Offline Multitask Representation Learning for Reinforcement Learning

Haque Ishfaq [1,2], Thanh Nguyen-Tang [3], Songtao Feng [4], Raman Arora [3], Mengdi Wang [5], Ming Yin [5], Doina Precup [1,2]

[1]Mila – Quebec AI Institute [2]McGill University [3]Johns Hopkins University [4]University of Florida [5]Princeton University

Highlights

- We study offline multitask representation learning in reinforcement learning.
- Learner is provided with offline datasets from different tasks with shared representation.
- We prove our proposed algorithm can learn near-accurate model and near-optimal policies.
- We show theoretical benefits of using learned representation in downstream reward-free, offline and online RL tasks.

Setting

- We consider **low-rank** episodic MDPs $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$: two unknown embedding functions $\phi_h^*: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\mu_h^*: \mathcal{S} \to \mathbb{R}^d$ such that for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, $P_h^*(s' \mid s, a) = \langle \phi_h^*(s,a), \mu_h^*(s') \rangle$.
- **Value function** of policy π:

$$V_{h,P,r}^{\pi}(s) = \mathbb{E}_{(s_{h'},a_{h'})\sim(P,\pi)}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) | s_h = s\right].$$

- Upstream offline multitask learning: T tasks, each task t: $\mathcal{M}^t = (\mathcal{S}, \mathcal{A}, H, P^t, r^t)$.

$$P_h^{(*,t)}(s' \mid s, a) = \langle \phi_h^*(s,a), \mu_h^{(*,t)}(s') \rangle, \quad \forall s, s' \in \mathcal{S}, a \in \mathcal{A}.$$

- We have access to offline dataset $\mathcal{D} = \bigcup_{t\in[T],h\in[H]} \mathcal{D}_h^{(t)}$, where $\mathcal{D}_h^{(t)} = \{(s_h^{(i,t)}, a_h^{(i,t)}, r_h^{(i,t)}, s_{h+1}^{(i,t)}\}_{i\in[n]}$ with $s_{h+1}^{(i,t)} \sim P_h^{(*,t)}(\cdot \mid s_h^{(i,t)}, a_h^{(i,t)})$ and $\mathcal{D}_h^{(t)}$ was collected using a *fixed behavior policy* π_t^b.
- Downstream target task $T+1$ with

$$P^{(*,T+1)}(s' \mid s, a) = \langle \phi_h^*(s,a), \mu_h^{(*,T+1)}(s') \rangle, \quad \forall s, s' \in \mathcal{S}, a \in \mathcal{A}.$$

