CSC2621 Topics in Robotics
Reinforcement Learning in Robotics

Week 2: Behavioral Cloning from Observation

Tingwu Wang, Dylan Turpin, Animesh Garg

Agenda

- Background
  - Problem Setting
  - Behavior Cloning / Dagger
  - Generative Adversarial Imitation Learning
- Motivation
- Behavior Cloning from Observation
  - Algorithm
  - Results
- Discussion

Problem Setting

- Imitation learning
  - Other names in different contexts:
    - Learning from demonstrations / Apprenticeship learning
  - **Input**:
    - Expert's perfect trajectories {(s_t, a_t)}
  - **Output**:
    - A policy network p(a_t | s_t)
  - **Goal**:
    - Can our agent be taught to **reproduce the skills** to solve a given task?
    - Why not reward / Why not use human designed rules?
      - Hard / not safe / not generalized

Behavior Cloning / Dagger

- Treat it as a regression problem
  - A **policy** network
    - Input: s_i
    - Output: a = p(a_i | s_i)
  - Find the **policy** parameterized by phi that fits the expert data

$$\phi^* = \arg\max_{\phi} \prod_{i=1}^{N} \pi_\phi(a_i|s_i)$$

  - How is the “dataset” {(a_i, s_i)} generated?
    - Two different problem settings

Behavior Cloning / Dagger

- Behavior cloning (BC)
  - Setting A
  - Ask an expert to generate the **expert dataset**.
  - The agent direct regresses on the **expert dataset**.
  - **Train on expert's state distribution**.
- Dataset Aggregation algorithm (Dagger)
  - Setting B
  - The learner samples the states {s_i}.
  - Then ask the expert to produce the correct actions {a_i}.
  - Repeat
  - Dagger: **Train on learner's state distribution**. It has a more powerful / kinder expert.

Generative Adversarial Imitation Learning

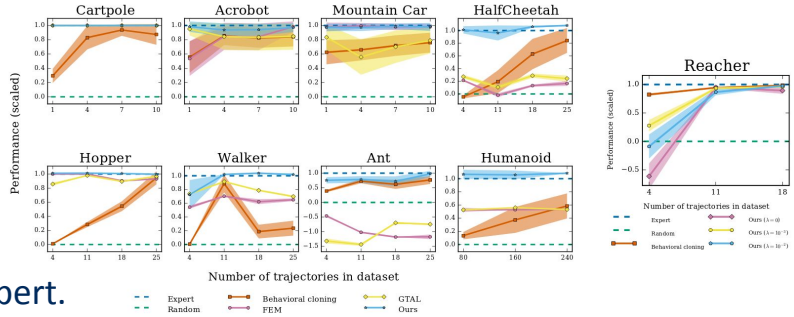
- Goes back to Setting A
  - Behavior cloning is good enough when:
    - Large amounts of data
    - Lower dimensional environments
  - Compounding error

- Inverse reinforcement learning (IRL)
  - Learns a cost / reward function that prioritizes entire trajectories.
  - Then learns the policy as a RL problem.
  - Mathematically proved that it introduces smaller compounding error.

Generative Adversarial Imitation Learning

- Generative Adversarial Imitation Learning (GAIL)
  - Learn the reward function using GAN (Generative Adversarial Network)
  - Discriminator assigns reward of 1.0 to expert's (s_t, a_t)
  - Discriminator assigns reward of 0.0 to learner's (s_t, a_t)

- Process
  - Learner generate new trajectories {(s_t, a_t)}.
  - Discriminator trains on trajectories of the learner and expert.
  - Discriminator assign rewards to learner's trajectories {(s_t, a_t)}.
  - Learner updates policy network.

