Off-Policy Reinforcement Learning with Delayed Rewards

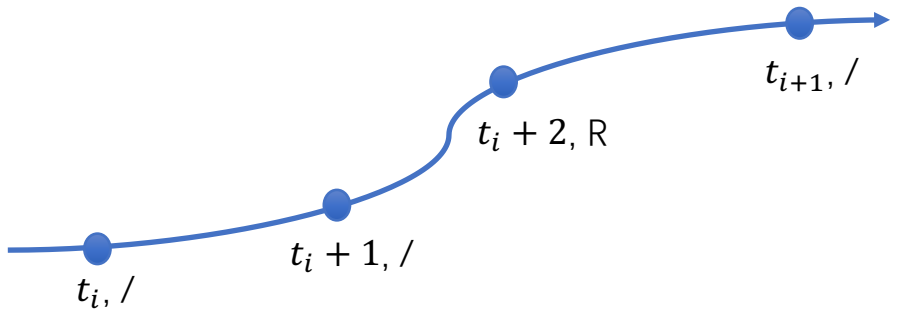
Beining Han, Zhizhou Ren, Zuofan Wu, Yuan Zhou, Jian Peng

ICML | 2022

Delayed Reward MDPs

Definition 2.1 (DRMDP). A Delayed Reward Markov Decision Process $M = (S, A, p, q_n, r, \gamma)$ is described by the following parameters.

1. The state and action spaces are S and A respectively.
2. The Markov transition function is $p(s'|s, a)$ for each $(s, a) \in S \times A$; the initial state distribution is $p(s_0)$.
3. The signal interval length is distributed according to $q_n(\cdot)$, *i.e.*, for the i-th signal interval, its length n_i is independently drawn from $q_n(\cdot)$.
4. The reward function r defines the expected reward generated for each signal interval; suppose $\tau_i = \tau_{t:t+n_i} = (s, a)_{t:t+n_i} = ((s_t, a_t), \ldots (s_{t+n-1}, a_{t+n_i-1}))$ is the state-action sequence during the i-th signal interval of length n_i, then the expected reward for this interval is $r(\tau_i)$.
5. The reward discount factor is γ.



Past-Invariant DRMDPs

Fact 2.2. *For any DRMDP, there exists an optimal policy* $\pi^* \in \Pi_\tau$. *However, there exists some DRMDP such that all of its optimal policies* $\pi^* \notin \Pi_s$.

Definition 2.3 (PI-DRMDP)**.** A Past-Invariant Delayed Reward Markov Decision Process is a DRMDP $M = (S, A, p, q_n, r, \gamma)$ whose reward function r satisfies the following Past-Invariant (PI) condition: for any two trajectory segments τ_1 and τ_2 of the same length, and for any two equal-length trajectory segments τ_1' and τ_2' such that the concatenated trajectories $\tau_a \circ \tau_b'$ are feasible under the transition dynamics p for all $a, b \in \{1, 2\}$, it holds that

$$r(\tau_1 \circ \tau_1') > r(\tau_1 \circ \tau_2') \iff r(\tau_2 \circ \tau_1') > r(\tau_2 \circ \tau_2').$$

Non-Markovian Rewards

- Normal off-policy algorithms (SAC [Haarnoja et al., 2018]) cannot handle non-Markovian rewards.
  - Biased critic estimate.
  - Fixed point ambiguity.
  - Large learning variance.

Algorithmic Framework

$$\mathcal{Q}^{\pi}(\tau_{t_i:t+1}) := \mathbb{E}_{(\tau,n)\sim\pi}\left[\sum_{j=i}^{\infty}\gamma^{t_{j+1}-t-1}r(\tau_j)\middle|\tau_{t_i:t+1}\right]$$

which is learned by minimizing the following function.

$$\mathcal{L}_{\phi} := \mathbb{E}_{D}\left[(R_t + \gamma\hat{\mathcal{Q}}_{\phi}(\tau_{t_j:t+2}) - \mathcal{Q}_{\phi}(\tau_{t_i:t+1}))^2\right]$$

Fact 3.1. *For any distribution D with non-zero measure for any $\tau_{t_i:t+1}$, $\mathcal{Q}^{\pi}(\tau_{t_i:t+1})$ is the unique fixed point of the MSE problem in Eq. (4). More specifically, when fixing $\hat{\mathcal{Q}}_{\phi}$ as the corresponding $\mathcal{Q}^{\pi}$, the solution of the MSE problem is still $\mathcal{Q}^{\pi}$.*

Algorithmic Framework

Proposition 3.3. (Policy Improvement Theorem for PI-DRMDP) *For any policy $\pi_k \in \Pi_s$, the policy iteration w.r.t. $\mathcal{Q}^{\pi_k}$, i.e., for any s_t and some feasible $\tau_{t_i:t}$,*

$$\pi_{k+1}(a_t|s_t, t-t_i) = \arg\max_a \mathcal{Q}^{\pi_k}(\tau_{t_i:t} \circ (s_t, a)),$$

produces policy $\pi_{k+1} \in \Pi_s$ such that $\forall \tau_{t_i:t+1}$, it holds that

$$\mathcal{Q}^{\pi_{k+1}}(\tau_{t_i:t+1}) \geq \mathcal{Q}^{\pi_k}(\tau_{t_i:t+1}),$$

