Rethinking Goal-conditioned Supervised Learning and Its Connection to Offline RL

Rui Yang[1], Yiming Lu[1], Wenzhe Li[1], Hao Sun[2], Meng Fang[3], Yali Du[4], Xiu Li[1], Lei Han[5], Chongjie Zhang[1]

[1]Tsinghua University, [2]University of Cambridge, [3]Eindhoven University of Technology,

[4]King's College London, [5]Tencent Robotics X

Background

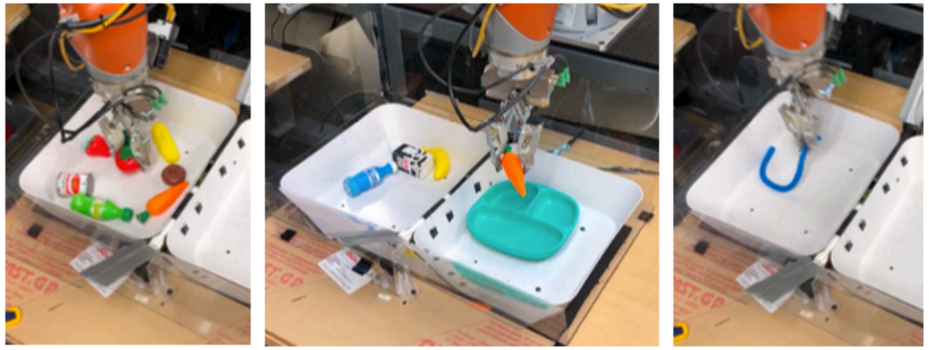
- Goal-conditioned RL (GCRL) encourages agents to reach multiple goals and learn general policies

Chebotar Y et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. ICML 2021

Background

- Goal-conditioned RL (GCRL) encourages agents to reach multiple goals and learn general policies

- Current GCRL algorithms require intense online interactions (dangerous & expensive)

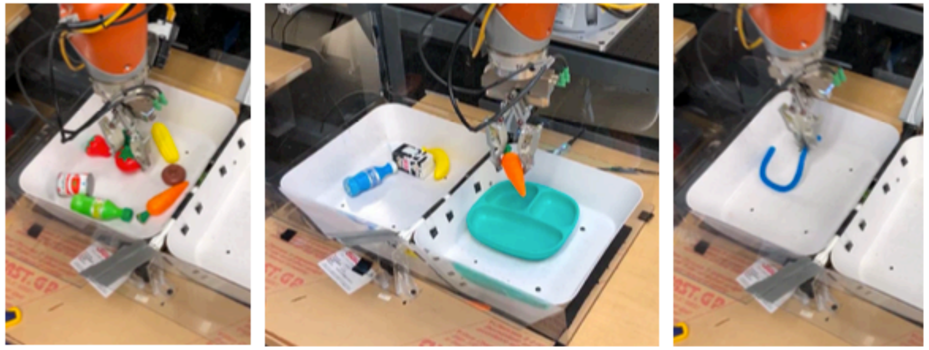


Chebotar Y et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. ICML 2021

Background

- Goal-conditioned RL (GCRL) encourages agents to reach multiple goals and learn general policies

- Current GCRL algorithms require intense online interactions (dangerous & expensive)



- Solution: learning goal-conditioned policies from offline datasets

Chebotar Y et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. ICML 2021

Formulation

- Offline Goal-conditioned RL (GCRL)
- Goal-augmented MDP: $(S, A, \boldsymbol{G}, P, r, \gamma)$
- State-to-goal mapping $\boldsymbol{\phi}$: $\boldsymbol{S} \rightarrow \boldsymbol{G}$,

Formulation

- Offline Goal-conditioned RL (GCRL)

- Goal-augmented MDP: $(S, A, \boldsymbol{G}, P, r, \gamma)$

- State-to-goal mapping $\boldsymbol{\phi}$: $\boldsymbol{S} \rightarrow \boldsymbol{G}$,

- Reward function: $$r(s_t, a_t, g) = \begin{cases} 1, & \|\phi(s_t) - g\|_2^2 \leq \epsilon \\ 0, & otherwise \end{cases}$$

