Rethinking Optimal Transport in Offline Reinforcement Learning (Asadulaev et al.)

Discussions by: Jethro (Cheuk Sau) Au, Suhwan Bong

11/18/2025

Background: Reinforcement Learning

Reinforcement Learning

Framework for modelling decision-making process as MDPs.
Goal: Find a policy that maximizes the expected cumulative reward:

$$J(\pi) \triangleq \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right].$$

Critic Function

We also define $d^\pi(s)$ as the distribution over the state space when following policy π. To estimate the expected cumulative reward for a given policy π, the critic function $Q^\pi(s, a)$ is used:

$$Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim T(s,a),\, a' \sim \pi(s')}\left[Q^\pi(s', a')\right].$$

Remark: Critic Learning in Deep RL

The critic can be learned by minimizing the mean squared Bellman error over an experience replay dataset $\mathcal{D} = \{(s_i, a_i, s_i', r_i)\}$, which consists of trajectories generated by the policy π. This objective is given by:

$$\min_{\phi}\ \mathbb{E}_{(s,a,s') \sim \mathcal{D}}\left[r(s, a) + \gamma\ \mathbb{E}_{a' \sim \pi_\theta(s')} Q_\phi^\pi(s', a') - Q_\phi^\pi(s, a)\right]^2.$$

Background: Offline RL & OT

Offline RL and Behavioural Cloning

Implementation of online RL or collection of Rl data can be too costly.

Critic Q can be reframed as a supervised learning approach be offline collected action-state data D by expert policy β to estimate Q in an offline manner.

Challenge with Offline RL

Learned policy may have **distribution shift** if it expert selected actions not representative of $\mathcal{D}$!

Past Work: OT for Behavior Cloning (WBRAC)

Optimal Transport (OT) measures, such as the Wasserstein-1 distance.
Author: EMD distance no mechanism to infer importance of each action, and too complicated to calculate.

$$\min_\pi \max_{\|f\|_L \leq 1} \mathbb{E}_{s\sim\mathcal{D},a\sim\pi(s)}\Big[\underbrace{-Q^\pi(s,a)}_{\text{Critic}}\Big] + \alpha\Big(\underbrace{\mathbb{E}_{(s,a)\sim\mathcal{D}}\big[f(s,a)\big] - \mathbb{E}_{s\sim\mathcal{D},a\sim\pi(s)}\big[f(s,a)\big]}_{\textit{Wasserstein-1}\text{ distance}}\Big) \quad (8)$$

where α is a behavior cloning coefficient.

Motivation: OT Extensions

Maximin OT Formulation (Korotin et al. 2022)

Simultaneously computes OT distance and OT map T:

$$\max_f \min_T \mathbb{E}_{x\sim\mu}[c(x, T(x)) - f(T(x))] + \mathbb{E}_{y\sim\nu}[f(y)]$$

where $T : X \to Y$ and f is the Kantorovich potential. The Rockafellar interchange theorem enables this efficient neural solution for strong and weak OT.

Extremal OT Formulation (Gazdieva et al. 2023)

Balances matching strictness with w:

$$\mathrm{EOT}_w(\mu,\nu) := \min_{\pi \le w\cdot\nu,\ \pi\in\Pi(\mu,\nu)} \int c(x,y)\,\mathrm{d}\pi(x,y)$$

where only a $1/w$ fraction of ν must be matched.
$w > 1$: Only the closest $1/w$ fraction of ν is matched; rest is ignored.

w	Matched ν	Description
1	100%	Full match
2	50%	Closest half
3	33%	Closest third

Methodology: Entire RL as OT problem

Step 1: RL as Monge/Kantovorich OT

Offline RL can be framed as an optimal transport (OT) problem:

$$\min_{\pi_{\#} d(s) = \beta(\cdot|s)} \mathbb{E}_{s \sim \mathcal{D},\, a \sim \pi(s)} \left[-Q^{\pi}(s, a) \right]$$

Step 2: Stitching with Partial OT

To focus on *optimal* actions within β, we relax to a partial alignment:

$$\min_{\pi_{\#} d(s) \leq w\, \beta(\cdot|s)} \mathbb{E}_{s \sim \mathcal{D},\, a \sim \pi(s)} \left[-Q^{\pi}(s, a) \right]$$

The dual for the partial OT problem can be formulated as:

$$\max_{f \geq 0} \mathbb{E}_{s \sim \mathcal{D},\, a \sim \pi} f_c(s, a) + w\, \mathbb{E}_{s \sim \mathcal{D},\, a \sim \beta} f(s, a)$$

where f_c is a cost-related function parameterized by neural networks, and w is the unbalance coefficient.

- ⋆ **Key distinction with other OT papers:** Instead of adding OT regularization, the entire policy optimization is cast as an OT problem, mapping only to optimal actions as defined by Q.

