expert $\pi^\star \in \Pi$, Log-Loss BC ensures with prob at least $1-\delta$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 8R \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Special case: Stationary policies (or parameter sharing)

- Ex: $\pi_\theta(x_h) = \arg\max_{a \in \mathcal{A}} \langle \theta, \phi(x_h, a) \rangle$ ($\theta \in \mathbb{R}^d$; d parameters)

Interpreting the main theorem

Corollary (Horizon-independent regret for Log Loss BC)

For any deterministic expert $\pi^\star \in \Pi$, Log-Loss BC ensures with prob at least $1-\delta$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 8R \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Special case: Stationary policies (or parameter sharing)

- Ex: $\pi_\theta(x_h) = \arg\max_{a\in\mathcal{A}} \langle \theta, \phi(x_h, a) \rangle$ ($\theta \in \mathbb{R}^d$; d parameters)
- Typically have $\log|\Pi| = O(1)$.
  - All stationary policies in a tabular MDP: $\log|\Pi| = |\mathcal{X}| \log|\mathcal{A}|$.
  - Linear policies: $\log|\Pi| = \widetilde{O}(d)$.

Interpreting the main theorem

Corollary (Horizon-independent regret for Log Loss BC)

For any deterministic expert $\pi^\star \in \Pi$, Log-Loss BC ensures with prob at least $1-\delta$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 8R \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Special case: Stationary policies (or parameter sharing)

- Ex: $\pi_\theta(x_h) = \arg\max_{a \in \mathcal{A}} \langle \theta, \phi(x_h, a) \rangle$ ($\theta \in \mathbb{R}^d$; d parameters)
- Typically have $\log|\Pi| = O(1)$.
  - All stationary policies in a tabular MDP: $\log|\Pi| = |\mathcal{X}| \log|\mathcal{A}|$.
  - Linear policies: $\log|\Pi| = \widetilde{O}(d)$.
- **Sparse rewards**: For $R = O(1)$, $O(\frac{1}{\varepsilon})$ trajectories suffice for ε-optimal policy.
- **Dense rewards**: For $R = H$, $O(\frac{H}{\varepsilon})$ trajectories suffice.

Log-Loss BC beats the curse of horizon!

($O(H)$ under dense rewards)

Main result (deterministic case)

Theorem (Horizon-independent regret; deterministic case)

For any deterministic expert policy $\pi^\star$ and potentially stochastic imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 4R \cdot D^2_{\mathsf{Hel}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big).$$

No explicit dependence on horizon!

Corollary (Horizon-independent regret for Log-Loss BC)

For any deterministic expert $\pi^\star \in \Pi$, Log-Loss BC ensures with prob at least $1-\delta$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 8R \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Remarks:

- *Squared* Hellinger critical; $J(\pi^\star) - J(\widehat{\pi}) \leq R \cdot D_{\mathsf{TV}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big)$ trivially but no fast rate.
- Benefits of log-loss in RL: [**F** & Krishnamurthy '21], [Wang et al. '23/24], [Ayoub et al. '24].

Main result (stochastic case)

Define "variance" of expert policy: $\sigma^2_{\pi^\star} = \sum_{h=1}^{H} \mathbb{E}^{\pi^\star}\left[(V_h^{\pi^\star}(x_h) - Q_h^{\pi^\star}(x_h, a_h))^2 \right]$

Main result (stochastic case)

Define "variance" of expert policy: $\sigma_{\pi^\star}^2 = \sum_{h=1}^{H} \mathbb{E}^{\pi^\star}\left[\left(V_h^{\pi^\star}(x_h) - Q_h^{\pi^\star}(x_h, a_h)\right)^2\right]$

Theorem (Horizon-independent regret; stochastic case)

For any expert policy $\pi^\star$ and imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq \sqrt{6\sigma_{\pi^\star}^2 \cdot D^2_{\mathsf{Hel}}\left(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\right)} + \widetilde{O}\left(R \cdot D^2_{\mathsf{Hel}}\left(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\right)\right).$$

Main result (stochastic case)