Goal of Upstream and Downstream Tasks

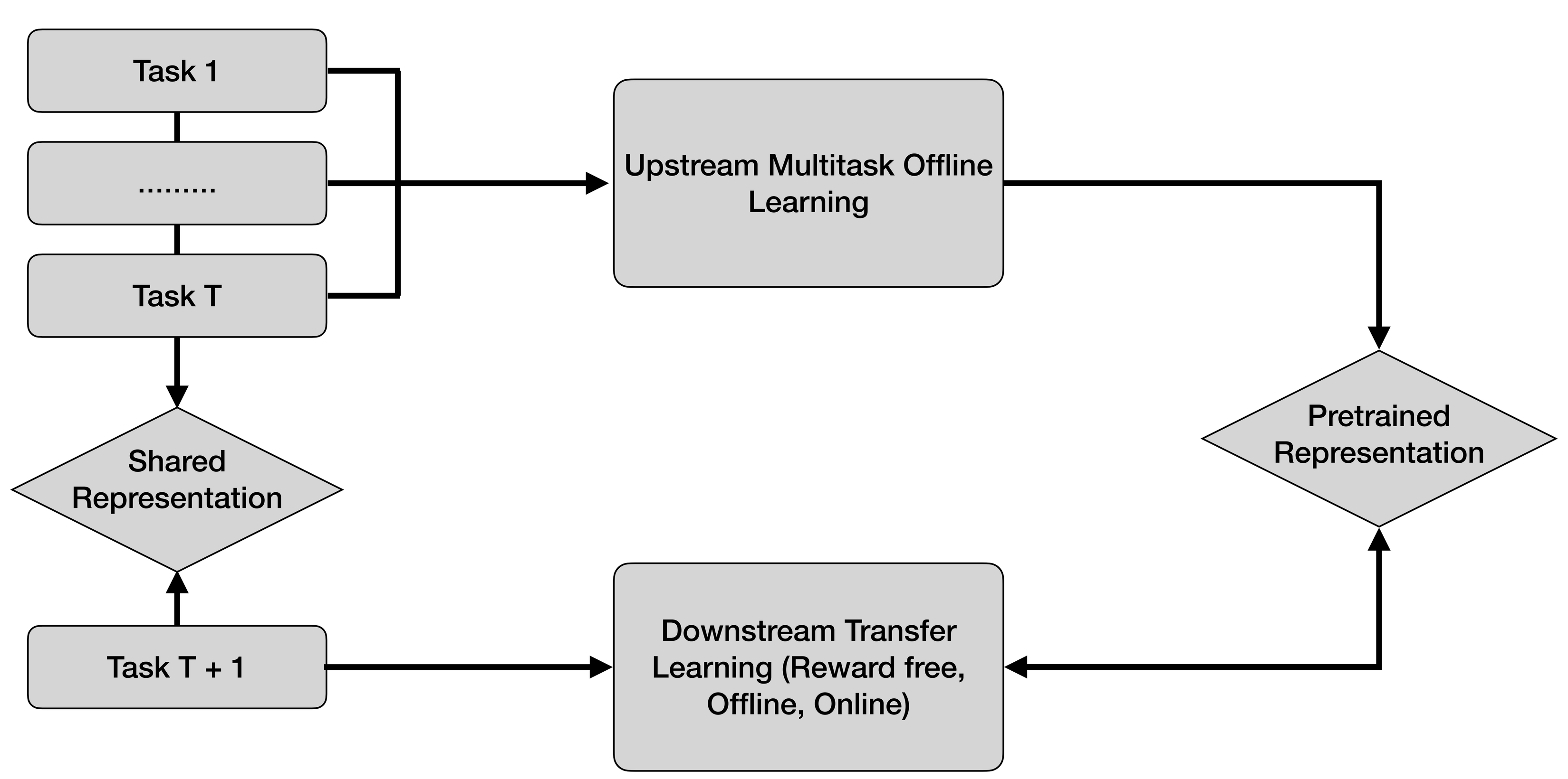

Figure 1. Upstream and Downstream Task Overview

Algorithm: Multitask Offline Representation Learning (MORL)

- Learning jointly via offline Multitask Maximum Likelihood Estimation (MLE) oracle

$$\left(\widehat{\phi}_h, \widehat{\mu}_h^{(1)}, \dots, \widehat{\mu}_h^{(T)}\right) = \underset{\phi_h\in\Phi, \mu_h^{(1)},\dots,\mu_h^{(T)}\in\Psi}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{t=1}^{T} \log\left(\left\langle \phi_h(s_h^{(i,t)}, a_h^{(i,t)}), \mu_h^t(s_{h+1}^{(i,t)} \right\rangle\right).$$

