Using Ensembles to address Bootstrapping Error in Offline Reinforcement Learning

Marco A. Gallo

29-06-2022

Outline

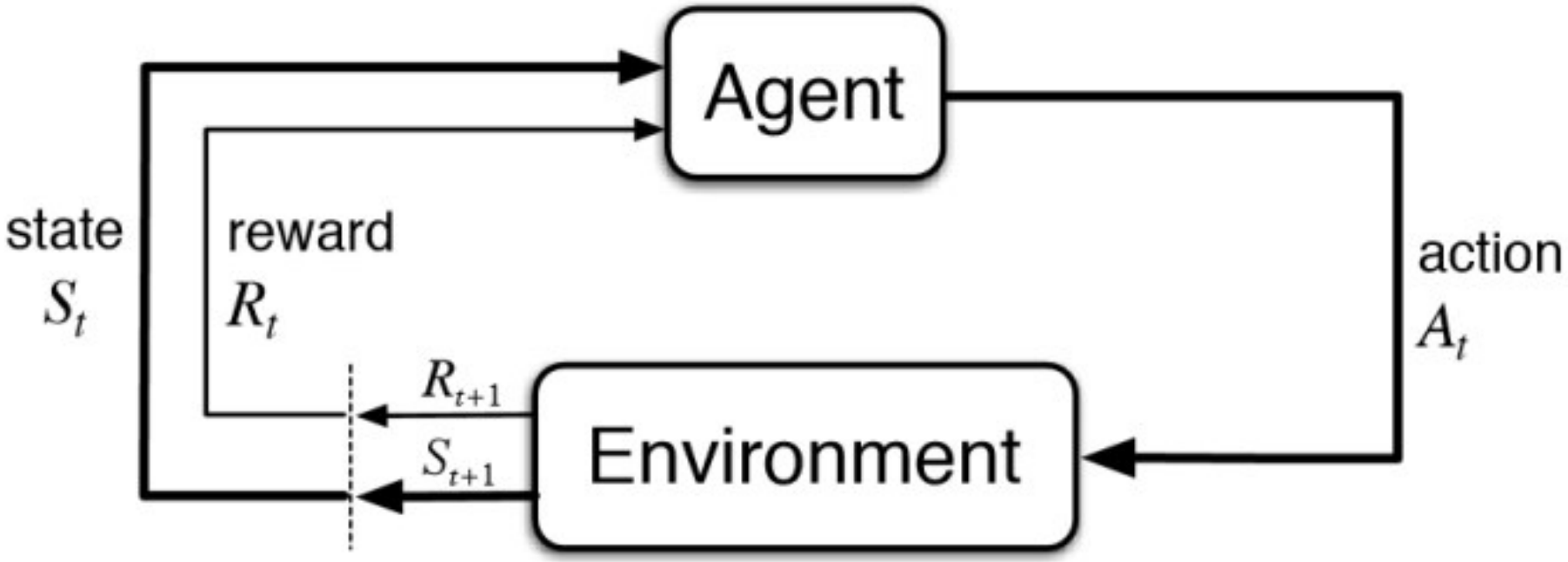
Reinforcement Learning (RL)

- An agent seeking an optimal policy $\pi(s, a)$ - a mapping from states to action probabilities ($s \in S$, $a \in A$)
- Used in sequential decision making problems modeled as Markov decision process (MDP), enriched with a reward function $R(s, a)$