Define "variance" of expert policy: $\sigma^2_{\pi^\star} = \sum_{h=1}^{H} \mathbb{E}^{\pi^\star}\left[(V_h^{\pi^\star}(x_h) - Q_h^{\pi^\star}(x_h, a_h))^2 \right]$

Theorem (Horizon-independent regret; stochastic case)

For any expert policy $\pi^\star$ and imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq \sqrt{6\sigma^2_{\pi^\star} \cdot D^2_{\mathsf{Hel}}\left(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\right)} + \widetilde{O}\left(R \cdot D^2_{\mathsf{Hel}}\left(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\right)\right).$$

Corollary (Horizon-independent regret for Log-Loss BC; stochastic case)

For any expert policy $\pi^\star \in \Pi$, Log Loss BC achieves with prob at least $1-\delta$:

$$J(\pi^\star) - J(\widehat{\pi}) \lesssim \sqrt{\frac{\sigma^2_{\pi^\star} \cdot \log(|\Pi|\delta^{-1})}{n}} + R\log(n) \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Main result (stochastic case)

Define "variance" of expert policy: $\sigma^2_{\pi^\star} = \sum_{h=1}^{H} \mathbb{E}^{\pi^\star}\left[\left(V_h^{\pi^\star}(x_h) - Q_h^{\pi^\star}(x_h, a_h)\right)^2\right]$

Corollary (Horizon-independent regret for Log-Loss BC; stochastic case)

For any expert policy $\pi^\star \in \Pi$, Log Loss BC achieves with prob at least $1-\delta$:

$$J(\pi^\star) - J(\widehat{\pi}) \lesssim \sqrt{\frac{\sigma^2_{\pi^\star} \cdot \log(|\Pi|\delta^{-1})}{n}} + R\log(n) \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Slower rate: $1/\sqrt{n}$ instead of $1/n$, but recovers deterministic case ($\sigma^2_{\pi^\star} = 0$).

Still horizon-independent:

1. Always have $\sigma^2_{\pi^\star} \leq R^2$ (law of total variance).
2. $\implies$ For ε-optimal policy, $\frac{R^2 \log(|\Pi|\delta^{-1})}{\varepsilon^2}$ trajectories suffice.

But worse dependence on R than deterministic case (fundamental).

Proof sketch

Theorem (Horizon-independent regret; deterministic case)

For any deterministic expert policy $\pi^\star$ and potentially stochastic imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 4R \cdot D^2_{\mathsf{Hel}}\left(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\right).$$

Proof sketch

Theorem (Horizon-independent regret; deterministic case)

For any deterministic expert policy $\pi^\star$ and potentially stochastic imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 4R \cdot D^2_{\mathsf{Hel}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big).$$

Define *trajectory-level* distance function between (potentially stochastic) policies π, π':

$$\rho(\pi \parallel \pi') := \mathbb{E}^{\pi}\, \mathbb{E}_{a'_1 \sim \pi'(x_1), \ldots, a'_H \sim \pi'(x_H)} \big[\mathbb{I}\{\exists h : a_h \neq a'_h\}\big],$$

Proof sketch

Theorem (Horizon-independent regret; deterministic case)

For any deterministic expert policy $\pi^\star$ and potentially stochastic imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \le 4R \cdot D^2_{\mathsf{Hel}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big).$$

Define *trajectory-level* distance function between (potentially stochastic) policies π, π':

$$\rho(\pi \parallel \pi') := \mathbb{E}^{\pi}\, \mathbb{E}_{a'_1 \sim \pi'(x_1), \ldots, a'_H \sim \pi'(x_H)}\big[\mathbb{I}\{\exists h : a_h \neq a'_h\}\big],$$

Proof:

1. For all (potentially stochastic) policies $\pi^\star$ and $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \le R \cdot \rho\big(\pi^\star \parallel \widehat{\pi}\big).$$

Proof sketch

Theorem (Horizon-independent regret; deterministic case)

For any deterministic expert policy $\pi^\star$ and potentially stochastic imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 4R \cdot D^2_{\mathsf{Hel}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big).$$

