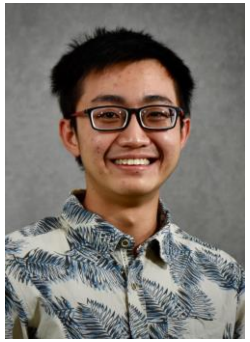
On Efficient Online Imitation Learning via Classification

Chicheng Zhang

University of Arizona

Joint work with Yichen Li (University of Arizona)

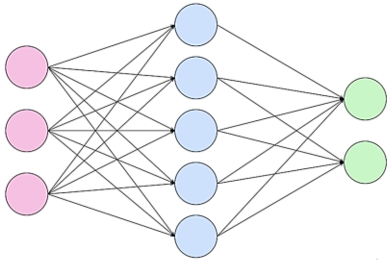


Imitation learning (IL)

Expert feedback

Imitation learner

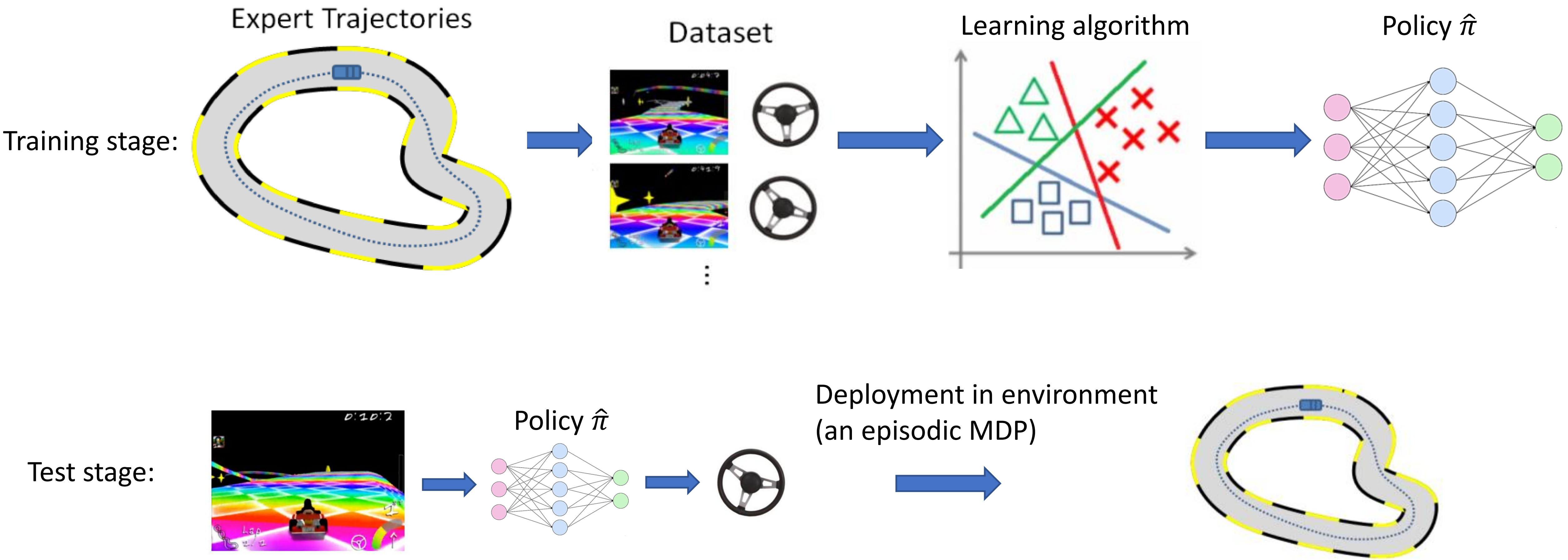
Policy $\hat{\pi}$



- Applications:
  - Autonomous driving
  - Robot control
  - Game playing

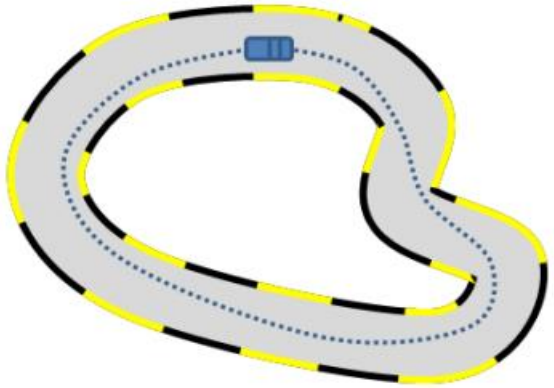
- Sidesteps exploration challenges in reinforcement learning