- for each $t \in \{1, \dots, T\}$ define:
  - estimated transition kernel: $\widehat{P}_h^{(t)}(s' \mid s, a) = \langle \widehat{\phi}_h(s,a), \widehat{\mu}_h^{(t)}(s') \rangle$.
  - empirical covariance matrix: $\widehat{\Sigma}_{h,\widehat{\phi}}^{(t)} = \sum_{i=1}^n \widehat{\phi}_h(s_h^{(i,t)}, a_h^{(i,t)})\widehat{\phi}_h(s_h^{(i,t)}, a_h^{(i,t)})^\top + \lambda I$.
  - penalty term: $\widehat{b}_h^{(t)}(s_h, a_h) = \min\left\{\alpha\|\widehat{\phi}_h(s_h,a_h)\|_{(\widehat{\Sigma}_{h,\widehat{\phi}}^{(t)})^{-1}}, 1\right\}$.
  - Get policy $\widehat{\pi}_t = \operatorname{argmax}_\pi V_{\widehat{P}^{(t)}, r^t - \widehat{b}^{(t)}}$
- **Output:** $\widehat{\phi}, \widehat{P}^{(1)}, \dots, \widehat{P}^{(T)}, \widehat{\pi}_1, \dots, \widehat{\pi}_T$

Theoretical Result on Upstream Task

Definition 1 (Multi-task relative condition number). For task t and time step h, we define $C_{t,h}^*(\pi_t, \pi_t^b)$ as the relative condition number under ϕ_h^*:

$$C_{t,h}^*(\pi_t, \pi_t^b) := \sup_{x\in\mathbb{R}^d} \frac{x^\top \mathbb{E}_{(s_h,a_h)\sim(P^{(*,t)},\pi_t)}[\phi_h^*(s_h,a_h)\phi_h^*(s_h,a_h)^\top]x}{x^\top \mathbb{E}_{(s_h,a_h)\sim(P^{(*,t)},\pi_t^b)}[\phi_h^*(s_h,a_h)\phi_h^*(s_h,a_h)^\top]x}.$$

We define $C_t^* := \max_{h\in[H]} C_{t,h}^*(\pi_t, \pi_t^b)$ and $C^* := \max_{t\in[T]} C_t^*$.

***Theorem* 1.** *Under realizability assumption, with probability at least $1-\delta$, for any step $h \in [H]$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{(s_h,a_h)\sim(P^{(*,t)},\pi_t^b)}\left[\left\|\widehat{P}_h^{(t)}(\cdot \mid s_h, a_h) - P_h^{(*,t)}(\cdot \mid s_h, a_h)\right\|_{TV}\right] \le \sqrt{\frac{2\log(2|\Phi||\Psi|^T nH/\delta)}{nT}},$$

where $\widehat{\phi}, \widehat{P}^{(1)}, \dots, \widehat{P}^{(T)}$ be the output of MORL.

In addition, if we set $\alpha = \sqrt{2n\omega\zeta_n + \lambda d}$, $\lambda = cd\log(|\Phi||\Psi|^T nH/\delta)$ with $\zeta_n := \frac{2\log(2|\Phi||\Psi|^T nH/\delta)}{n}$ and c being a constant, where we assume that $\omega := \max_t \max_{s,a}(1/\pi_t^b(a \mid s)) < \infty$, then under realizability assumption, with probability at least $1-\delta$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\left[V_{P^{(*,t)},r^t}^{\pi_t} - V_{P^{(*,t)},r^t}^{\widehat{\pi}_t}\right] \le \omega\alpha dH\sqrt{\frac{C^*}{n}} + 2dH^2\sqrt{\frac{\lambda C^*}{n}} + \omega H^2\sqrt{\frac{dC^*\zeta_n}{T}} + \alpha\sqrt{\frac{d}{n}} + 2H\sqrt{\frac{\omega\zeta_n}{T}},$$

where $\{\widehat{\pi}_t\}_{t\in[T]}$ is the output of the algorithm MORL.

Connecting Upstream and Downstream tasks

- **Assumptions:** reachability of behavior policies, compact state space, smoothness of transition probabilities, and approximate linear combination.