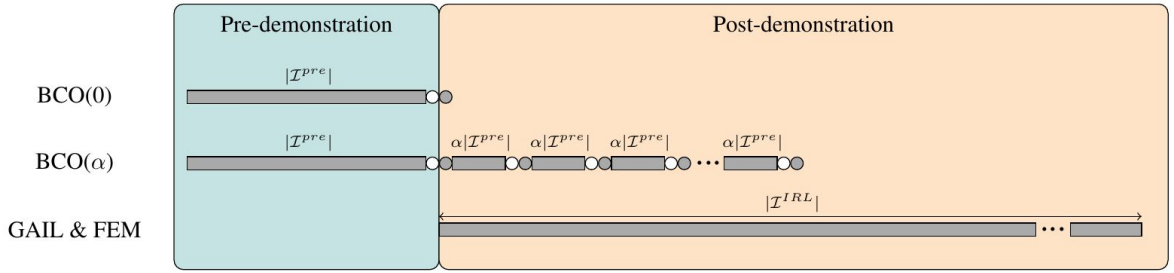


Motivation

- BC / GAIL / Dagger
  - They all requires the access of the actions, which is not the case when:
    - Imitation learning from motion captured data
    - Virtual Reality Teleoperation
    - Noisy data / model mismatch / retargeting
- Instead of expert's perfect trajectories {(s_t, a_t)}
  - **Input**:
    - expert's perfect trajectories without actions {(s_t)}

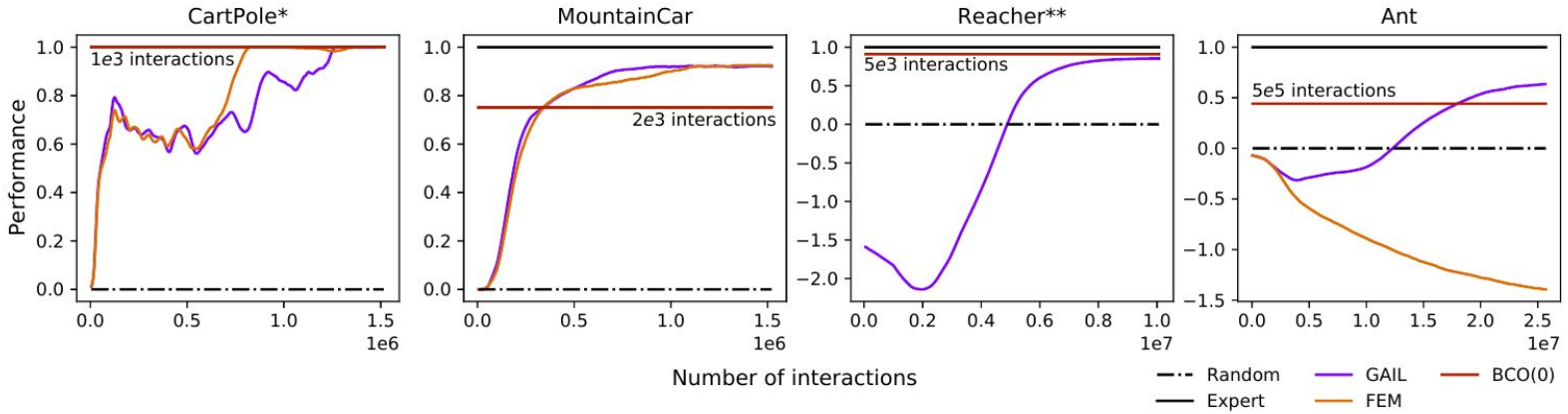
Behavior Cloning from Observation

- The idea of behavior cloning from observation (BCO):
  - If the actions won't come from the expert, then the learner must come to infer the actions
- Inverse dynamics
  - Forward dynamics:
    - s_t ← f(s_{t-1}, a_{t-1})
  - Inverse dynamics:
    - a_{t-1}t ← f(s_{t-1}, s_t)
- Essentially
  - Inverse dynamics + BC
- BCO (alpha) variant



Results

- Comparison on 4 environments



Discussion

- Pros:
  - Proposed to solve a problem of a new setting.
- Cons:
  - Could have a more comprehensive result sections
    - Right figure from [1]
    - Below figure from [2]

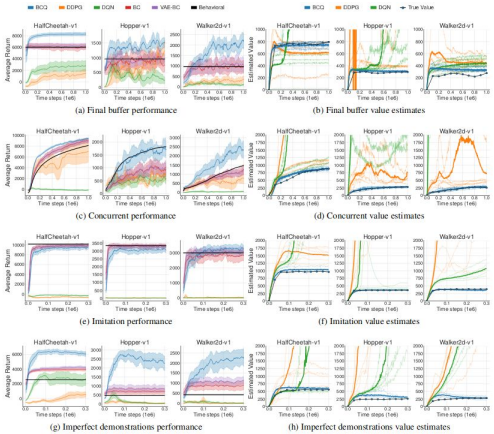


Table 1: Final performance for 18 environments of the bench-marked algorithms. All the algorithms are run for 200k time-steps. Blue refers to the best methods using ground truth dynamics, red to the best MBRL algorithms, and green to the best MFRL algorithms. The results show the mean and standard deviation averaged over 4 random seeds and a window size of 5000 times-steps.

