# Notes on offline reinforcement learning

Compiled by D. Gueorguiev 1/14/24

## Introductory Notes and Problem Statement

The Reinforcement learning algorithms provide a fundamentally *online* learning paradigm which is one of the biggest obstacles to their adoption in various use cases. The process of reinforcement learning involves iteratively collecting experience by interacting with the environment, typically with the latest learned policy, and the using that experience to improve the policy. In many settings, the online interaction is not really useful or even applicable, because data collection is difficult to guarantee in reliable manner. Furthermore, even in domains where online interaction is feasible, we might still prefer to utilize previously collected data – for example if the domain is with large number of dimensions and complex search space so that effective generalization would require large datasets.

The success of machine learning across a range of problems can be partially attributed to the advent of scalable data-driven learning methods which improve the generalization quality as they are trained with more data. Online reinforcement learning is difficult to reconcile with this conclusion. For low-dimensional and linear models using reinforcement learning usually this is not a significant problem. However, the issue becomes obvious when deep neural nets are incorporated in the reinforcement learning algorithm. So the pertinent question is: can data-driven learning be incorporated with the reinforcement learning objectives such that well-chosen offline data is used to drive the learning process?

Environment

update

(b) off-policy reinforcement learning

buffer

Figure 1a: In online reinforcement learning, the policy is updated with streaming data online collected by the policy itself.

Environment

update

(a) online reinforcement learning

Figure 1b: In off-policy setting, the agent’s experience is appended to the data buffer (also called replay buffer) , and each new policy collects additional data, such that is composed of samples from , and all of this data is used to train an updated new policy .

In contrast, offline reinforcement learning employs a dataset collected by some (potentially unknown) behavior policy . The dataset is collected once, and is not altered during training, which makes it feasible to use large previous collected datasets. The training process does not interact with the MDP at all, and the policy is only deployed after being fully trained.

Figure 1c: In offline reinforcement learning a dataset is collected by some (potentially unknown) behavior policy . The dataset is collected once, and is not altered during training, which makes it feasible to use large previously collected datasets. The training process does not interact with the MDP at all, and the policy is only deployed after being fully trained.

learn

Environment

(c) offline reinforcement learning

buffer

Environment

data collected **once** with **any** policy

training phase

A data-driven offline reinforcement learning as pictured in Figure 1c poses significant algorithmic challenges.

## References

[Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems, Sergey Levine, Aviral Kumar, George Tucker, Justin Fu, 2020](https://github.com/dimitarpg13/self_supervised_learning/blob/main/literature/OfflineReinforcementLearningTutorialReviewLevine2020.pdf)