Conservative Data Sharing for Multi-Task Offline RL

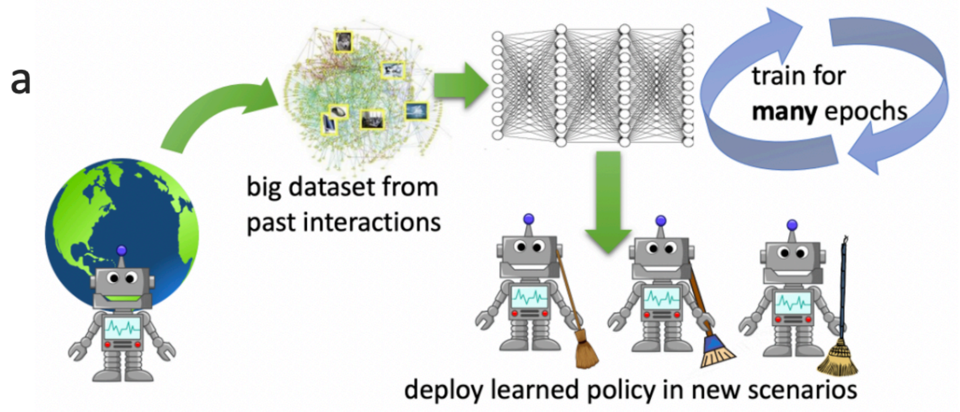
Tianhe Yu*, Aviral Kumar*, Yevgen Chebotar, Karol Hausman, Sergey Levine, Chelsea Finn

Stanford Artificial Intelligence · Google AI · BAIR Berkeley Artificial Intelligence Research

Using all data for all tasks doesn't always work in multi-task offline RL. We devise a scheme, CDS, to enable intelligent data sharing.

Offline Reinforcement Learning

- **Goal**: Learn a (good) policy directly from a fixed dataset of interactions
- Several advances in handling problems: distributional shift, overestimation, etc.



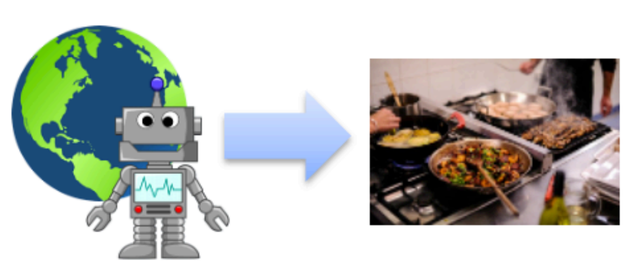
However, current algorithms cannot leverage heterogenous "general" datasets, to solve multiple problems

Can we devise techniques to leverage diverse, heterogenous data?

Multi-Task Learning to the Rescue

Specialist in one task — Single-task RL

Generalist in various tasks — Multi-task RL



solving one task using an offline dataset

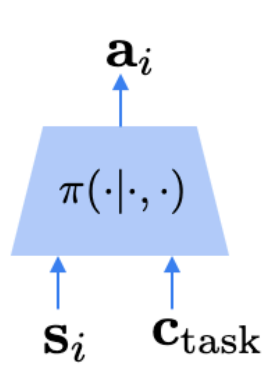
solving multiple tasks using given data

Can we apply multi-task RL on large offline datasets?

Goal of multi-task RL: learn a single policy that solves multiple tasks more efficiently than learning each task independently.

Multi-Task Offline RL: Key Ingredients

- **Parameter Sharing:**
  - Train a context-conditioned policy, context identifies the task



Optimization issues, right way to condition, etc.

- **Data Sharing:**
  - Relabel data from one task to the other
  - Can be effective in **sequential** problems due to temporal stitching
  - Is widely adopted in goal-conditioned RL (so why not in multi-task RL?)

This work!

Could data sharing help in multi-task offline RL?

Could it hurt?

What is the right way to share data?

Does Data Sharing Always Help?