	Pendulum	InvertedPendulum	Acrobot	CartPole	Mountain Car	Reacher
Random	-202.6 ± 249.3	-205.1 ± 13.6	-374.5 ± 17.1	38.4 ± 32.5	-105.1 ± 1.8	-45.7 ± 4.8
iLQG	**160.8 ± 29.8**	**-0.0 ± 0.0**	-195.5 ± 28.7	**199.3 ± 0.6**	**-55.9 ± 8.3**	**-6.0 ± 2.6**
GT-CEM	**170.5 ± 35.2**	**-0.2 ± 0.1**	**13.9 ± 40.5**	**199.9 ± 0.1**	**-58.0 ± 2.9**	**-3.6 ± 1.2**
GT-RS	**171.5 ± 31.8**	**-0.0 ± 0.0**	**2.5 ± 39.4**	**200.0 ± 0.0**	-68.5 ± 2.2	-25.7 ± 3.5
RS	164.4 ± 9.1	**-0.0 ± 0.0+**	**-4.9 ± 5.4**	**200.0 ± 0.0+**	-71.3 ± 0.5	-27.1 ± 0.6
MB-MF	157.5 ± 13.2	-182.3 ± 24.4	-92.5 ± 15.8	**199.7 ± 1.2**	4.2 ± 18.5	-15.1 ± 1.7
PETS-CEM	167.4 ± 53.0	-20.5 ± 28.9	**12.5 ± 29.0+**	**199.5 ± 3.0**	-57.9 ± 3.6	-12.3 ± 5.2
PETS-RS	167.9 ± 35.8	-12.1 ± 25.1	-71.5 ± 44.6	195.0 ± 28.0	-78.5 ± 2.1	-40.1 ± 6.9
ME-TRPO	**177.3 ± 1.9+**	-126.2 ± 86.6	-68.1 ± 6.7	160.1 ± 69.1	-42.5 ± 26.6	-13.4 ± 0.2
GPS	162.7 ± 7.6	-74.6 ± 97.8	-193.3 ± 11.7	14.4 ± 18.6	-10.6 ± 32.1	-19.8 ± 0.9
PILCO	-132.6 ± 410.1	-194.5 ± 0.8	-394.4 ± 1.4	-1.9 ± 155.9	-59.0 ± 4.6	-13.2 ± 5.9
SVG	141.4 ± 62.4	-183.1 ± 9.0	-79.7 ± 6.6	82.1 ± 31.9	-27.6 ± 32.6	-11.0 ± 1.0
MB-MPO	171.2 ± 26.9	**-0.0 ± 0.0+**	-87.8 ± 12.9	**199.3 ± 2.3**	-30.6 ± 34.8	**-5.6 ± 0.8**
SLBO	173.5 ± 2.5	-240.4 ± 7.2	-75.6 ± 8.8	78.0 ± 166.6	**44.1 ± 6.8**	**-4.1 ± 0.1+**
PPO	**163.4 ± 8.0**	-40.8 ± 21.0	-95.3 ± 8.9	86.5 ± 7.8	21.7 ± 13.1	-17.2 ± 0.9
TRPO	**166.7 ± 7.3**	-27.6 ± 15.8	-147.5 ± 12.3	47.3 ± 15.7	-37.2 ± 16.4	-10.1 ± 0.6
TD3	**161.4 ± 14.4**	-224.5 ± 0.4	-64.3 ± 6.9	196.0 ± 3.1	-60.0 ± 1.2	-14.0 ± 0.9
SAC	**168.2 ± 9.5**	**-0.2 ± 0.1**	**-52.9 ± 2.0**	**199.4 ± 0.4**	**52.6 ± 0.6+**	**-6.4 ± 0.5**

	HalfCheetah	Swimmer-v0	Swimmer	Ant	Ant-ET	Walker2D
Random	-288.3 ± 65.8	1.2 ± 11.2	-9.5 ± 11.6	473.8 ± 40.8	124.6 ± 145.0	-2456.9 ± 345.3
iLQG	**2142.6 ± 137.7**	47.8 ± 2.4	306.7 ± 0.8	9739.8 ± 745.0	1506.2 ± 459.4	-1186.2 ± 126.3
GT-CEM	**14777.2 ± 13964.2**	**111.0 ± 4.6**	**335.9 ± 1.1**	**12115.3 ± 209.7**	226.0 ± 178.6	**7719.7 ± 486.7**
GT-RS	815.7 ± 38.5	35.8 ± 3.0	42.2 ± 5.3	2709.1 ± 631.1	**2519.0 ± 469.8**	-1641.4 ± 137.6
RS	421.0 ± 55.2	31.1 ± 2.0	92.8 ± 8.1	535.5 ± 37.0	**239.9 ± 81.7**	-2060.3 ± 228.0
MB-MF	126.9 ± 72.7	51.8 ± 30.9	284.9 ± 25.1	134.2 ± 50.4	85.7 ± 27.7	-2218.1 ± 437.7
PETS-CEM	**2795.3 ± 879.9**	22.1 ± 25.2	306.3 ± 37.3	1165.5 ± 226.9	81.6 ± 145.8	**260.2 ± 536.9**
PETS-RS	966.9 ± 471.6	42.1 ± 20.2	170.1 ± 8.1	**1852.1 ± 141.0+**	130.0 ± 148.1	**312.5 ± 493.4+**
ME-TRPO	**2283.7 ± 900.4**	30.1 ± 9.7	**336.3 ± 15.8+**	282.2 ± 18.0	42.6 ± 21.1	-1609.3 ± 657.5
GPS	52.3 ± 41.7	14.5 ± 5.6	-35.3 ± 8.4	445.5 ± 212.9	**275.4 ± 309.1**	-1730.8 ± 441.7
PILCO	-41.9 ± 267.0	-13.8 ± 16.1	-18.7 ± 10.3	770.7 ± 153.0	N. A.	-2693.8 ± 484.4
SVG	336.6 ± 387.6	**77.2 ± 99.0**	75.2 ± 85.3	377.9 ± 33.6	185.0 ± 141.6	-1430.9 ± 230.1
MB-MPO	**3639.0 ± 1185.8**	**85.0 ± 98.9+**	268.5 ± 125.4	705.8 ± 147.2	30.3 ± 22.3	-1545.9 ± 216.5
SLBO	1097.7 ± 166.4	41.6 ± 18.4	125.2 ± 93.2	718.1 ± 123.3	**200.0 ± 40.1**	-1277.7 ± 427.5
PPO	17.2 ± 84.4	**38.0 ± 1.5**	306.8 ± 4.2	321.0 ± 51.2	80.1 ± 17.3	-1893.6 ± 234.1
TRPO	-12.0 ± 85.5	**37.9 ± 2.0**	215.7 ± 10.4	323.3 ± 24.9	116.8 ± 47.3	-2286.3 ± 373.3
TD3	3614.3 ± 82.1	**40.4 ± 8.3**	**331.1 ± 0.9**	**956.1 ± 66.9**	259.7 ± 1.0	**-73.8 ± 769.0**
SAC	**4000.7 ± 202.1+**	**41.2 ± 4.6**	309.8 ± 4.2	506.7 ± 165.2	**2012.7 ± 571.3+**	-415.9 ± 588.1