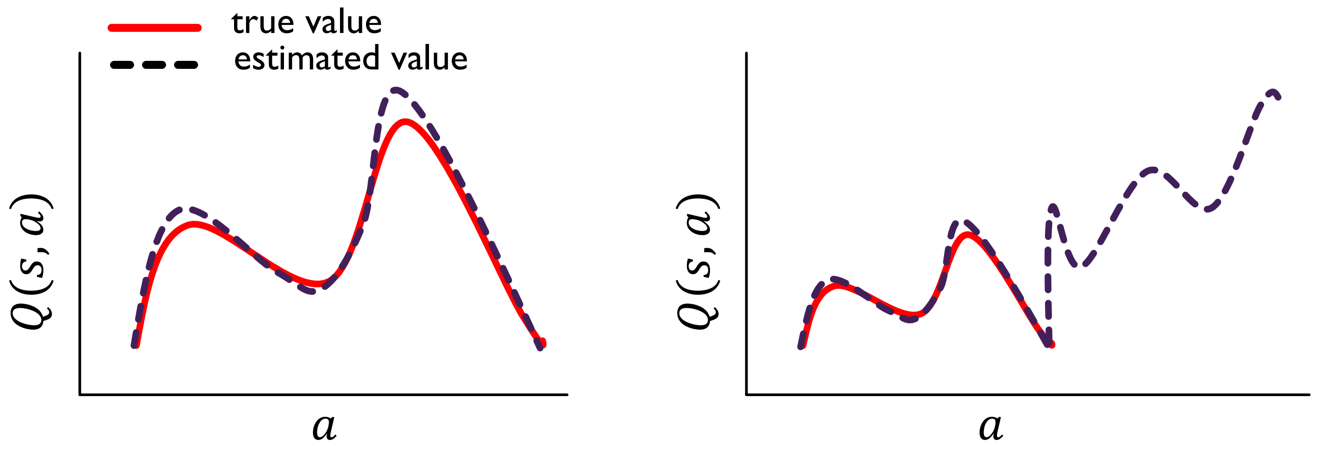
- Objective： learning goal-conditioned policies from offline dataset $D = \{(s_t, a_t, g, r_t)\}$ to maximize

$$J(\pi) = E_{s_0, g, \pi}[\Sigma_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}, g)]$$

Challenges

- Distribution Shift

- Learning with offline dataset $\mathcal{D}$ only guarantees predictions on the data distribution



- Overestimation on OOD actions

$$\mathcal{B}^{\pi}Q(s,a) = r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)}[\mathbb{E}_{\pi(a'|s')}[Q^{\pi}(s',a')]]$$
$$\pi = \arg\max_{\pi} \mathbb{E}_{s\sim D,\pi(a|s)}[Q(s,a)]$$

Fakoor R, et al. Continuous doubly constrained batch reinforcement learning. NeurIPS, 2021.

Challenges

- Generalization
- GCRL needs to reach multiple goals rather than overfitting to a single one

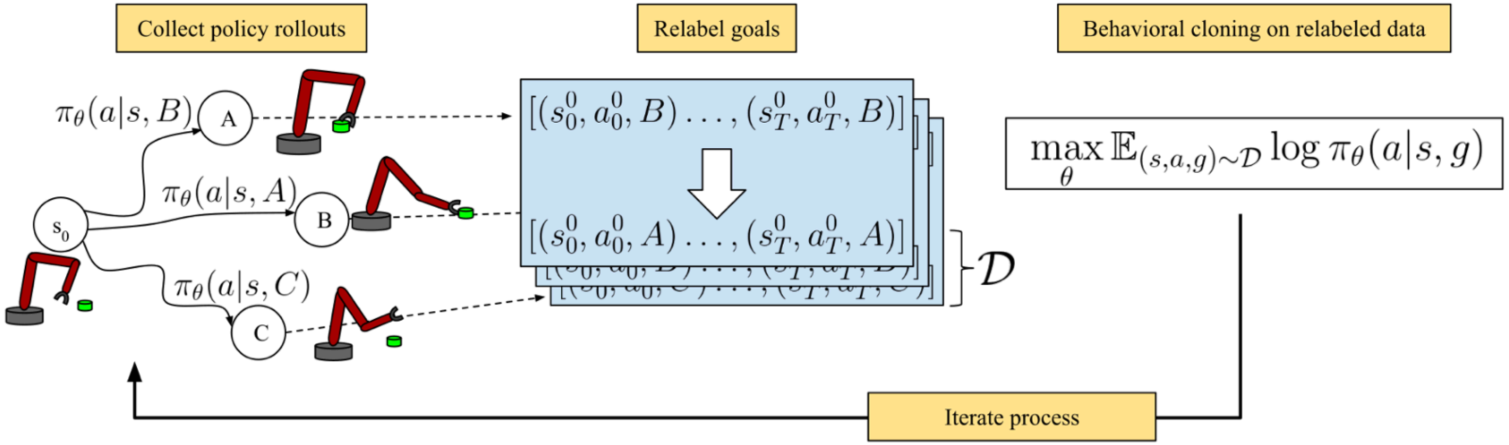
- Multi-modality
- In the offline dataset, there are generally multiple valid trajectories from a state to a goal, which may hinder learning a good policy

- Sparse Reward
- When the data is collected by random policy, there is little learning information for offline GCRL

Solving Offline GCRL via Supervised Learning

- Goal-conditioned Supervised Learning (GCSL)
- Relabeling data similar to Hindsight Experience Replay
- Imitation learning on relabeled data

$$J_{GCSL}(\pi) = \mathbb{E}_{(s_t, a_t, \phi(s_i)) \sim D_{relabel}} [\log \pi(a_t | s_t, \phi(s_i))]$$