Example: learning to drive from demonstrations



(Images from Stephane Ross's slides)

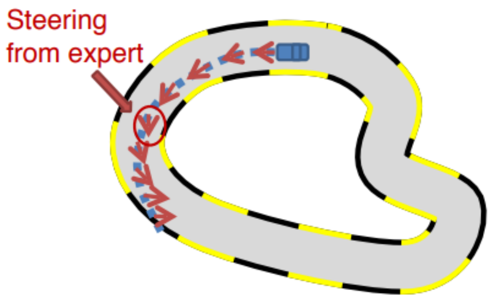
Offline vs. Interactive imitation learning



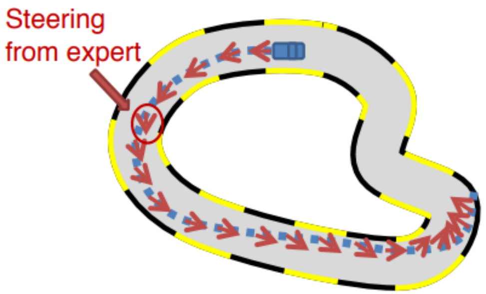
- Offline IL (behavior cloning): learner receives demonstrations ahead of time

- Interactive IL: learner adaptively *queries* expert for demonstrations

Round 1

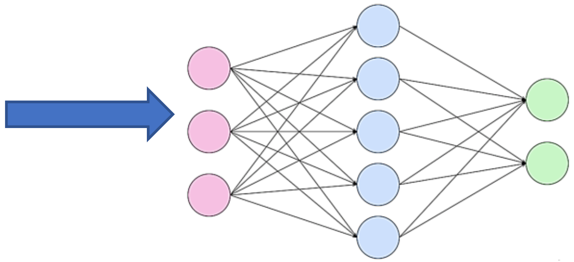


Round 2



.....

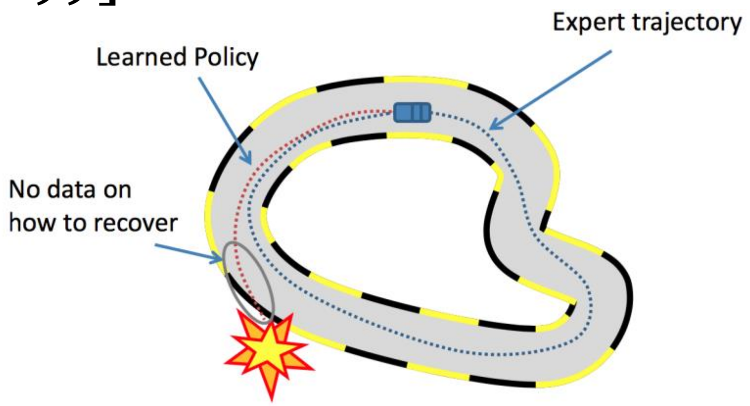
Policy $\hat{\pi}$



- Goal: learn a policy competitive with expert, with low:
  - Sample complexity (#expert demonstrations)
  - Interaction round complexity

Loss choice in imitation learning

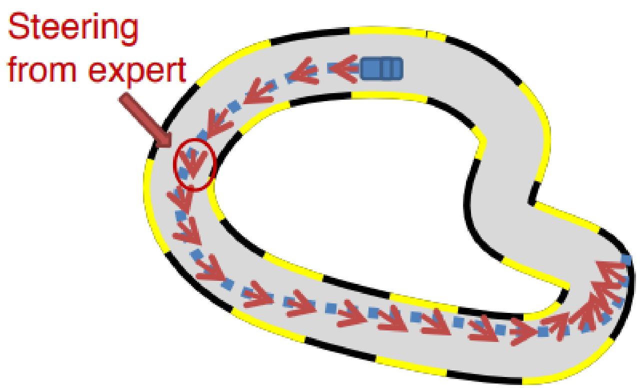
- Assume discrete action space
- Offline IL objective: $L_0(\pi) = \mathrm{E}_{s\sim d_{\pi^E}}[I(\pi(s) \neq \pi^E(s))]$
- d_{π^E}: average state distribution experienced by π^E
- Issue: compounding error (covariate shift)