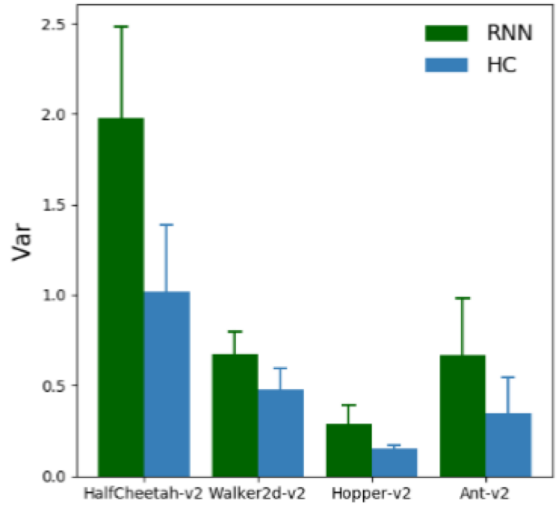
which implies that $\mathcal{J}(\pi_{k+1}) \geq \mathcal{J}(\pi_k)$.

HC-Decomposition

- However, we find vanilla implementation has unsatisfactory performance.

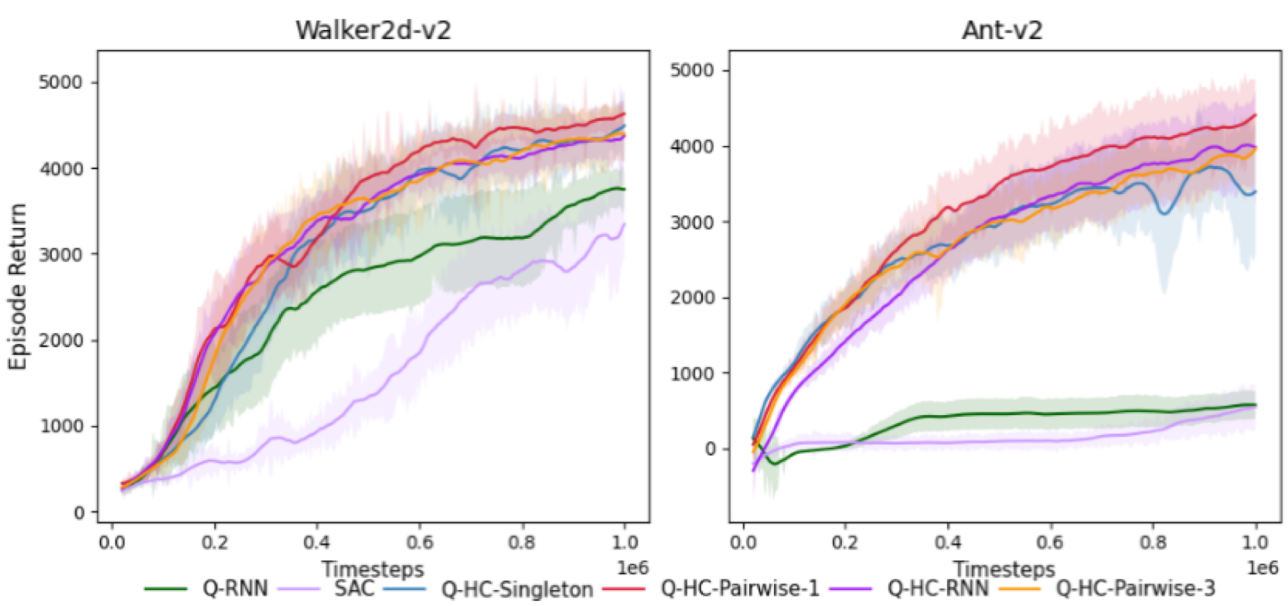
$$\mathcal{Q}_\phi(\tau_{t_i:t+1}) = H_\phi(\tau_{t_i:t}) + C_\phi(s_t, a_t),$$

- Motivated by the Markovian dynamics in DRMDPs.
- Advantages:
  1. Less variance in policy gradients.
  2. Easier optimization in value evaluation.