Task Name	#Eps.	QT-Opt	$f_{I_{\text{orig}}}$, rand QT-Opt MultiTask	$f_{I_{\text{orig}}}$, rebal	$f_{I_{\text{all}}}$, rand DataShare MultiTask
lift-any	635K	0.88	0.94	0.85	0.62
lift-banana	9K	0.04	0.13	0.38	0.09
lift-bottle	11K	0.02	0.16	0.66	0.15
lift-sausage	5K	0.02	0.10	0.38	0.15
lift-milk	6K	0.01	0.13	0.42	0.13
lift-box	6K	0.00	0.12	0.16	0.08
lift-can	6K	0.01	0.16	0.46	0.07
lift-carrot	80K	0.71	0.41	0.72	0.37
place-any	30K	N/A	**0.86**	0.74	0.30
place-bottom	5K	N/A	0.43	0.57	0.30
place-top-right	4K	N/A	0.16	**0.55**	0.08
place-top-left	4K	N/A	0.23	**0.75**	0.19

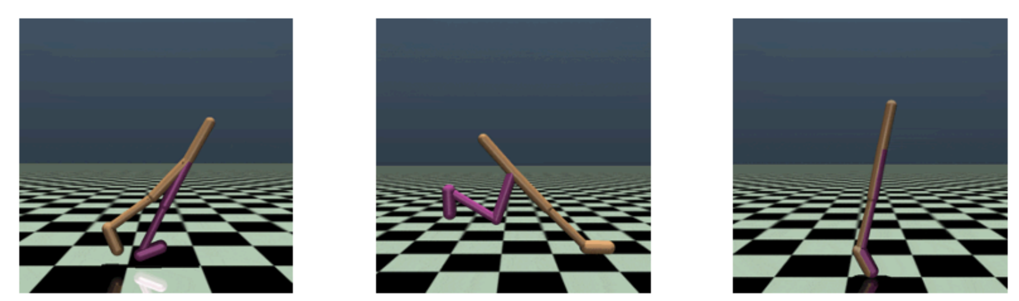
With sharing

Without sharing

Data sharing can hurt in some cases!

MT-OPT: Continuous Multi-Task Robotic Reinforcement Learning at Scale, Kalashnikov et. al., CoRL 2021

When does naively sharing data hurt?

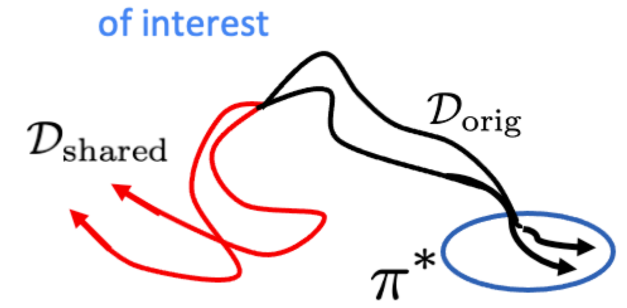


Dataset types / Tasks	Dataset Size	Avg Return No Sharing	Avg Return Sharing All	$D_{\text{KL}}(\pi,\pi_\beta)$ No Sharing	$D_{\text{KL}}(\pi,\pi_\beta)$ Sharing All
medium / `run forward`	27646	297.4	**848.7**	**6.53**	11.78
medium / `run backward`	31298	207.5	**600.4**	**4.44**	10.13
medium / `jump`	100000	351.1	**776.1**	**5.57**	21.27
average task performance	N/A	285.3	**747.7**	**5.51**	14.39
medium-replay / `run forward`	109900	590.1	**701.4**	**1.49**	7.76
medium / `run backward`	31298	614.7	**756.7**	**1.91**	12.2
expert / `jump`	5000	**1575.2**	885.1	**3.12**	27.5
average task performance	N/A	**926.6**	781	**2.17**	15.82

Reduced performance

Increased divergence

- Sharing data across tasks generally helps
- It hurts performance when sharing data increases deviation (divergence) from the optimal policy of the task of interest



CDS: Conservative Data Sharing

How can we balance the various factors that affect the performance of data sharing?

A simple approach works:

Share transitions with high **conservative Q-values.**

Q-values $\hat{Q}^{\pi}$ obtained via

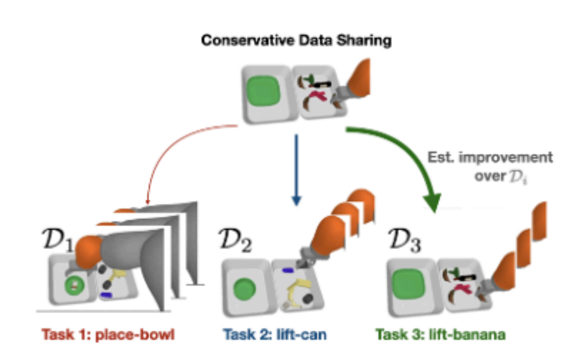
$$\widehat{r}(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) - D(\pi,\pi_\beta)(\mathbf{s},\mathbf{a}) \quad \text{[BRAC]}$$

$$\min_{Q} \;\; \alpha E_\pi[Q(\mathbf{s},\mathbf{a})] + \mathrm{TD}(\mathbf{s},\mathbf{a},Q) \quad \text{[CQL]}$$