- A better objective (e.g. Ke et al, 2020):

$$L(\pi) = \mathrm{E}_{s\sim d_{\pi}}[I(\pi(s) \neq \pi^E(s))]$$

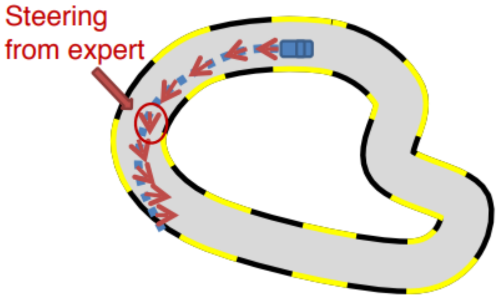
- Can be optimized in *interactive IL*



The DAgger reduction framework for interactive IL (Ross-Gordon-Bagnell'11)

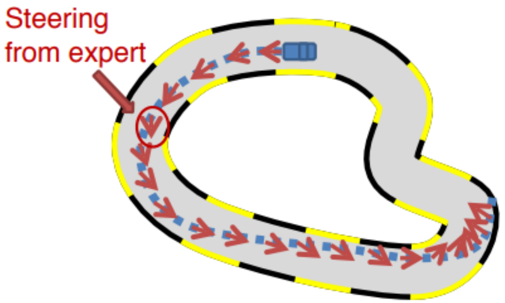
- Goal: optimize imitation loss $L(\pi) = \mathrm{E}_{s\sim d_\pi}[I(\pi(s) \neq \pi^E(s))]$
- DAgger (Data Aggregation) simulates a N-round online learning game:

Learned policy π_n



Approximation of $F_n(\pi) := \mathrm{E}_{s\sim d_{\pi_n}}[I(\pi(s) \neq \pi^E(s))]$

Learned policy π_{n+1}



Approximation of $F_{n+1}(\pi) := \mathrm{E}_{s\sim d_{\pi_{n+1}}}[I(\pi(s) \neq \pi^E(s))]$

- Minimizing $\sum_{n=1}^{N} F_n(\pi_n) \Leftrightarrow$ Minimizing $\sum_{n=1}^{N} L(\pi_n)$

DAgger: guarantees & limitations

- Theorem (simplified): if the sequence of policies $\{\pi_n\}_{n=1}^N$ satisfies that

$$\mathrm{SReg}(N) = \sum_{n=1}^N F_n(\pi_n) - \min_{\pi \in B} \sum_{n=1}^N F_n(\pi) \le R(N), \qquad (*)$$

$\mathrm{E}_{s \sim d_{\pi_n}}[I(\pi(s) \neq \pi^E(s))]$

then outputting $\hat{\pi} \sim \mathrm{Uniform}(\{\pi_n\}_{n=1}^N)$ has $L(\hat{\pi}) \le \mathrm{bias}(B, \pi^E) + \frac{R(N)}{N}$.

Approximability of π^E using B

- How to achieve (*) with small $R(N)$?
  - (Ross-Gordon-Bagnell'11) and subsequent works: Assume some parametrization of π, and optimize for a convex surrogate of $F_n(\pi)$
- Issues:
  - convex surrogate may result in poor approximation of 0-1 error minimizer [Ben-David et al '12]
  - π may not have a parametrization amenable for optimization (e.g. decision trees)

This work: provable regret minimization in classification-based IL

Question: how can we provably achieve

$$\mathrm{SReg}(N) = \sum_{n=1}^{N} F_n(\pi_n) - \min_{\pi \in B} \sum_{n=1}^{N} F_n(\pi) \le R(N), (*)$$

$$\mathrm{E}_{s \sim d_{\pi_n}}[I(\pi(s) \neq \pi^E(s))]$$

?

A classical online classification problem?

We show:

- any (possibly randomized) proper learning algorithm must have $R(N) = \Omega(N)$ in the worst case
- an improper learning framework, *Logger*, that allows the design of IL algorithms with $R(N) = o(N)$
- efficient improper learning algorithms with sample complexity / interaction round complexity guarantees, using *Logger,* using *offline classification oracle*

Result 1: failure of proper learning

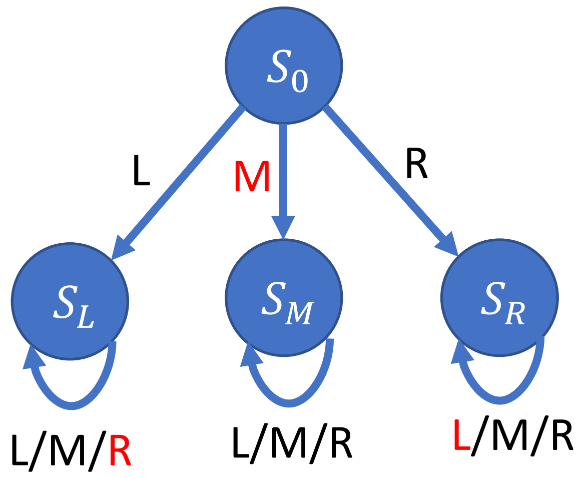
- **Theorem:** there exists an episodic MDP and benchmark policy class B, such that for any $\{\pi_n\}_{n=1}^{N} \subset B$,

$$\mathrm{SReg}(N) = \sum_{n=1}^{N} F_n(\pi_n) - \min_{\pi \in B} \sum_{n=1}^{N} F_n(\pi) = \Omega(N)$$

- Key observation: different from standard online classification, in online IL, F_n may (adversarially) adapt to the choice of π_n
- Similar to (Cover'66)'s impossibility result

Result 1: failure of proper learning

- Episodic MDP M with episode length $H \geq 2$
- Expert π^E
- Benchmark policy class $B = \{h_L \equiv L,\ h_R \equiv R\}$



- For $\{\pi_n\}_{n=1}^{N} \subset B$:
- For every n, $F_n(\pi_n) = 1$ (e.g. $\pi_n = h_L$, trajectory = $(S_0, S_L, \ldots, S_L)$)
- Meanwhile, $\min\left(F_n(h_L), F_n(h_R)\right) \leq \frac{1}{H}$
- These imply that $\mathrm{SReg}(N) \geq \left(1 - \frac{1}{H}\right)\frac{N}{2}$

Result 2: improper learning framework *Logger*

- Define mixed policy classes:

$$\Pi_B = \{\pi_w := \sum_{h \in B} w[h] h(\cdot \mid s) : w \in \Delta^B\}$$

  - Executing π_w: randomly following a policy $\sim$ *w at every step*
  - Implicitly used in (Syed-Schapire'10)
- **Theorem:** algorithmic framework *Logger* (Linear lOss aGGrEgation), when taking online linear optimization (OLO) algorithm A with *deterministic* regret $R(N)$ as input, outputs $\{\pi_n\}_{n=1}^N \subset \Pi_B$ s.t.

$$\mathrm{SReg}(N) \le R(N)$$

- e.g. A = Hedge (Freund-Schapire'97), Follow-the-Regularized-Leader (FTRL), …

Result 2: improper learning framework *Logger*

- Key observation:

$$F_n(\pi_w) = \mathrm{E}_{s\sim d_{\pi_n}} \mathrm{E}_{a\sim\pi_w(\cdot|s)}\left[I\left(a \neq \pi^E(s)\right)\right]$$
$$= \sum_{h\in B} w[h]\, \mathrm{E}_{s\sim d_{\pi_n}}\left[I\left(h(s) \neq \pi^E(s)\right)\right]$$
$$=: \ell_n(w)$$

Algorithm *Logger(A):*

For $n = 1,2, \dots, N$:

  $\pi_n = \pi_{w_n}$, with w_n being the output of A

  Update A with (unbiased estimates) of ℓ_n

$$\Rightarrow R(N)$$
$$\geq \sum_{n=1}^{N} \ell_n(w_n) - \min_{w\in\Delta^B} \sum_{n=1}^{N} \ell_n(w)$$
$$= \sum_{n=1}^{N} F_n(\pi_n) - \min_{\pi\in B} \sum_{n=1}^{N} F_n(\pi)$$
$$= \mathrm{SReg}(N)$$

Result 3: oracle-efficient regret minimization for IL

- Assume offline classification oracle O for policy class B:

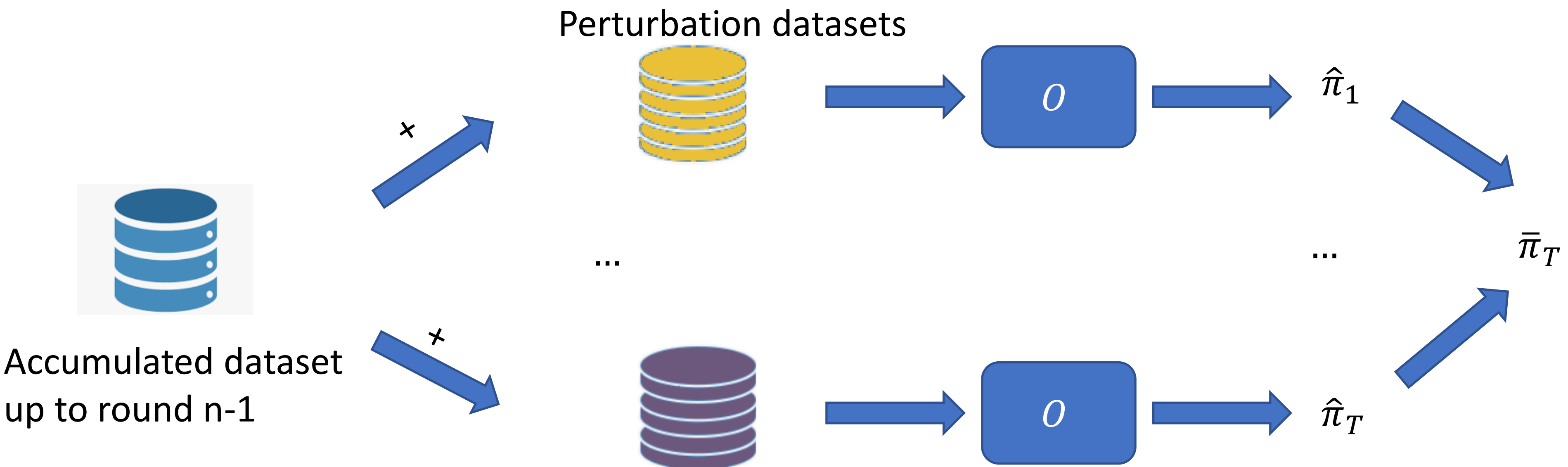
$$D = \langle (x_1, y_1), \ldots, (x_n, y_n) \rangle \longrightarrow \boxed{O} \longrightarrow \mathrm{argmin}_{h \in B} \mathrm{E}_D[I(h(x) \neq y)]$$

- Useful computational abstraction for designing efficient online learning algorithms (e.g. Langford-Zhang’07, Syrgkanis et al’16, Rakhlin-Sridharan’16)

- Can we use it to design efficient regret minimization algorithms for IL?

Result 3: oracle-efficient regret minimization for IL

- Challenge: existing adversarial oracle-efficient online learning algorithms use *proper learning* (e.g. Syrgkanis et al '16)
  - unavoidably suffers linear regret in IL (Result 1)
- Workaround: utilize an equivalence between an in-expectation version of Follow-the-Perturbed-Leader (FTPL) and Follow-the-Regularized-Leader (Abernethy et al '14) ⇒ our *Mixed-FTPL* algorithm



Result 3: oracle-efficient regret minimization for IL

- **Theorem:** Assuming B satisfies a small-separator condition (Syrgkanis et al ’16). *Logger,* when taking $A =$ Mixed-FTPL,
  - outputs $\{\pi_n\}_{n=1}^{N} \subset \Pi_B$ such that $\mathrm{SReg}(N) \leq O\left(\sqrt{N}\right)$;
  - calls the offline classification oracle for $O(N^2)$ times

- See full paper for detailed sample complexity & interaction round complexity analysis, and comparison with behavior cloning

Conclusions

- We established fundamental results of (efficient) regret minimization in classification-based online imitation learning, which puts imitation learning into firmer theoretical foundations

Future work

- Investigate sample complexity and interaction round complexity lower bounds for online imitation learning
- Relax the (small separator) assumption for designing efficient algorithms
- Empirical evaluation of the algorithms

Thank you!