Ghosh D, et al. Learning to Reach Goals via Iterated Supervised Learning. ICLR 2021

Revisiting GCSL

- GCSL alleviates OOD actions and sparse rewards naturally

$$J_{GCSL}(\pi) = \mathbb{E}_{(s_t, a_t, \phi(s_i)) \sim D_{relabel}}[\log \pi(a_t | s_t, \phi(s_i))]$$

Ghosh D, et al. Learning to Reach Goals via Iterated Supervised Learning. ICLR 2021

Revisiting GCSL

- GCSL alleviates OOD actions and sparse rewards naturally

$$J_{GCSL}(\pi) = \mathbb{E}_{(s_t, a_t, \phi(s_i)) \sim D_{relabel}}[\log \pi(a_t | s_t, \phi(s_i))]$$

- But GCSL only considers the last-step reward and weights all relabeled transitions equally

$$\mathbb{E}[\log \pi(a|s, g')] \quad \Longrightarrow \quad \mathbb{E}[r(s_T, a_T, g)]$$

Ghosh D, et al. Learning to Reach Goals via Iterated Supervised Learning. ICLR 2021

Revisiting GCSL

- GCSL alleviates OOD actions and sparse rewards naturally

$$J_{GCSL}(\pi) = \mathbb{E}_{(s_t, a_t, \phi(s_i)) \sim D_{relabel}}[\log \pi(a_t | s_t, \phi(s_i))]$$

- But GCSL only considers the last-step reward and weights all relabeled transitions equally

$$\mathbb{E}[\log \pi(a|s, g')] \longrightarrow \mathbb{E}[r(s_T, a_T, g)]$$

- We tackle this problem via weighted supervised learning

$$\mathbb{E}[\Sigma_t \gamma^t r(s_t, a_t, g)]$$

Ghosh D, et al. Learning to Reach Goals via Iterated Supervised Learning. ICLR 2021

Revisiting GCSL

- GCSL alleviates OOD actions and sparse rewards naturally

$$J_{GCSL}(\pi) = \mathbb{E}_{(s_t, a_t, \phi(s_i)) \sim D_{relabel}} [\log \pi(a_t | s_t, \phi(s_i))]$$

- But GCSL only considers the last-step reward and weights all relabeled transitions equally

$$\mathbb{E}[\log \pi(a|s, g')] \longrightarrow \mathbb{E}[r(s_T, a_T, g)]$$

$$\downarrow$$

- We tackle this problem via weighted supervised learning

$$\mathbb{E}[\boldsymbol{w} \log \pi(a|s, g')] \longleftarrow \mathbb{E}[\Sigma_t \gamma^t r(s_t, a_t, g)]$$

Ghosh D, et al. Learning to Reach Goals via Iterated Supervised Learning. ICLR 2021

Algorithm

- Weighted Goal-conditioned Supervised Learning（WGCSL）

$$J_{WGCSL}(\pi) = E_{g\sim p(g),\tau\sim\pi_b(\cdot|g),t\sim[0,T],i\sim[t,T]}\left[w_{i,t} log\pi_\theta(a_t|s_t,\phi(s_i))\right]$$

- $w_{i,t}$ includes 3 parts：

Algorithm

- Weighted Goal-conditioned Supervised Learning (WGCSL)

$$J_{WGCSL}(\pi) = E_{g\sim p(g),\tau\sim\pi_b(\cdot|g),t\sim[0,T],i\sim[t,T]}\left[w_{i,t} log\pi_\theta(a_t|s_t,\phi(s_i))\right]$$

- $w_{i,t}$ includes 3 parts:

① Discounted Relabeling Weight (DRW): γ^{i-t}

Theorem 1. *Assume a finite-horizon discrete MDP, a stochastic discrete policy π which selects actions with non-zero probability and a sparse reward function $r(s_t, a_t, g) = 1[\phi(s_t) = g]$, where ϕ is the state-to-goal mapping and $1[\phi(s_t) = g]$ is an indicator function. Given trajectories $\tau = (s_1, a_1, \cdots, s_T, a_T)$ and discount factor $\gamma \in (0, 1]$, let the weight $w_{t,i} = \gamma^{i-t}, t \in [1, T], i \in [t, T]$, then the following bounds hold:*

$$J_{surr}(\pi) \geq T \cdot J_{WGCSL}(\pi) \geq T \cdot J_{GCSL}(\pi),$$

where $J_{surr}(\pi) = \frac{1}{T}\mathbb{E}_{g\sim p(g),\tau\sim\pi_b(\cdot|g)}\left[\sum_{t=1}^{T} \log \pi(a_t|s_t, g) \sum_{i=t}^{T} \gamma^{i-1} \cdot 1[\phi(s_i) = g]\right]$ is a surrogate function of $J(\pi)$.