Define *trajectory-level* distance function between (potentially stochastic) policies π, π':

$$\rho(\pi \parallel \pi') := \mathbb{E}^{\pi}\, \mathbb{E}_{a'_1 \sim \pi'(x_1), \ldots, a'_H \sim \pi'(x_H)}\big[\mathbb{I}\{\exists h : a_h \neq a'_h\}\big],$$

Proof:

1. For all (potentially stochastic) policies $\pi^\star$ and $\widehat{\pi}$,
$$J(\pi^\star) - J(\widehat{\pi}) \leq R \cdot \rho\big(\pi^\star \parallel \widehat{\pi}\big).$$

2. Whenever $\pi^\star$ is deterministic, Hellinger distance satisfies
$$D^2_{\mathsf{Hel}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big) \geq \frac{1}{4} \cdot \rho\big(\widehat{\pi} \parallel \pi^\star\big).$$

Proof sketch

Theorem (Horizon-independent regret; deterministic case)

For any deterministic expert policy $\pi^\star$ and potentially stochastic imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 4R \cdot D^2_{\mathsf{Hel}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big).$$

Define *trajectory-level* distance function between (potentially stochastic) policies π, π':

$$\rho(\pi \parallel \pi') := \mathbb{E}^{\pi}\,\mathbb{E}_{a'_1 \sim \pi'(x_1), \ldots, a'_H \sim \pi'(x_H)}\big[\mathbb{I}\{\exists h : a_h \neq a'_h\}\big],$$

Proof:

1. For all (potentially stochastic) policies $\pi^\star$ and $\widehat{\pi}$,
$$J(\pi^\star) - J(\widehat{\pi}) \leq R \cdot \rho\big(\pi^\star \parallel \widehat{\pi}\big).$$
2. Whenever $\pi^\star$ is deterministic, Hellinger distance satisfies
$$D^2_{\mathsf{Hel}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big) \geq \frac{1}{4} \cdot \rho\big(\widehat{\pi} \parallel \pi^\star\big).$$
3. Trajectory-level distance is symmetric: $\rho\big(\widehat{\pi} \parallel \pi^\star\big) = \rho\big(\pi^\star \parallel \widehat{\pi}\big)$.

$\blacksquare$

See also [Rajaraman et al '21] (directly minimizes $\rho\big(\widehat{\pi} \parallel \pi^\star\big)$ for linear policies)

Outline

1. **Revisit analysis of behavior cloning (focusing on horizon)**
2. **Behavior cloning is better than you might think**
3. **Discussion**
    - Is this real?
    - Implications for online vs. offline IL