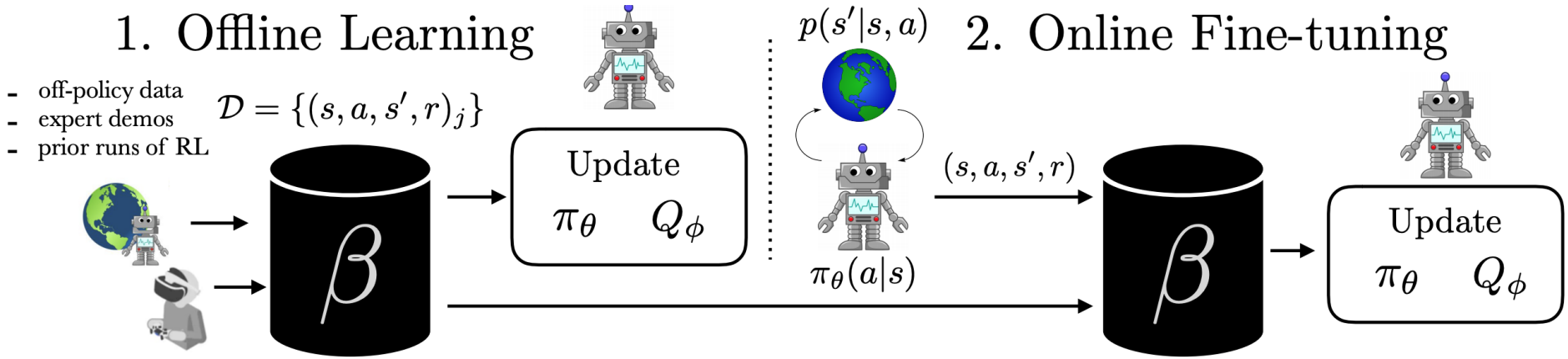
RL Elements

1. $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$ **(Expected discounted reward)**
2. $Q^{\pi}(s, a) = \mathbb{E}\left[R_t | s_t = s, a_t = a\right]$ **(State-action value function)**
3. $Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a)$ **(Optimal value function)**

Reinforcement Learning (RL) - Online



Reinforcement Learning (RL) - Offline

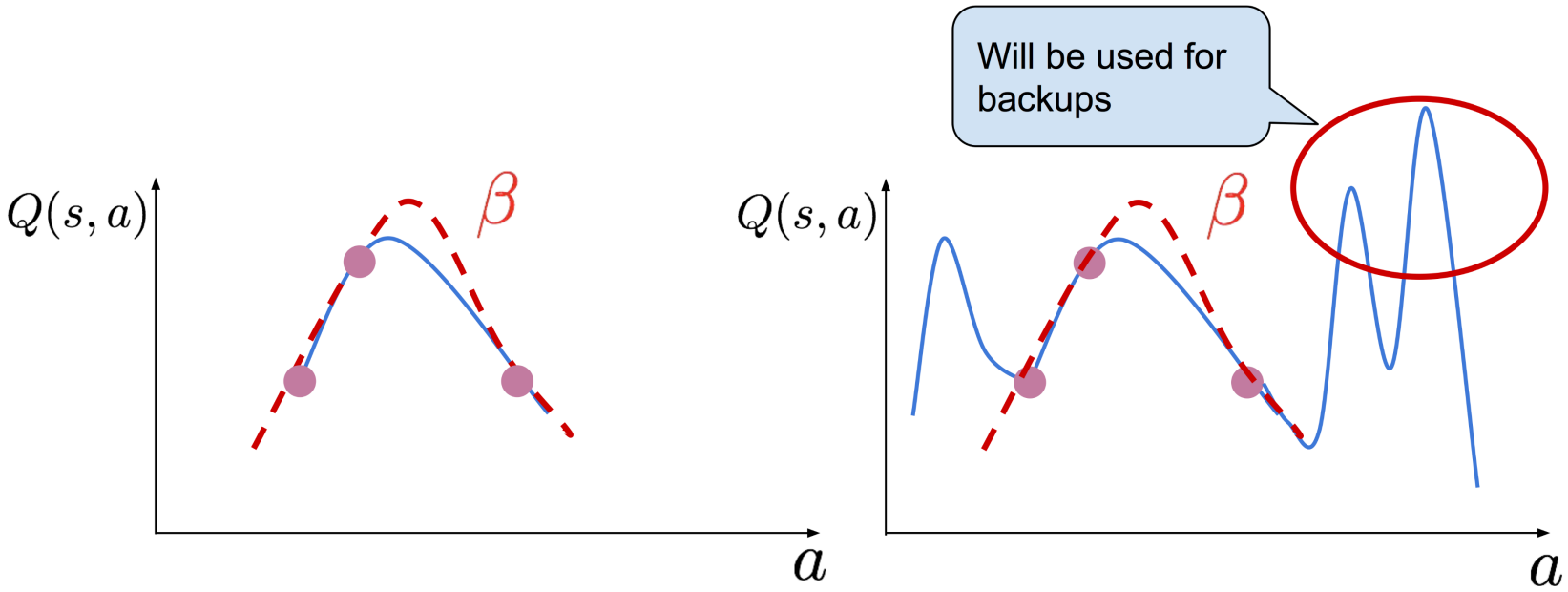


- Also called Batch Reinforcement Learning
- *Pure Batch* RL methods
- *Growing Batch* RL methods

Detrimental factors in Offline RL

- Function approximation errors in Deep RL (Neural Networks)
- Different state visitation frequencies under training and testing distributions
- Bootstrapping error (Kumar et al., 2019)

Bootstrapping Error



- Bellman optimality operator forms both the targets and the estimates for the Q-function regression
- Out-of-distribution (OOD) actions have arbitrarily wrong estimates
- Naive max over next state action pair in Bellman targets selects them, and error is propagated backwards! happens off-policy generally - but offline it cannot be corrected with ground truth values

References

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32.