Algorithm

- Weighted Goal-conditioned Supervised Learning（WGCSL）

$$J_{WGCSL}(\pi) = E_{g\sim p(g),\tau\sim\pi_b(\cdot|g),t\sim[0,T],i\sim[t,T]}\left[w_{i,t} log\pi_\theta(a_t|s_t,\phi(s_i))\right]$$

- $w_{i,t}$ includes 3 parts：

① Discounted Relabeling Weight (DRW)： γ^{i-t}

② Goal-conditioned Exponential Advantage Weight (GEAW)： $\exp_{clip}\left(A\left(s_t, a_t, \phi(s_i)\right)\right)$

Exponential advantage weight is a commonly used technique in offline RL

$$\pi_{k+1} = \arg\max_{\pi\in\Pi} \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\mathbf{s})}[A^{\pi_k}(\mathbf{s},\mathbf{a})]$$
$$\text{s.t. } D_{\mathrm{KL}}(\pi(\cdot|\mathbf{s})||\pi_\beta(\cdot|\mathbf{s})) \leq \epsilon.$$

Wang Q et al. Exponentially weighted imitation learning for batched historical data. Advances in NeurIPS,2018

Algorithm

- Weighted Goal-conditioned Supervised Learning （WGCSL）

$$J_{WGCSL}(\pi) = E_{g\sim p(g),\tau\sim\pi_b(\cdot|g),t\sim[0,T],i\sim[t,T]}\left[w_{i,t} log\pi_\theta(a_t|s_t,\phi(s_i))\right]$$

- $w_{i,t}$ includes 3 parts：

① Discounted Relabeling Weight (DRW)： γ^{i-t}

② Goal-conditioned Exponential Advantage Weight (GEAW)： $\exp_{clip}\left(A\left(s_t,a_t,\phi(s_i)\right)\right)$

③ Best-Advantage Weight (BAW)： $\epsilon\left(A\left(s_t,a_t,\phi(s_i)\right)\right) = \begin{cases} 1, & A\left(s_t,a_t,\phi(s_i)\right) > \hat{A} \\ \epsilon_{min}, & otherwise \end{cases}$

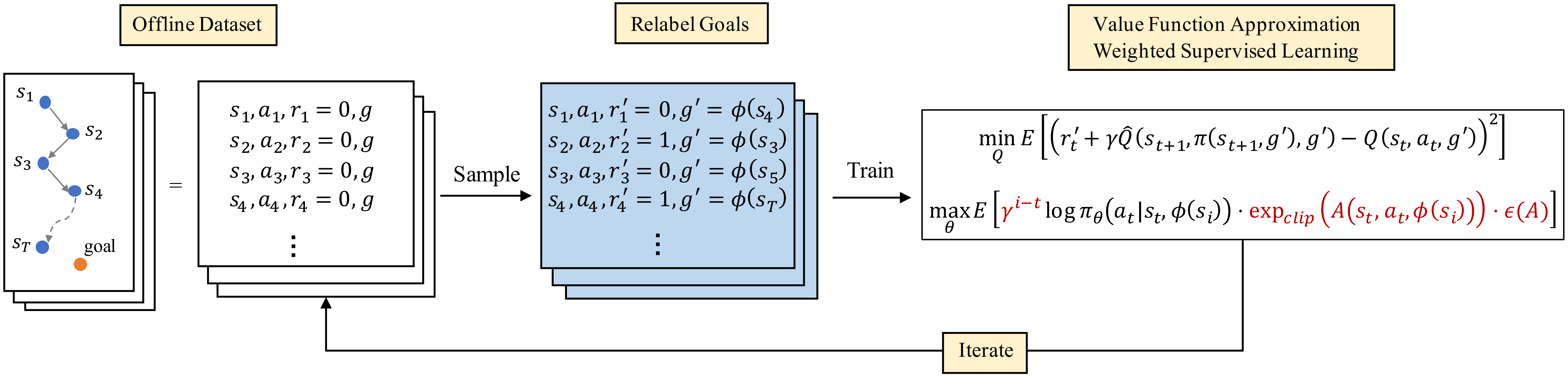
BAW selects the data to alleviate the multi-modality problem

In our implementation, $\epsilon_{min} = 0.05$

Curriculum learning: $\hat{A}$ is set as N percentile of advantage values, N gradually increases from 0 to 80

Algorithm

- Weighted Goal-conditioned Supervised Learning（WGCSL）



Experiments

- Experimental Settings
- Ten sparse reward goal-conditioned tasks
- Offline datasets are collected by online-trained HER agents (namely 'expert') and random policy ('random')
- 2×10^6 transitions for 4 harder tasks and 1×10^5 for others

Data Set	**PointReach**	**PointRooms**	**Reacher**	**SawyerReach**	**SawyerDoor**
Random	1.33	1.32	1.25	2.26	4.30
Expert	32.22	29.11	27.56	30.93	27.01
Data Set	**FetchReach**	**FetchPush**	**FetchSlide**	**FetchPick**	**HandReach**
Random	0.71	3.19	0.16	1.76	0.00
Expert	36.69	31.35	1.58	17.44	0.50

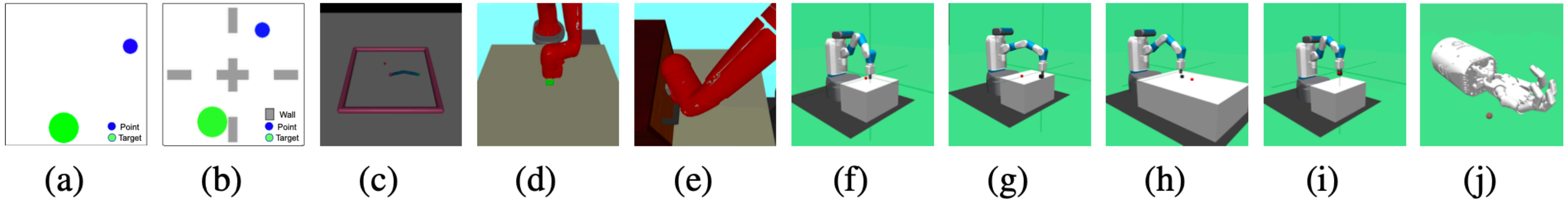
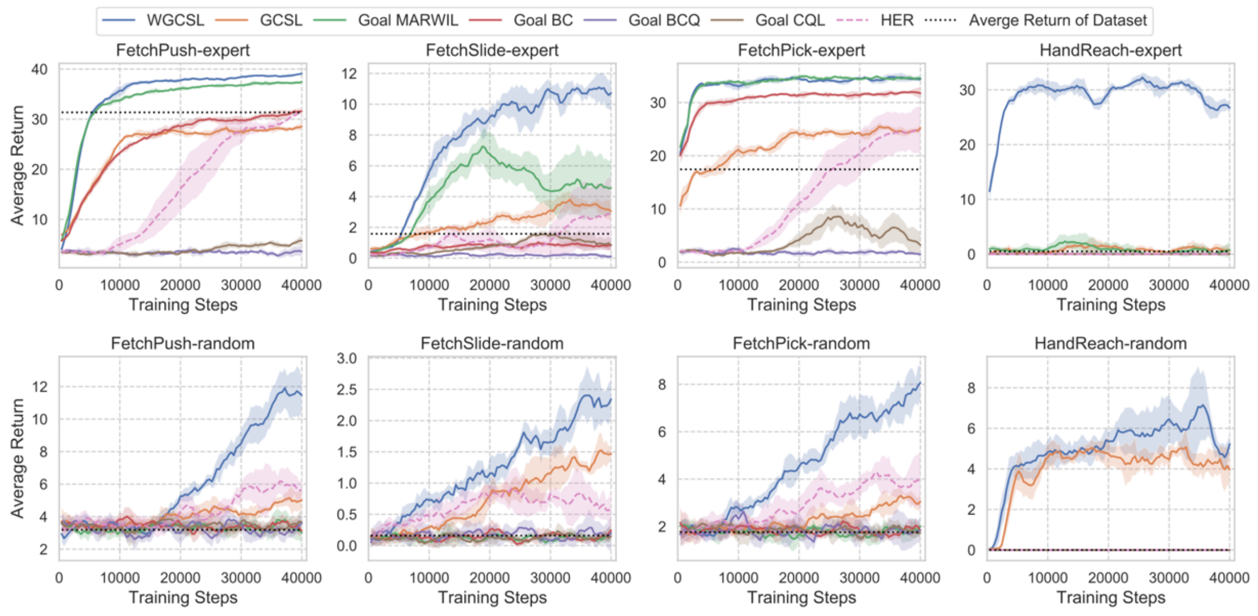


Figure 3: Goal-conditioned tasks: (a) PointReach, (b) PointRooms, (c) Reacher, (d) SawyerReach, (e) SawyerDoor, (f) FetchReach, (g) FetchPush, (h) FetchSlide, (i) FetchPick, (j) HandReach.

Experiments

- Experiment Results

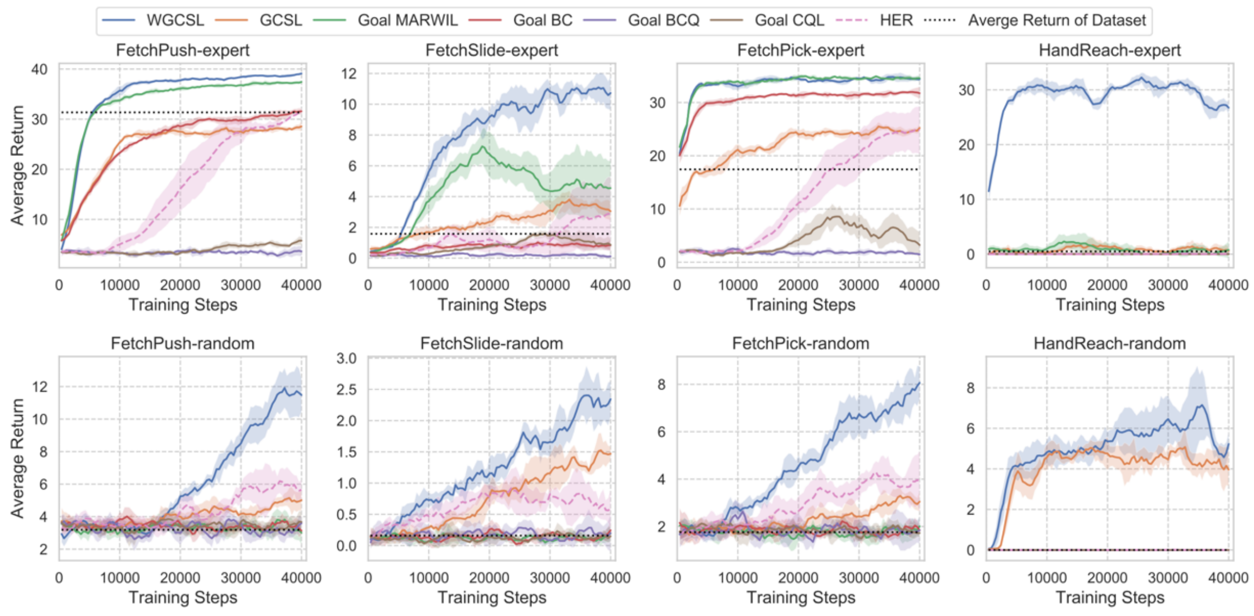


Task Name	WGCSL	GCSL	g-MARWIL	g-BC	g-BCQ	g-CQL	HER
PointReach-e	**44.40** ±0.14	39.27 ±0.48	**42.95** ±0.15	39.36 ±0.48	40.42 ±0.57	39.75 ±0.33	24.68 ±7.07
PointRooms-e	**36.15** ±0.85	33.05 ±0.54	**36.02** ±0.57	33.17 ±0.52	32.37 ±1.77	30.05 ±0.38	12.41 ±8.59
Reacher-e	**40.57** ±0.20	36.42 ±0.30	38.89 ±0.17	35.72 ±0.37	39.57 ±0.08	**42.23** ±0.12	8.27 ±4.33
SawyerReach-e	**40.12** ±0.29	33.65 ±0.38	37.42 ±0.31	32.91 ±0.31	**39.49** ±0.33	19.33 ±0.45	26.48 ±6.23
SawyerDoor-e	42.81 ±0.23	35.67 ±0.09	40.03 ±0.16	35.03 ±0.20	40.13 ±0.75	**45.86** ±0.11	**44.09** ±0.65
FetchReach-e	**46.33** ±0.04	41.72 ±0.31	**45.01** ±0.11	42.03 ±0.25	35.18 ±3.09	1.03 ±0.26	**46.73** ±0.14
FetchPush-e	**39.11** ±0.17	28.56 ±0.96	**37.42** ±0.22	31.56 ±0.61	3.62 ±0.96	5.76 ±0.83	31.53 ±0.47
FetchSlide-e	**10.73** ±1.09	3.05 ±0.62	4.55 ±1.79	0.84 ±0.35	0.12 ±0.10	0.86 ±0.38	2.86 ±2.40
FetchPick-e	**34.37** ±0.51	25.22 ±0.85	**34.56** ±0.54	31.75 ±1.19	1.46 ±0.29	3.23 ±2.52	24.79 ±4.49
HandReach-e	**26.73** ±1.20	0.57 ±0.68	0.81 ±1.59	0.06 ±0.03	0.04 ±0.04	0.00 ±0.00	0.05 ±0.07
PointReach-r	**44.30** ±0.24	30.80 ±1.74	7.67 ±1.97	1.37 ±0.09	1.78 ±0.14	1.52 ±0.26	**45.17** ±0.13
PointRooms-r	**35.52** ±0.80	24.10 ±0.81	4.67 ±0.80	1.43 ±0.18	1.61 ±0.17	1.29 ±0.37	**36.16** ±1.16
Reacher-r	**41.12** ±0.11	22.52 ±0.77	15.35 ±1.95	1.66 ±0.30	2.52 ±0.28	2.54 ±0.17	34.48 ±8.12
SawyerReach-r	**41.05** ±0.19	14.86 ±3.27	11.30 ±2.12	0.58 ±0.21	1.36 ±0.14	1.18 ±0.29	**39.27** ±2.16
SawyerDoor-r	**36.82** ±3.20	25.86 ±1.12	25.33 ±1.46	3.73 ±0.83	9.82 ±1.08	4.36 ±0.86	28.85 ±1.99
FetchReach-r	**46.50** ±0.09	38.26 ±0.24	30.86 ±8.49	0.84 ±0.31	0.19 ±0.04	0.97 ±0.23	**47.01** ±0.07
FetchPush-r	**11.48** ±1.03	5.01 ±0.64	3.01 ±0.71	3.14 ±0.25	3.60 ±0.42	3.67 ±0.65	5.65 ±0.64
FetchSlide-r	**2.34** ±0.28	**1.47** ±0.12	0.19 ±0.11	0.25 ±0.13	0.20 ±0.29	0.15 ±0.09	0.59 ±0.30
FetchPick-r	**8.06** ±0.71	3.05 ±0.42	2.01 ±0.46	1.84 ±0.17	1.84 ±0.58	1.73 ±0.17	3.91 ±1.29
HandReach-r	**5.23** ±0.55	**3.96** ±0.81	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.01

- WGCSL outperforms other baselines consistently

Experiments

- Experiment Results

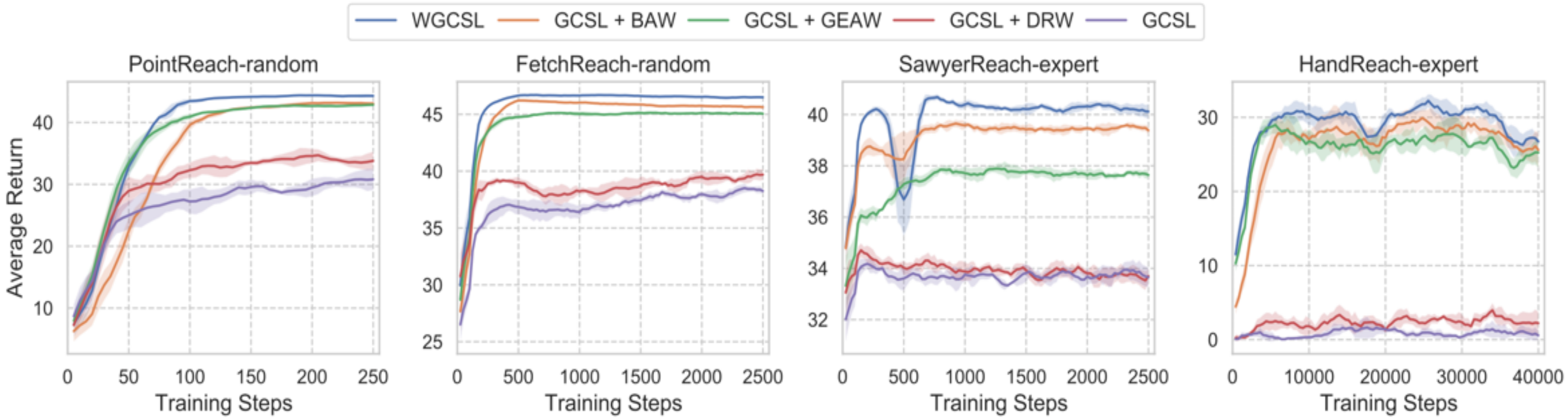


Task Name	WGCSL	GCSL	g-MARWIL	g-BC	g-BCQ	g-CQL	HER
PointReach-e	**44.40** ±0.14	39.27 ±0.48	**42.95** ±0.15	39.36 ±0.48	40.42 ±0.57	39.75 ±0.33	24.68 ±7.07
PointRooms-e	**36.15** ±0.85	33.05 ±0.54	**36.02** ±0.57	33.17 ±0.52	32.37 ±1.77	30.05 ±0.38	12.41 ±8.59
Reacher-e	**40.57** ±0.20	36.42 ±0.30	38.89 ±0.17	35.72 ±0.37	39.57 ±0.08	**42.23** ±0.12	8.27 ±4.33
SawyerReach-e	**40.12** ±0.29	33.65 ±0.38	37.42 ±0.31	32.91 ±0.31	**39.49** ±0.33	19.33 ±0.45	26.48 ±6.23
SawyerDoor-e	42.81 ±0.23	35.67 ±0.09	40.03 ±0.16	35.03 ±0.20	40.13 ±0.75	**45.86** ±0.11	**44.09** ±0.65
FetchReach-e	**46.33** ±0.04	41.72 ±0.31	**45.01** ±0.11	42.03 ±0.25	35.18 ±3.09	1.03 ±0.26	**46.73** ±0.14
FetchPush-e	**39.11** ±0.17	28.56 ±0.96	**37.42** ±0.22	31.56 ±0.61	3.62 ±0.96	5.76 ±0.83	31.53 ±0.47
FetchSlide-e	**10.73** ±1.09	3.05 ±0.62	4.55 ±1.79	0.84 ±0.35	0.12 ±0.10	0.86 ±0.38	2.86 ±2.40
FetchPick-e	**34.37** ±0.51	25.22 ±0.85	**34.56** ±0.54	31.75 ±1.19	1.46 ±0.29	3.23 ±2.52	24.79 ±4.49
HandReach-e	**26.73** ±1.20	0.57 ±0.68	0.81 ±1.59	0.06 ±0.03	0.04 ±0.04	0.00 ±0.00	0.05 ±0.07
PointReach-r	**44.30** ±0.24	30.80 ±1.74	7.67 ±1.97	1.37 ±0.09	1.78 ±0.14	1.52 ±0.26	**45.17** ±0.13
PointRooms-r	**35.52** ±0.80	24.10 ±0.81	4.67 ±0.80	1.43 ±0.18	1.61 ±0.17	1.29 ±0.37	**36.16** ±1.16
Reacher-r	**41.12** ±0.11	22.52 ±0.77	15.35 ±1.95	1.66 ±0.30	2.52 ±0.28	2.54 ±0.17	34.48 ±8.12
SawyerReach-r	**41.05** ±0.19	14.86 ±3.27	11.30 ±2.12	0.58 ±0.21	1.36 ±0.14	1.18 ±0.29	**39.27** ±2.16