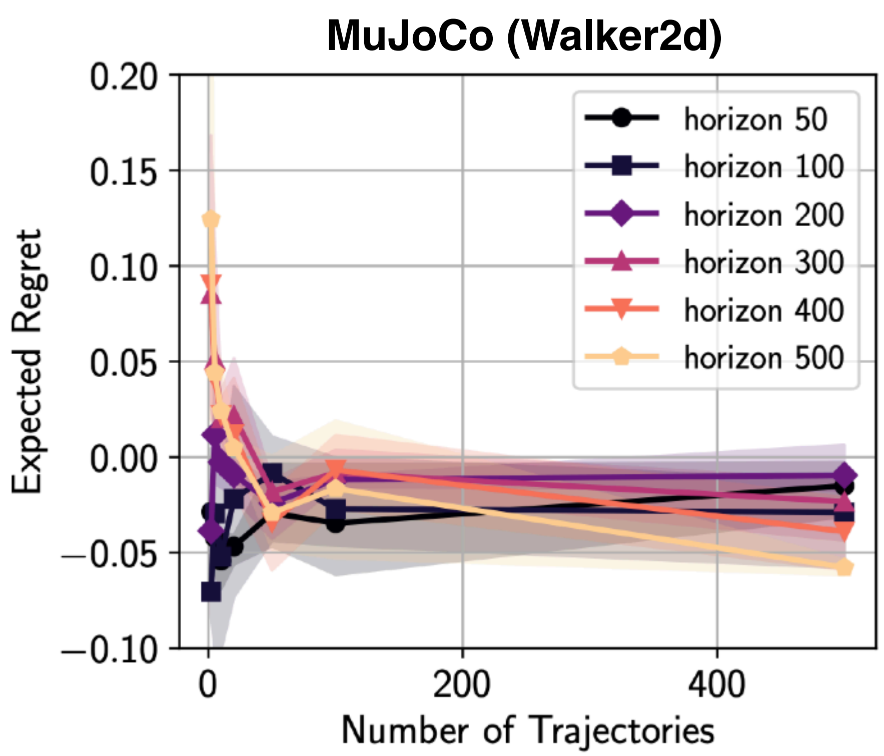
Outline

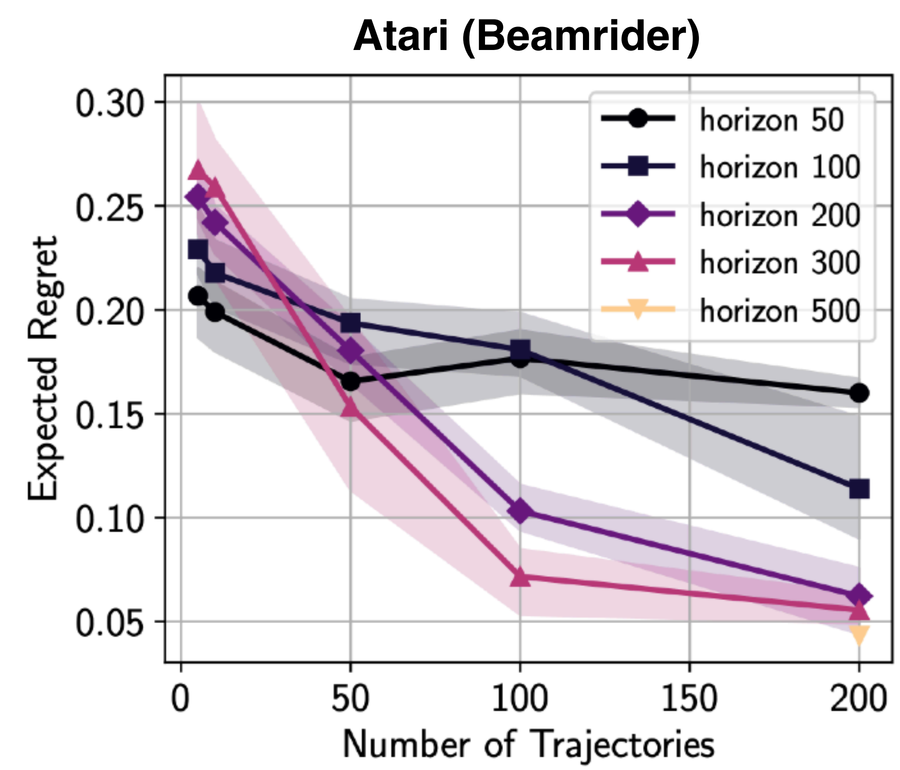
1. **Revisit analysis of behavior cloning (focusing on horizon)**
2. **Behavior cloning is better than you might think**
3. **Discussion**
    - Is this real?
    - Implications for online vs. offline IL

Experiments: Control

Setup:

- Multiple environments (discrete + continuous control, language)
- Train (stationary) expert $\pi^\star$ using RL.
- Train imitator $\widehat{\pi}$ using log-loss BC w/ same policy network architecture (randomly initialized).
- Repeat process for varying values of H (normalizing so $R = O(1)$).



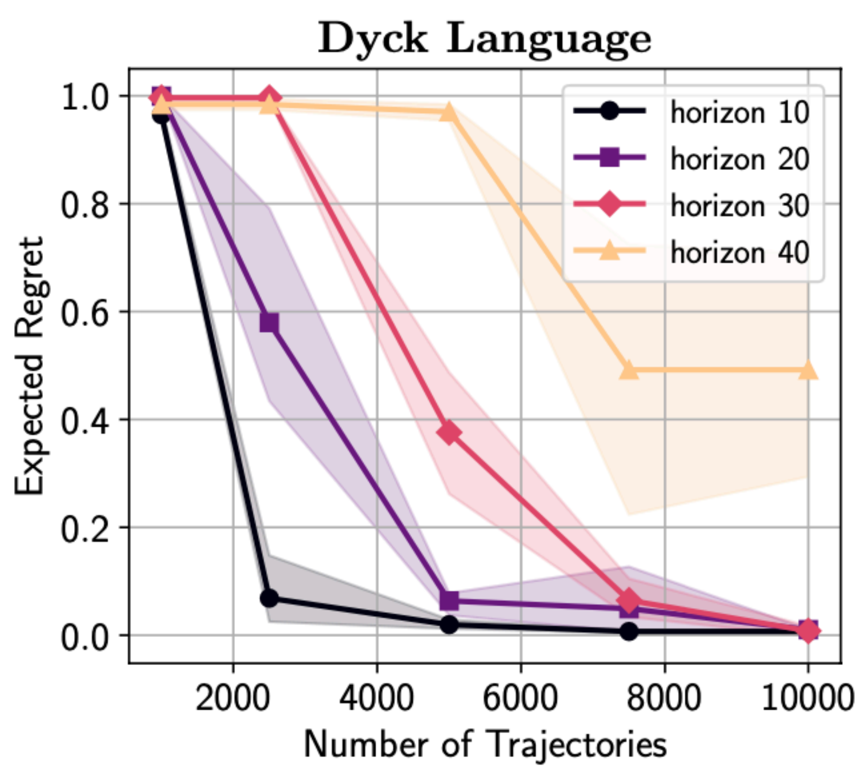


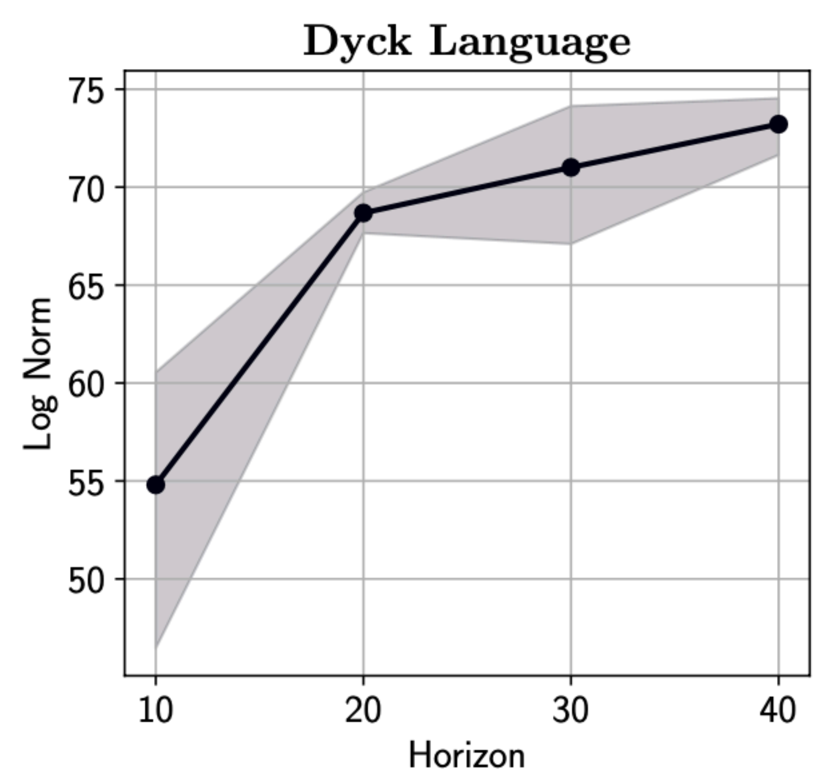
Learning curves are constant or improving as a function of horizon H!