***Lemma* 2.** *Under the above assumptions, the output $\widehat{\phi}$ of MORL is a ξ_{down}-approximate feature for MDP $\mathcal{M}^{T+1}$ where $\xi_{down} = \xi + \frac{C_L C_R \nu}{\kappa}\sqrt{\frac{2T\log(2|\Phi||\Psi|^T nH/\delta)}{n}}$, i.e. there exist a time-dependent unknown (signed) measure $\widehat{\mu}^*$ over $\mathcal{S}$ such that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have*

$$\|P_h^{(*,T+1)}(\cdot|s,a) - \langle \widehat{\phi}_h(s,a), \widehat{\mu}_h^*(\cdot)\rangle\|_{TV} \le \xi_{down}.$$

Downstream RL: Reward-free Exploration

***Theorem* 3.** *Under the above assumptions, after collecting K_{RFE} trajectories during the exploration phase, with probability at least $1-\delta$, the output of the planning phase, policy π satisfies*

$$\mathbb{E}_{s_1\sim\mu}[V_1^*(s_1,r) - V_1^\pi(s_1,r)] \le c'\sqrt{d^3H^4\log(dK_{RFE}H/\delta)/K_{RFE}} + 6H^2\xi_{down}. \quad (1)$$

If the linear combination misspecification error ξ satisfies $\widetilde{O}(\sqrt{d^3/K_{RFE}})$ and the number of trajectories in the offline dataset for each upstream task is at least $\widetilde{O}(TK_{RFE}/d^3)$, then, provided K_{RFE} is at least $O(H^4d^3\log(dH\delta^{-1}\epsilon^{-1})/\epsilon^2)$, with probability $1-\delta$, the policy π will be an ϵ-optimal policy for any given reward during the planning phase.

Algorithm	Sample Complexity	Task
FLAMBE [Agarwal et al., 2020]	$\widetilde{O}(\frac{H^{22}d^7K^9}{\epsilon^{10}})$	Single task
MOFFLE [Modi et al., 2021]	$\widetilde{O}(\frac{H^7d^{11}K^{14}}{\min\{\epsilon^2\eta,\eta^5\}})$	Single task
RAFFLE [Cheng et al., 2023]	$\widetilde{O}(\frac{H^5d^4K}{\epsilon^2})$	Single task
This work	$\widetilde{O}(\frac{H^4d^3}{\epsilon^2})$	Multi-task

Table 1. Sample complexities of different approaches to learning an ϵ-optimal policy for the reward-free RL setting with low-rank MDPs.

Downstream RL: Offline RL and Online RL

Downstream Offline Task:

Assumption 4 (Feature coverage). There exists an absolute constant κ_ρ such that for all $h \in [H]$ and $\phi_h \in \Phi_h$, $\lambda_{\min}(\mathbb{E}_\rho[\phi_h(s_h,a_h)\phi_h(s_h,a_h)^\top|s_1 = s]) \ge \kappa_\rho$.

***Theorem* 5** (Downstream offline task). *Under the above assumptions and the sample size $N_{off} \ge 40/\kappa_\rho \cdot \log(4dH/\delta)$, with probability at least $1-\delta$, the suboptimality gap of offline downstream task is at most*

$$V_{P^{(*,T+1)},r}^{\pi^*}(s) - V_{P^{(*,T+1)},r}^{\widehat{\pi}}(s) \le O\left(\kappa_\rho^{-1/2}H^2d\sqrt{\frac{\log(HdN_{off}\max(\xi_{down},1)/\delta)}{N_{off}}} + \kappa_\rho^{-1/2}H^2d^{1/2}\xi_{down}\right).$$

Downstream Online Task:

***Theorem* 6** (Downstream online task). *Let $\widetilde{\pi}$ be the uniform mixture of $\pi^1, \dots, \pi^{N_{on}}$. Under the above assumptions, with probability $1-\delta$, the suboptimality gap of online downstream task satisfies*

$$V_{P^{(*,T+1)},r}^* - V_{P^{(*,T+1)},r}^{\widetilde{\pi}} \le \widetilde{O}(H^2d^{3/2}N_{on}^{-1/2} + H^2d\xi_{down}).$$

For more details check the paper!

The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)

haque.ishfaq@mail.mcgill.ca