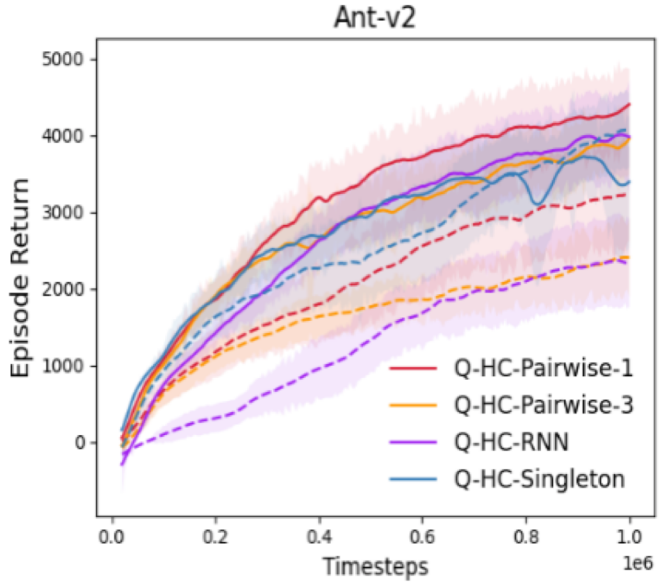


Design Evaluation

$$\mathcal{L}_\phi^{\mathrm{HC}} = \mathbb{E}_D[(R_t + \gamma(\hat{H}_\phi(\tau_{t_i:t+1}) + \hat{C}_\phi(s_{t+1}, a'_{t+1})) - (H_\phi(\tau_{t_i:t}) + C_\phi(s_t, a_t)))^2] + \lambda L_{\mathrm{reg}}(H_\phi),$$

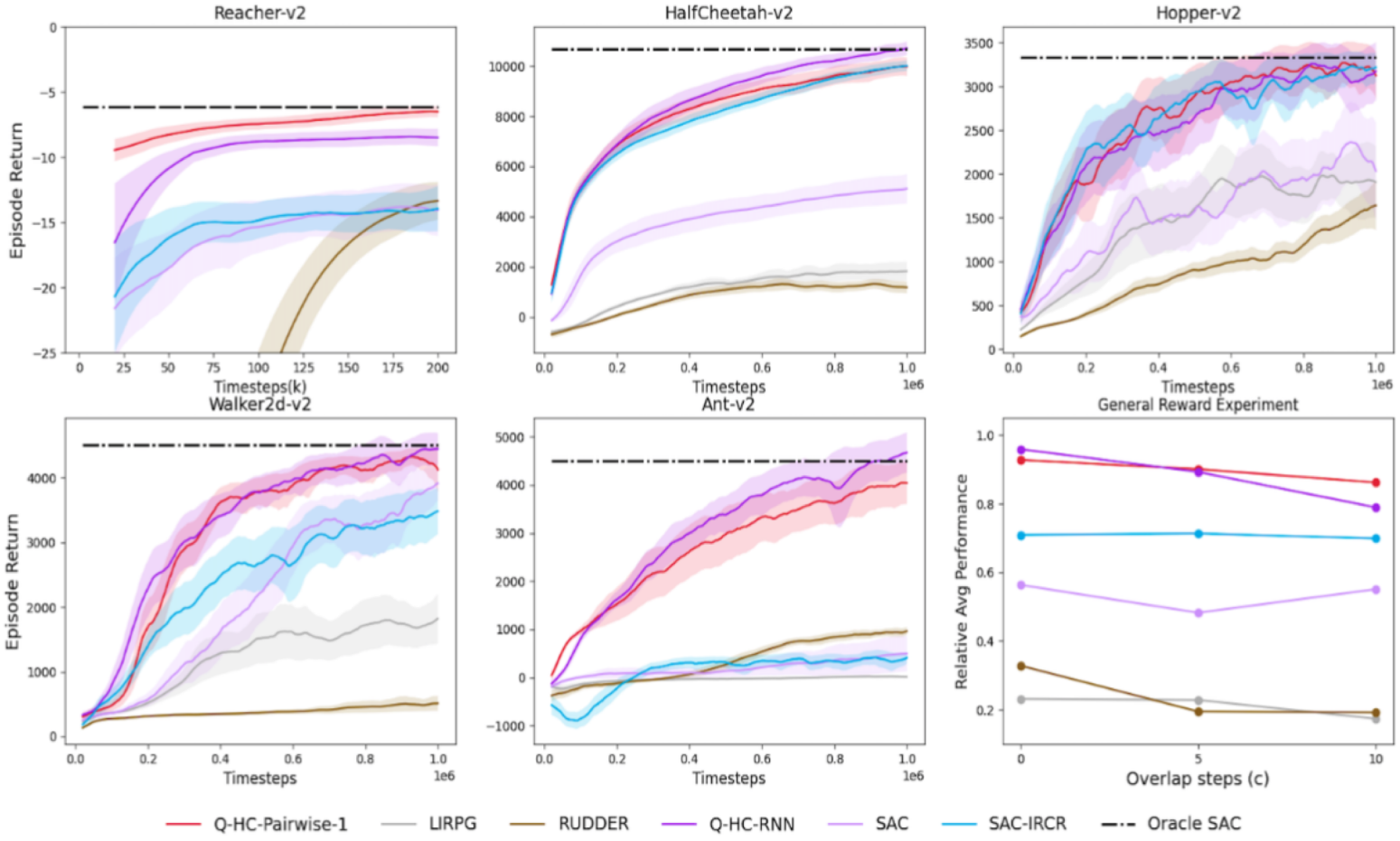


(a) Performance on High-Dimensional Tasks



(b) Ablation on Regulation

Comparative Evaluation



Reference

1. Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. InInternational conference on machine learning 2018 Jul 3 (pp. 1861-1870). PMLR.