Intuition: Conservative Q-values adequately balance:
(a) "goodness" of data (i.e., rewards in the data),
(b) distributional shift

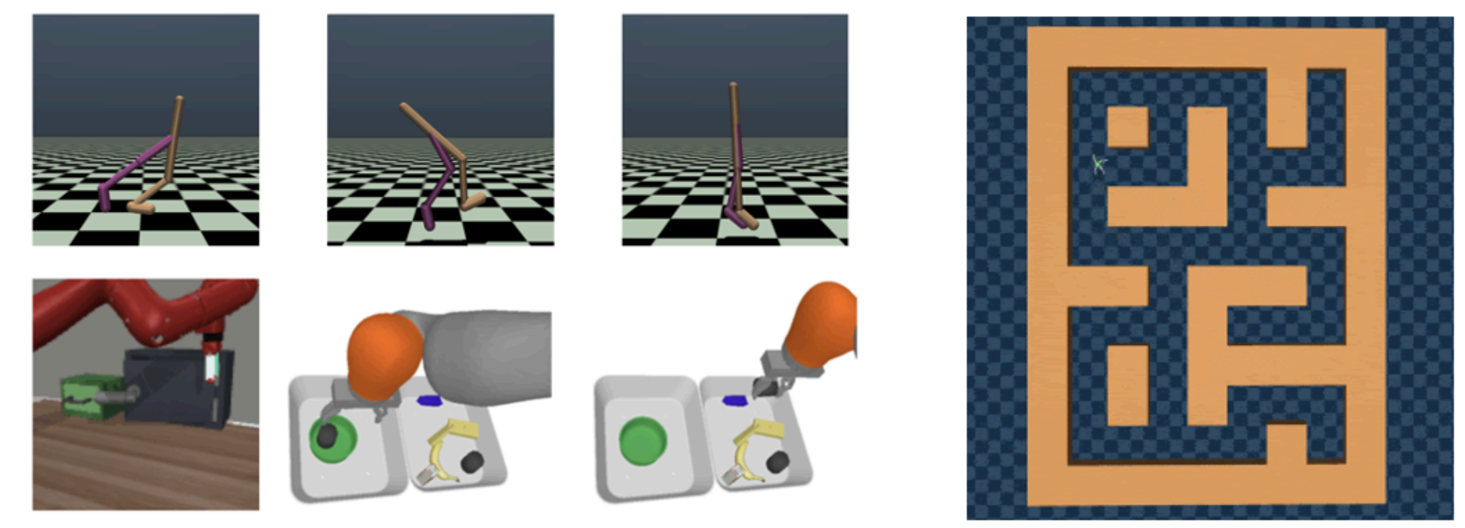
All we need to do now is to relabel **many** datapoints to increase sample size.....

CDS: relabels transitions if the Q-value of a transition shared from task j to task i exceeds the top-k percentile of the Q-values of all datapoints for task i.



Experimental Evaluation

Wide range of tasks: locomotion, navigation and manipulation tasks



Experiment results (low-dimensional inputs)