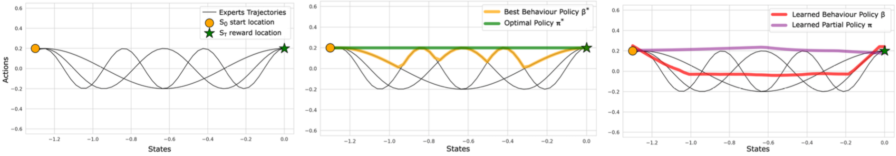
Methodology: Partial Policy Learning

$$\max_{f\geq 0}\min_{\pi}\mathbb{E}_{s\sim\mathcal{D},a\sim\pi(s)}\Big[\underbrace{-Q^{\pi}(s,a)}_{\text{Cost}}-\underbrace{f(s,a)\Big]+w\mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[f(s,a)\Big]}_{\text{Constraints}}.$$

Algorithm 1 Partial Policy Learning

Input: Dataset $\mathcal{D}(s,a,r,s')$
Initialize $Q_\phi, \pi_\theta, f_\omega, \beta$
for k in 1...N **do**
  $(s,a,r,s') \leftarrow \mathcal{D}$: sample a batch of transitions from the dataset.
  $Q^{k+1} \leftarrow$ Update cost function Q_ϕ^π using the Bellman update in (2).
  $f^{k+1} \leftarrow$ Update f_ω using outputs of π_θ and samples from dataset: $\arg\min_f -\mathbb{E}_{s\sim\mathcal{D},a\sim\pi^k(s)}[f^k(s,a)] + w\mathbb{E}_{s,a\sim\mathcal{D}}[f^k(s,a)]$
  $\pi^{k+1} \leftarrow$ Update policy π_θ as a transport map: $\arg\min_\pi \mathbb{E}_{s\sim\mathcal{D},a\sim\pi^k(s)}[-Q^k(s,a) - f^k(s,a)]$.
end for

Toy Experiments



(a) (b) (c)

- Compared to offline RL (b), new method demonstrated superior performance by extracting and exploiting the insights from the data, ignoring all inefficient actions.

D4RL Experiments

Table 1: Averaged normalized scores on Antmaze-v2 tasks. Reported scores are the results of the final 100 evaluations and 5 random seeds.

Dataset	IQL	OTR+IQL	CQL	PPL^{CQL}(Ours)	ReBRAC	PPL^{R}(Ours)
umaze	87.5 ± 2.6	83.4 ± 3.3	86.3 ±3.7	90±2.6	97.8 ± 1.0	98.0 ±14
umaze-diverse	62.2 ± 13.8	68.9 ± 13.6	34.6 ±20.9	40±2.6	88.3 ± 13.0	93.6±6.1
medium-play	71.2 ± 7.3	70.5 ± 6.6	63.0 ±9.8	67.3±10.1	84.0 ± 4.2	90.2 ±3.1
medium-diverse	70.0 ± 10.9	70.4 ± 4.8	59.6 ±3.5	65.3±8.0	76.3 ± 13.5	84.8 ±14.7
large-play	39.6 ± 5.8	45.3 ± 6.9	20.0 ±10.8	25.6±3.7	60.4 ± 26.1	76.8 ±4.0
large-diverse	47.5 ± 9.5	45.5 ± 6.2	20.0 ±5.1	23.6 ±11.0	54.4 ± 25.1	76.6 ± 7.4
Total	378	384	283.5	311.8	461.2	**520**

- The novel method provides state-of-the-art results for all datasets on this task, and gives a significant improvement of (+16) and (+21) points for the large environments.
- They consistently outperform the previous best OT-based offline RL algorithm, OTR+IQL.

D4RL Experiments

Table 2: Averaged normalized scores on `MuJoCo` tasks. Reported scores are the results of the final 10 evaluations and 5 random seeds.

	Dataset	BC	One-RL	CQL	IQL	OTR+IQL	TD3+BC	ReBRAC	PPL^{R}(Ours)
M	Half.	42.6	48.4	44.0	47.4	43.3	48.3	65.6	64.95±0.2
	Hopper	52.9	59.6	58.5	66.3	78.7	59.3	102.0	93.49±7.2
	Walker	75.3	81.1	72.5	78.3	79.4	65.5	82.5	85.66±0.6
MR	Half.	36.6	38.1	45.5	44.2	41.3	44.6	51.0	51.1±0.3
	Hopper	18.1	97.5	95.0	94.7	84.8	60.9	98.1	100.0±2
	Walker	26.0	49.5	77.3	73.9	66.0	81.8	77.3	78.66±2.0
ME	Half.	55.2	93.4	91.6	86.7	89.6	90.7	101.1	104.85±0.1
	Hopper	52.5	103.3	105.4	91.5	93.2	98.0	107.0	109.0±1.2
	Walker	107.5	113.0	108.8	109.6	109.3	110.1	112.3	111.74±1.1
	Total	467.7	684.9	698.6	692.6	685.6	659.2	796.9	**799.45**

- We can interpret that the new method lies between behavior cloning and direct maximization of the Q function.

Critical Reflection and Limitations

- **Choice of ω:** ω controls the policy's support (action range) and thus action selection, but it is set arbitrarily across tasks. Task-adaptive tuning may improve performance.
- **OT baselines:** Lack of comparisons with other OT-based RL methods under matched datasets, metrics, and compute.
- **Ablations:** Lack of a comprehensive ablation study (components, training objectives, hyperparameters).
- **Reporting:** Minor inconsistencies in formatting and presentation of results.

Discussion and Future Directions

- The authors introduces a novel algorithm for offline RL using optimal transport.
- The novel algorithm effectively selects and maps the best expert actions for each given state.
- Using the authors' formulation, other OT methods also can be integrated into RL. (e.g., Various regularizations or general costs)
- Weak Neural OT can be relevant in RL where stochastic behavior is preferred for exploration in the presence of multimodal goals.