Results not in this talk

- We also design an algorithm with improved interaction round complexity by utilizing the predictability of the losses (e.g. Cheng et al, 2018, 2020)

- We also show computational hardness of *dynamic regret* minimization in the *Logger* framework

$$\mathrm{DReg}(N) = \sum_{n=1}^{N} \left(F_n(\pi_n) - \min_{\pi \in B} F_n(\pi) \right)$$

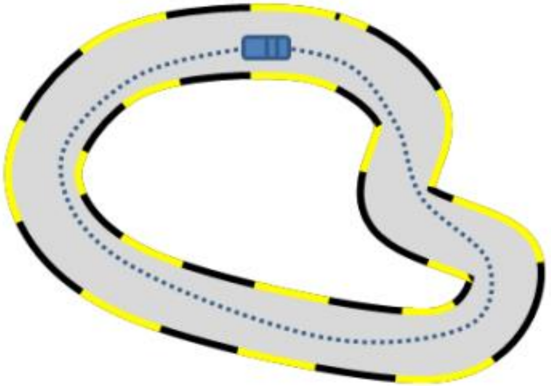
Backup

Notations

- Markov decision process M
- State space S
- Action space A
- Expert policy π^E

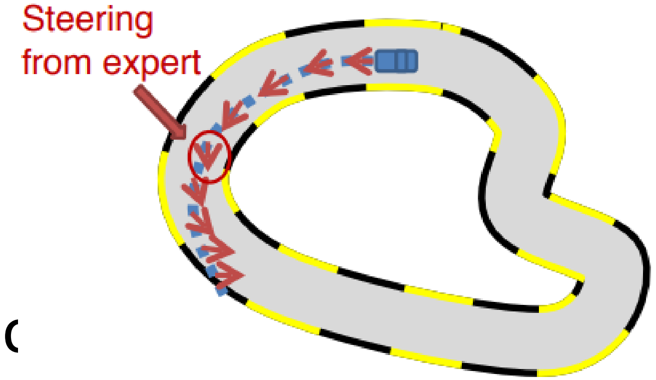
- Occupancy distribution d_π

Offline vs. Interactive imitation learning

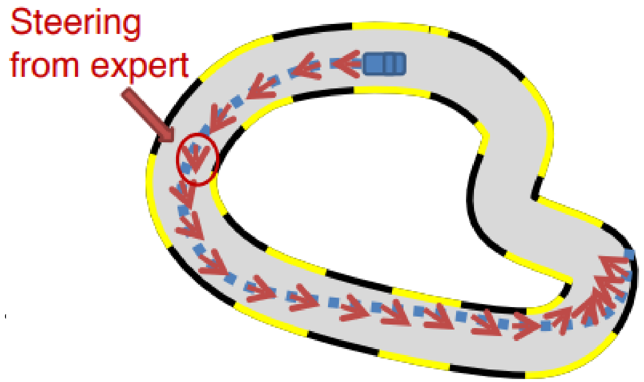


- Offline IL (behavior cloning): learner receive expert demons
  ahead of time
- Interactive IL: learner interactively *queries* experts for demc

Round 1

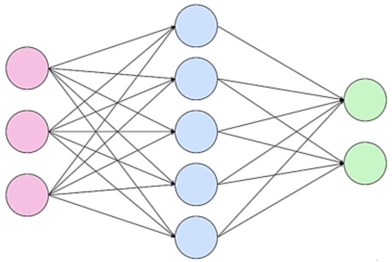


Round 2



.....

Policy π



- Gc eti mall:
  - Sample complexity: # expert demonstrations
  - Interaction round complexity : # interaction rounds

The DAgger reduction framework for interactive IL [Ross et al, 2011]

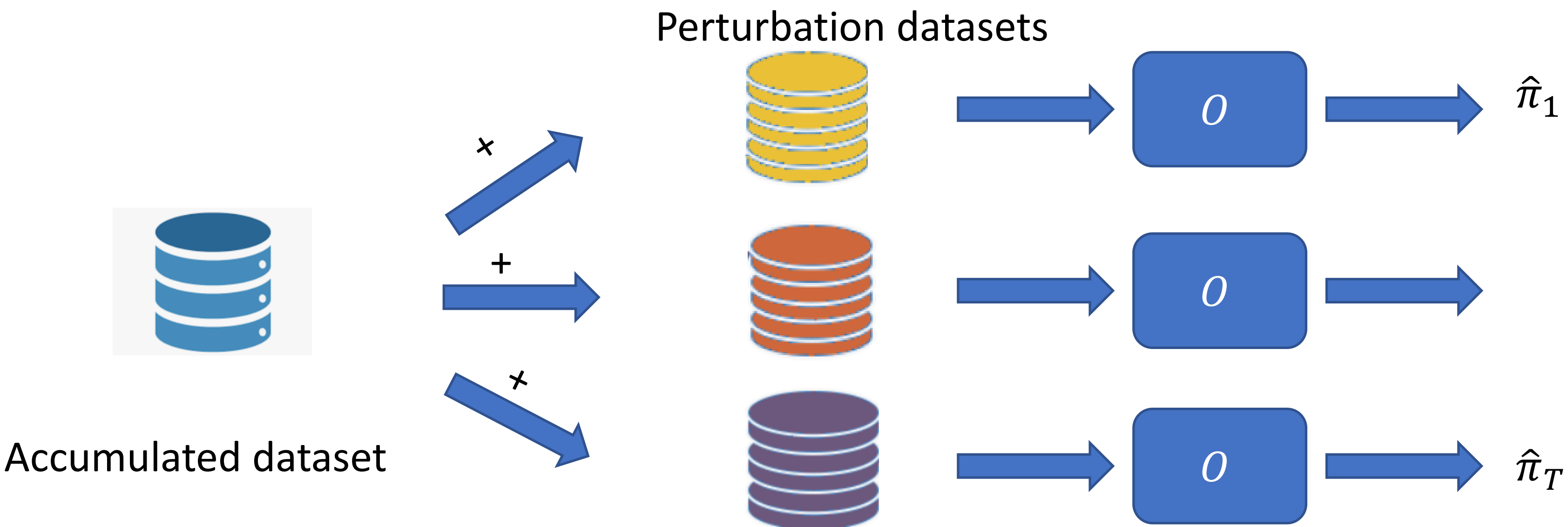
- Goal: find policy π that optimizes imitation loss $L(\pi) = \mathrm{E}_{s \sim d_\pi}[I(\pi(s) \neq \pi^E(s))]$
- DAgger (Data Aggregation) simulates a N-round online learning game:

- For n=1,..,N:
  - Learned policy π_n
  - $F_n(\pi) := \mathrm{E}_{s \sim d_{\pi_n}}[I(\pi(s) \neq \pi^E(s))]$ = loss at round n
  - Observe approximation of $F_n(\pi)$ by executing π_n and query expert π^E

- Minimizing $\sum_{n=1}^N F_n(\pi_n) \Leftrightarrow$ Minimizing $\sum_{n=1}^N L(\pi_n)$

Oracle-efficient regret minimization algorithms for IL

- Challenge: existing adversarial oracle-efficient online learning algorithms use *proper learning* (e.g. Syrgkanis et al ’16)
  - unavoidably suffers linear regret in IL setting
- Workaround: utilize an equivalence between an in-expectation version of Follow-the-Perturbed-Leader and Follow-the-Regularized-Leader (Abernethy et al ’14)



Result 3: oracle-efficient regret minimization for IL

- Assume cost-sensitive classification (CSC) oracle O for policy class B:

$$D = \langle (x_1, c_1), \dots, (x_n, c_n) \rangle \longrightarrow \boxed{O} \longrightarrow \mathrm{argmin}_{h \in B} \mathrm{E}_D[c(h(x))]$$

- Useful computational abstraction for designing efficient online learning algorithms (e.g. Langford-Zhang'07, Syrgkanis et al'16, Rakhlin-Sridharan'16)

- Can we use it to design efficient regret minimization algorithms for IL?

Result 3: oracle-efficient regret minimization for IL

- **Theorem:** Assuming B satisfies a small-separator condition (Syrgkanis et al '16). *Logger,* when taking $A =$ Mixed-FTPL,
  - outputs $\{\pi_n\}_{n=1}^{N} \subset \Pi_B$ such that $\mathrm{SReg}(N) \le O\left(\sqrt{N}\right)$;
  - calls the offline classification oracle for $O(N^2)$ times

- See full paper for detailed sample complexity & interaction round complexity analysis, and comparison with behavior cloning
- We also design an algorithm with improved interaction round complexity by utilizing the predictability of the losses (e.g. Cheng et al, 2018, 2020)

Conclusions and future work

- We established fundamental statistical limits of regret minimization in classification-based online imitation learning, which puts imitation learning into firmer theoretical foundations

- (Not covered in this talk) We also show computational hardness of *dynamic regret* minimization in the *Logger* framework

$$\mathrm{DReg}(N) = \sum_{n=1}^{N} \left(F_n(\pi_n) - \min_{\pi \in B} F_n(\pi) \right)$$