SawyerDoor-r	**36.82** ±3.20	25.86 ±1.12	25.33 ±1.46	3.73 ±0.83	9.82 ±1.08	4.36 ±0.86	28.85 ±1.99
FetchReach-r	**46.50** ±0.09	38.26 ±0.24	30.86 ±8.49	0.84 ±0.31	0.19 ±0.04	0.97 ±0.23	**47.01** ±0.07
FetchPush-r	**11.48** ±1.03	5.01 ±0.64	3.01 ±0.71	3.14 ±0.25	3.60 ±0.42	3.67 ±0.65	5.65 ±0.64
FetchSlide-r	**2.34** ±0.28	**1.47** ±0.12	0.19 ±0.11	0.25 ±0.13	0.20 ±0.29	0.15 ±0.09	0.59 ±0.30
FetchPick-r	**8.06** ±0.71	3.05 ±0.42	2.01 ±0.46	1.84 ±0.17	1.84 ±0.58	1.73 ±0.17	3.91 ±1.29
HandReach-r	**5.23** ±0.55	**3.96** ±0.81	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.01

- WGCSL outperforms other baselines consistently
- WGCSL can even learn reasonable policies from random datasets

Experiments

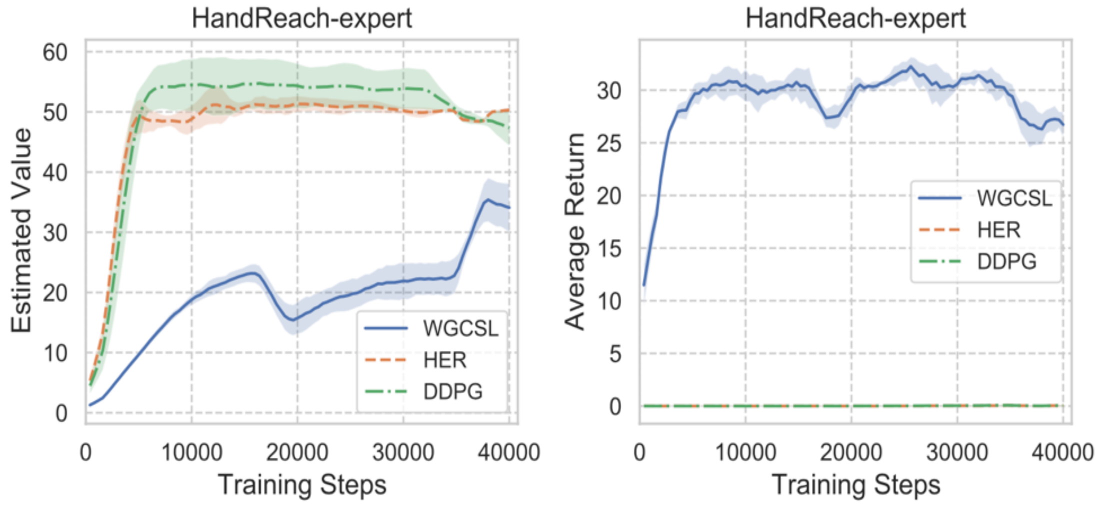
- Ablation Studies



- BAW, GEAW, DRW are all effective on top of GCSL
- Learned policy can be improved by combining all three weights.

Experiments

- Value Estimation



- DDPG and HER exhibit large estimated values
- WGCSL has a more robust value approximation

Summary

- We propose WGCSL, a weighted supervised learning method for offline goal-conditioned RL
- We provide a benchmark and offline datasets
- WGCSL outperforms current approaches significantly in learning efficiency and performance

Summary

- We propose WGCSL, a weighted supervised learning method for offline goal-conditioned RL
- We provide a benchmark and offline datasets
- WGCSL outperforms current approaches significantly in learning efficiency and performance

Please refer to our paper for more details and analysis of our method

Website

Paper