Environment	Tasks / Dataset type	CDS (ours)	CDS (basic)	HIPI [16]	Sharing All	No Sharing
walker2d	run forward / medium-replay	**1057.9**±121.6	968.6±188.6	695.5±61.9	701.4±47.0	590.1±48.6
	run backward / medium	564.8±47.7	594.5±22.7	626.0±48.0	**756.7**±76.7	614.7±87.3
	jump / expert	1418.2±138.4	1501.8±115.1	**1603.7**±146.8	885.1±152.9	1575.2±70.9
	average	1013.6±71.5	**1021.6**±76.9	975.1±45.1	781.0±100.8	926.6±37.7
Meta-World [90]	door open / medium-replay	**58.4%**±9.3%	30.1%±16.6%	26.5%±20.5%	34.3%±17.9%	14.5%±12.7
	door close / expert	**65.3%**±27.7%	41.5%±28.2%	1.3%±5.3%	48.3%±27.3%	4.0%±6.1%
	drawer open / expert	**57.9%**±16.2%	39.4%±16.9%	41.2%±24.9%	55.1%±9.4%	16.0%±17.5%
	drawer close / medium-replay	98.8%±0.7%	86.3%±0.9%	62.2%±33.4%	**100.0%**±0%	99.0%±0.7%
	average	**70.1%**±8.1%	49.3%±16.0%	32.8%±18.7%	59.4%±5.7%	33.4%±8.3%
AntMaze [19]	large maze (7 tasks) / undirected	**22.8%** ± 4.5%	10.0% ± 5.9%	1.3% ± 2.3%	16.7% ± 7.0%	13.3% ± 8.6%
	large maze (7 tasks) / directed	**24.6%** ± 4.7%	0.0% ± 0.0%	11.8% ± 5.4%	20.6% ± 4.4%	19.2% ± 8.0%
	medium maze (3 tasks) / undirected	**36.7%** ± 6.2%	0.0% ± 0.0%	8.6% ± 3.2%	22.9% ± 3.6%	21.6% ± 7.1%
	medium maze (3 tasks) / directed	**18.5%** ± 6.0%	0.0% ± 0.0%	8.3% ± 9.1%	12.4% ± 5.4%	**17.0%** ± 3.2%

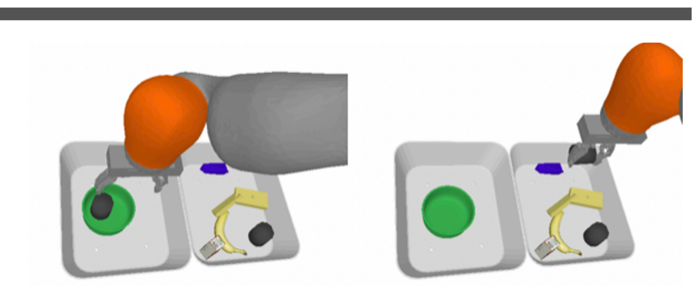
Relabeling Direction	CDS weight
door close → door open	0.46
drawer open → door open	0.10
drawer close → door open	0.02
drawer open → drawer close	0.35
door open → drawer close	0.26
door close → drawer close	0.22

CDS assigns high weights to more relevant tasks.

Does CDS prevent excessive distributional shift?

Environment	Dataset types / Tasks	$D_{\text{KL}}(\pi,\pi_\beta)$ No Sharing	$D_{\text{KL}}(\pi,\pi_\beta)$ Sharing All	CDS (ours)
walker2d	medium-replay / run forward	**1.49**	7.76	**1.49**
	medium / run backward	**1.91**	12.2	6.09
	expert / jump	3.12	27.5	**2.91**

- CDS reduces the KL divergence between the single-task optimal policy and the behavior policy after relabeling.



Vision-based results

Task Name	CDS (ours)	HIPI [16]	Skill [33]	Sharing All	No Sharing
`lift-banana`	**53.1%**±3.2%	48.3%±6.0%	32.1%±9.5%	41.8%±4.2%	20.0%±6.0%
`lift-bottle`	**74.0%**±6.3%	64.4%±7.7%	55.9%±9.6%	60.1%±10.2%	49.7%±8.7%
`lift-sausage`	**71.8%**±3.9%	71.0%±7.7%	68.8%±9.3%	70.0%±7.0%	60.9%±6.6%
`lift-milk`	**83.4%**±5.2%	79.0%±3.9%	68.2%±3.5%	72.5%±5.3%	68.4%±6.1%
`lift-food`	61.4%±9.5%	**62.6%**±6.3%	41.5%±12.1%	58.5%±7.0%	39.1%±7.0%
`lift-can`	65.5%±6.9%	**67.8%**±6.8%	50.8%±12.5%	57.7%±7.2%	49.1%±9.8%
`lift-carrot`	**83.8%**±3.5%	78.8%±6.9%	66.0%±7.0%	75.2%±7.6%	69.4%±7.6%
`place-bowl`	**81.0%**±8.1%	77.2%±8.9%	80.8%±6.9%	70.8%±7.8%	80.3%±8.6%
`place-plate`	85.8%±6.6%	83.6%±7.9%	78.4%±9.6%	78.7%±7.6%	**86.1%**±7.7%
`place-divider-plate`	**87.8%**±7.6%	78.0%±10.5%	80.8%±5.3%	79.2%±6.3%	85.0%±5.9%
average	**74.8%**±6.4%	71.1%±7.5%	62.3%±8.9%	66.4%±7.2%	60.8%±7.5%

Paper Link: