OPAL: Offline Preference-Based Apprenticeship Learning

Daniel Shin, Daniel S. Brown

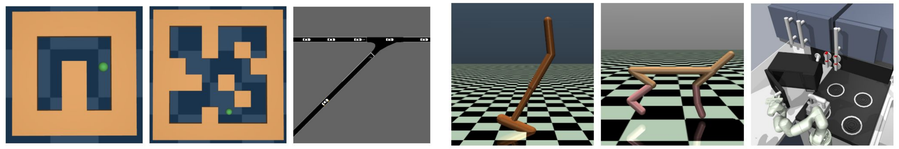
InterACT Laboratory

Berkeley UNIVERSITY OF CALIFORNIA

Motivation

To enable safe and efficient apprenticeship learning, we propose to leverage a large offline database of transitions to perform offline preference learning followed by offline RL. This enables robots to learn complex skills from humans, while avoiding expensive and possibly dangerous rollouts in the environment.

Are Offline RL Benchmarks Well Suited for Studying Offline Reward Learning?

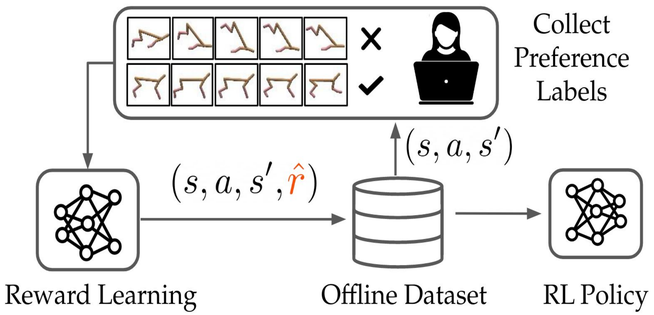


(a) U-Maze (b) Medium Maze (c) Flow Merge (d) Hopper (e) Halfcheetah (f) Franka Kitchen

We evaluate offline RL performance on D4RL benchmark tasks. We replace all rewards with the average over the dataset (Avg) or with zeros (Zero). Performance w.r.t. ground truth rewards significantly degrades on only a subset of the tasks (bolded). Un-bolded tasks contain little variation in the data and can be solved via direct imitation without having an informative reward function.

Task	AWR		BCQ		BEAR		CQL	
	Avg	Zero	Avg	Zero	Avg	Zero	Avg	Zero
flow-ring-random-v1	68.3	67.5	68.9	54.7	102.2	98.5	64.7	79.2
flow-merge-random-v1	-14.7	-14.5	6.1	2	-25.3	-22.4	-8.7	-30.6
maze2d-umaze	12.8	10.0	-23.7	-14.0	-8.6	18.6	-6.4	-20.6
maze2d-medium	-15.9	-11.5	-23.9	-18.2	-16.7	-24.8	-11.7	-28.1
halfcheetah-random	6.9	6.9	8.3	8.3	8.3	8.3	0.1	3.7
halfcheetah-medium-replay	71.7	69.9	71.4	77.6	58.7	61.5	24.3	-3.8
halfcheetah-medium	73.2	71.7	89.1	89.6	89.9	87.0	92.9	91.3
halfcheetah-medium-expert	6.4	6.1	102.2	101.1	47.5	44.6	41.6	51.7
halfcheetah-expert	6.9	6.7	99.5	94.0	47.2	85.1	2.9	33.0
hopper-random	11.7	1.1	17.6	19.0	20.7	7.1	51.9	-0.8
hopper-medium-replay	45.0	38.0	43.2	36.1	26.0	52.4	79.2	22.1
hopper-medium	69.4	55.9	102.3	93.6	36.1	112.9	129.3	123.0
hopper-medium-expert	26.5	25.3	132.2	60.8	23.3	23.8	49.7	53.0
hopper-expert	28.6	19.6	139.0	134.2	72.5	95.4	42.3	99.1
kitchen-complete	10.8	8.4	88.0	12.0	0.0	37.7	80.0	66.7
kitchen-mixed	20.0	32.9	85.7	28.6	117.7	128.9	63.5	111.1
kitchen-partial	138.2	94.1	78.4	39.2	367.6	379.4	196.1	130.7

OPAL Framework



The offline dataset is used to generate preference queries and also used for offline policy optimization using the rewards learned from preferences.

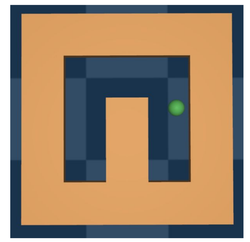
Active Learning Performance

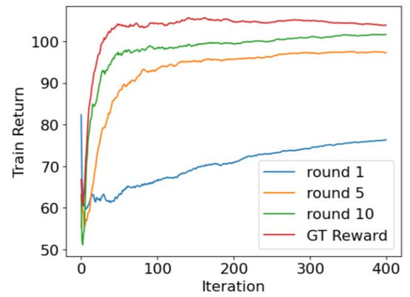
AWR (Peng et al., 2019)	Query Aquisition Method				
	T-REX	EnsemDis	EnsemInfo	DropDis	DropInfo
maze2d-umaze	78.2	**93.5**	89.2	88.0	61.3
maze2d-medium	71.6	**86.4**	71.0	52.6	44.4
hopper	69.0	77.5	72.8	87.6	**90.6**
halfcheetah	96.1	**113.7**	100.6	84.4	91.7
flow-merge-random	**110.1**	89.0	92.1	86.2	84.2
kitchen-complete	79.6	105.0	**158.8**	48.5	65.4

We investigate two query acquisition functions: *disagreement* and *information gain*, and two methods for obtaining uncertainty estimates: *ensembles* and *Bayesian dropout*. We evaluated all four combinations including Ensemble Disagreement (**EnsemDis**), Ensemble Information Gain (**EnsemInfo**), Dropout Disagreement (**DropDis**), Dropout Information Gain (**DropInfo**) in addition to a fixed set of randomly chosen queries (**T-REX**). We find that Ensemble Disagreement performs well overall.

OPAL Only Needs Small Numbers of Active Queries

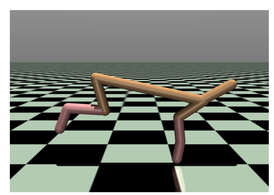
maze2d-umaze

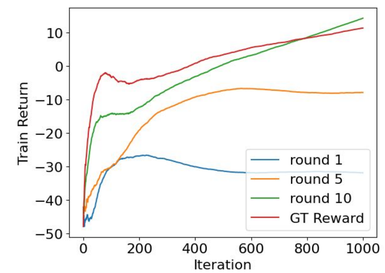




We only require **15 bits** of information to get performance comparable with using **1 million** samples of the ground truth rewards!

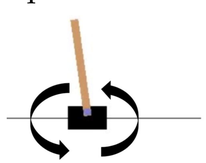
halfcheetah-random



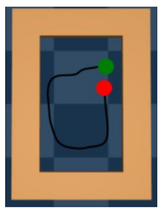


Customized Learned Behaviors

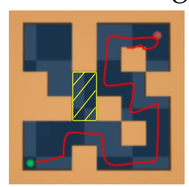
Cartpole Windmill



Orbit



Constrained Navigation



Check out our website here for more results and videos