Experiments: Language

Dyck language task:

- Goal of "agent": complete a valid word of a given length in a Dyck language Dyck_3.
    - Sequences of open and closed parentheses in '()', '[]', and '{}'.
    - Word is valid if parentheses are closed in correct order.
- Ex: '$([()])\{\}$' is a valid word, '$([)]$' and '$(((\{\}$' are not.
- Autoregressive task where $x_h = a_{1:h-1}$. Action a_h is next token.





Effect of horizon on performance is explained by supervised learning complexity for $\pi^\star$.

Outline

1. **Revisit analysis of behavior cloning (focusing on horizon)**
2. **Behavior cloning is better than you might think**
3. **Discussion**
   - Is this real?
   - Implications for online vs. offline IL

Main result (deterministic case)

Theorem (Horizon-independent regret; deterministic case)

For any deterministic expert policy $\pi^\star$ and potentially stochastic imitator policy $\widehat{\pi}$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 4R \cdot D^2_{\mathsf{Hel}}\big(\mathbb{P}^{\widehat{\pi}}, \mathbb{P}^{\pi^\star}\big).$$

No explicit dependence on horizon!

Corollary (Horizon-independent regret for Log-Loss BC)

For any deterministic expert $\pi^\star \in \Pi$, Log-Loss BC ensures with prob at least $1-\delta$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 8R \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Features:

- Tightest known guarantee for IL with general policy classes.
- Improves rate for vanilla BC ($RH \cdot \frac{\log(|\Pi|\delta^{-1})}{n}$) and Dagger ($\mu_{\mathsf{rec}} H \cdot \frac{\log(|\Pi|\delta^{-1})}{n}$).

Implications for online vs. offline IL

Theorem (Lower bound; deterministic expert case)

There exists Π with $|\Pi| = 2$ such that for any (online or offline) imitation learning algorithm, there exists a reward function $r = \{r_h\}_{h=1}^{H}$ with $r_h \in [0,1]$ (so $R \leq H$) and (optimal) deterministic expert policy $\pi^\star \in \Pi$ with $\mu_{\mathsf{rec}} = 1$ such that

$$\mathbb{E}\big[J(\pi^\star) - J(\widehat{\pi})\big] \gtrsim \frac{H}{n} = H \cdot \frac{\log|\Pi|}{n}.$$

In addition, the dynamics, rewards, and expert policies are stationary.

Implications for online vs. offline IL

Theorem (Lower bound; deterministic expert case)

There exists Π with $|\Pi| = 2$ such that for any (online or offline) imitation learning algorithm, there exists a reward function $r = \{r_h\}_{h=1}^{H}$ with $r_h \in [0,1]$ (so $R \leq H$) and (optimal) deterministic expert policy $\pi^\star \in \Pi$ with $\mu_{\mathsf{rec}} = 1$ such that

$$\mathbb{E}\left[J(\pi^\star) - J(\widehat{\pi})\right] \gtrsim \frac{H}{n} = H \cdot \frac{\log|\Pi|}{n}.$$

In addition, the dynamics, rewards, and expert policies are stationary.

Analogous lower bound for stochastic experts (Log-Loss BC regret is tight when $|\Pi| = 2$; online access cannot improve).

Implication: *Without further assumptions on Π, online imitation learning cannot improve upon offline imitation learning with Log-Loss BC*.

Online IL can still help for <u>some</u> classes Π [Rajaraman et al '20].

Benefits of online IL: No parameter sharing

Corollary (Horizon-independent regret for Log Loss BC)

For any deterministic expert $\pi^\star \in \Pi$, Log-Loss BC ensures with prob at least $1-\delta$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 8R \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Special case: Non-stationary policies (or no parameter sharing)

- Ex: $\pi_\theta(x_h) = \arg\max_{a \in \mathcal{A}} \langle \theta_h, \phi(x_h, a) \rangle$ ($\theta \in \mathbb{R}^d$; d parameters)
- Have $\Pi = \Pi_1 \times \cdots \Pi_H$, so $\log|\Pi| = \sum_{h=1}^{H} \log|\Pi_h| = O(H)$.
  - All non-stationary policies in a tabular MDP: $\log|\Pi| = H|\mathcal{X}|\log|\mathcal{A}|$.
  - Linear policies: $\log|\Pi| = \widetilde{O}(Hd)$.

Benefits of online IL: No parameter sharing

Corollary (Horizon-independent regret for Log Loss BC)

For any deterministic expert $\pi^\star \in \Pi$, Log-Loss BC ensures with prob at least $1-\delta$,

$$J(\pi^\star) - J(\widehat{\pi}) \leq 8R \cdot \frac{\log(|\Pi|\delta^{-1})}{n}.$$

Special case: Non-stationary policies (or no parameter sharing)

- Ex: $\pi_\theta(x_h) = \arg\max_{a \in \mathcal{A}} \langle \theta_h, \phi(x_h, a) \rangle$ ($\theta \in \mathbb{R}^d$; d parameters)
- Have $\Pi = \Pi_1 \times \cdots \Pi_H$, so $\log|\Pi| = \sum_{h=1}^{H} \log|\Pi_h| = O(H)$.
  - All non-stationary policies in a tabular MDP: $\log|\Pi| = H|\mathcal{X}|\log|\mathcal{A}|$.
  - Linear policies: $\log|\Pi| = \widetilde{O}(Hd)$.
- Log Loss BC gives $J(\pi^\star) - J(\widehat{\pi}) \lesssim \frac{RH}{n}$ at best.
- Dagger can get $J(\pi^\star) - J(\widehat{\pi}) \lesssim \frac{\mu_{\text{rec}} \cdot H}{n}$; not possible w/ offline [Rajaraman et al. '20].

Online IL still helps for non-stationary policies!

But for stationary policies, not possible to beat Log-Loss BC?

To what extent is online interaction beneficial?

Likely still beneficial, but in a problem-dependent, policy class-dependent sense.

To what extent is online interaction beneficial?

Likely still beneficial, but in a problem-dependent, policy class-dependent sense.

Some examples (see paper):

- Representational benefits
  - Learning to correct $\widehat{\pi}$ (Dagger-style) can be simpler representationally than directly learning $\pi^{\star}$.
  - Dagger can get away with smaller policy class: $\log|\Pi'| \ll \log|\Pi|$.

To what extent is online interaction beneficial?

Likely still beneficial, but in a problem-dependent, policy class-dependent sense.

Some examples (see paper):

- Representational benefits
  - Learning to correct $\widehat{\pi}$ (Dagger-style) can be simpler representationally than directly learning $\pi^{\star}$.
  - Dagger can get away with smaller policy class: $\log|\Pi'| \ll \log|\Pi|$.
- Exploration
  - Deliberately visit states that are informative for learning $\pi^{\star}$.
  - Ex: Try to discover state where $\pi^{\star}$ takes different action from all other $\pi \in \Pi$.

To what extent is online interaction beneficial?

Likely still beneficial, but in a problem-dependent, policy class-dependent sense.

Some examples (see paper):

- Representational benefits
  - Learning to correct $\widehat{\pi}$ (Dagger-style) can be simpler representationally than directly learning $\pi^\star$.
  - Dagger can get away with smaller policy class: $\log|\Pi'| \ll \log|\Pi|$.
- Exploration
  - Deliberately visit states that are informative for learning $\pi^\star$.
  - Ex: Try to discover state where $\pi^\star$ takes different action from all other $\pi \in \Pi$.

Further examples:

- Control-theoretic considerations (instability)
- Misspecification (benefits of inverse RL-style algos?)

Opportunity to ~align~ theory and practice!

Conclusion

arXiv: 2407.15007

Summary:

- Log-loss behavior cloning is horizon-independent.
- Instabilities of offline IL / benefits of online IL are probably real, but current theory may be too coarse to capture.
- **Opportunity for better theory + new algorithmic interventions**

Intersection of RL theory + language modeling:

- Error amplification in behavior cloning and autoregression: arXiv:2403.15371
- In-context exploration: arXiv:2403.15371
- Principled algorithms for alignment:
  - XPO: Exploratory preference optimization: arXiv:2405.21046
  - χPO: Rethinking KL-regularization and overoptimization: 2407.13399