	Walker2D-ET	Hopper	Hopper-ET	SlimHumanoid	SlimHumanoid-ET	Humanoid-ET
Random	-2.8 ± 4.3	-2572.7 ± 631.3	12.7 ± 7.8	-1172.9 ± 757.0	41.8 ± 47.3	50.5 ± 57.1
iLQG	**229.0 ± 74.7**	1157.6 ± 224.7	83.4 ± 21.7	13225.2 ± 1344.9	520.0 ± 240.9	255.0 ± 94.6
GT-CEM	**254.8 ± 233.4**	**3232.3 ± 192.3**	**256.8 ± 16.3**	**45979.8 ± 1654.9**	**1242.7 ± 676.0**	**1236.2 ± 668.0**
GT-RS	**207.9 ± 27.2**	-2467.2 ± 55.4	209.5 ± 46.8	8074.4 ± 441.1	361.5 ± 103.8	312.9 ± 167.8
RS	201.1 ± 10.5	-2491.5 ± 35.1	247.1 ± 6.1	-99.2 ± 388.5	332.8 ± 13.4	295.5 ± 10.9
MB-MF	**350.0 ± 107.6**	-1047.4 ± 1098.7	**926.9 ± 154.1**	-1320.2 ± 735.3	809.7 ± 57.5	776.8 ± 62.9
PETS-CEM	-2.5 ± 6.8	**1125.0 ± 679.6**	129.3 ± 36.0	1472.4 ± 738.3	355.1 ± 157.1	110.8 ± 91.0
PETS-RS	-0.8 ± 3.2	-1469.8 ± 224.1	205.8 ± 36.5	**2055.1 ± 771.5+**	320.7 ± 182.2	106.9 ± 102.6
ME-TRPO	-9.5 ± 4.6	**1272.5 ± 500.9**	4.9 ± 4.0	-154.9 ± 534.3	76.1 ± 8.8	72.9 ± 8.9
GPS	-2400.6 ± 610.8	-768.5 ± 200.9	-2303.9 ± 338.1	-592.6 ± 214.1	N. A.	N. A.
PILCO	N. A.	-1729.9 ± 1611.1	N. A.	N. A.	N. A.	N. A.
SVG	**252.4 ± 48.4**	-877.9 ± 427.9	435.2 ± 163.8	1096.8 ± 791.0	**1084.3 ± 77.0+**	811.8 ± 241.5
MB-MPO	-10.3 ± 1.4	333.2 ± 1189.7	8.3 ± 3.6	674.4 ± 982.2	115.5 ± 31.9	73.1 ± 23.1
SLBO	207.8 ± 108.7	-741.7 ± 734.1	**805.7 ± 142.4**	-588.9 ± 332.1	776.1 ± 252.5	**1377.0 ± 150.4**
PPO	306.1 ± 17.2	-103.8 ± 1028.0	758.0 ± 62.0	-1466.7 ± 278.5	454.3 ± 36.7	451.4 ± 39.1
TRPO	229.5 ± 27.1	-2100.1 ± 640.6	237.4 ± 33.5	-1140.9 ± 241.8	281.3 ± 10.9	289.8 ± 5.2
TD3	**3299.7 ± 1951.5+**	**2245.3 ± 232.4+**	1057.1 ± 29.5	**1319.1 ± 1246.1**	**1070.0 ± 168.3**	147.7 ± 0.7
SAC	**2216.4 ± 678.7**	726.4 ± 675.5	**1815.5 ± 655.1+**	**1328.4 ± 468.2**	843.6 ± 313.1	**1794.4 ± 458.3+**

[1] Wang, Tingwu et al. “Benchmarking Model-Based Reinforcement Learning.” *ArXiv* abs/1907.02057 (2019)

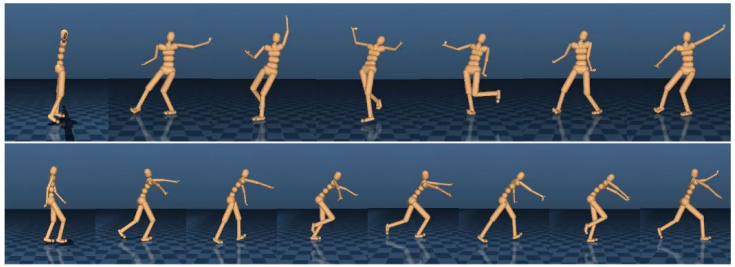
[2] Fujimoto, Scott, et al. "Off-policy deep reinforcement learning without exploration." *arXiv preprint arXiv:1812.02900* (2018).

Discussion

- Cons:
  - Some of the claims are not supported by empirical results nor theorems.
  - Missing baselines and perhaps limited novelty [3].

3.1 Validation of imitation of without actions, using partial observations

Imitation without ground-truth actions: To show that it is sufficient for the discriminator to condition on state information (including velocity) without accompanying actions, we examine imitation learning for a 2D planar walker (see Figure 3, left panel). The walker task consists of 10s episodes, terminating early if the walker torso falls below a threshold. The demonstration policy is continuously rewarded proportionally to the absolute difference between its horizontal velocity and a target speed (5m/s), minus a small control cost. See video of the demonstration, and see supplemental



[3] Merel, J., Tassa, Y., Srinivasan, S., Lemmon, J., Wang, Z., Wayne, G., & Heess, N. (2017). Learning human behaviors from motion capture by adversarial imitation. *arXiv preprint arXiv:1707.02201*.