Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism

Paria Rashidinejad*, Banghua Zhu, Cong Ma, Jiantao Jiao, Stuart Russell

Berkeley Artificial Intelligence Research
University of California, Berkeley

*paria.rashidinejad@berkeley.edu

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

The paper in 30 seconds

Goal. Learning to make decisions from an **arbitrary** and fixed previously-collected dataset without active data collection. The paper addresses two questions:

Question 1. Can we formulate our problem in a way to capture **any dataset composition**?

- Yes! By considering the deviation of data distribution and distribution of expert policy.

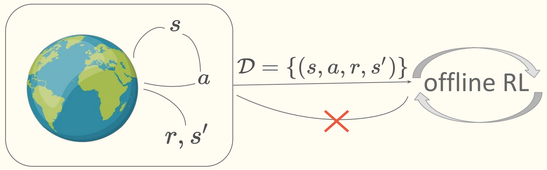
Question 2. Is there an algorithm that **optimally learns** to make decisions, regardless of the unknown dataset composition?

- (Almost) yes! **Pessimism in the face of uncertainty** achieves near-optimal performance, bridging the expert (imitation learning) and uniform coverage (vanilla offline RL) regimes.

Offline Reinforcement Learning

Agent's goal. Achieve competence in a task

- Using a previously-collected dataset
- Without access to further data collection



Advantages.

- Exploits large existing datasets
- Avoids policy deployment (costly and dangerous)

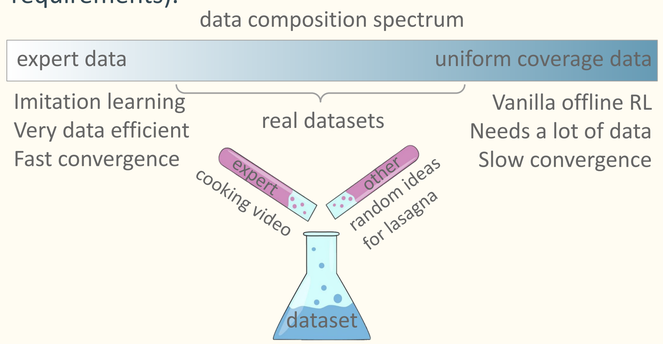
Offline dataset. A set of (s, a, r, s') collected from an unknown environment

Task: Cooking lasagna

- s : current state ⟶ sauteed onion in the pan
- a : current action ⟶ add meat
- r : reward ⟶ -1 if burnt
- s' : next state ⟶ onion + meat in the pan

Accommodating Different Data Compositions

Challenge. Existing methods require either expert or uniform coverage data compositions (strong requirements).



Main Questions

Question 1. Does an offline RL framework exist that captures the entire data composition?

- Yes, described by the ratio of target policy distribution over data distribution (weakest concentrability definition) denoted by C*.

Question 2. With this framework, is there an algorithm that handles any *possibly unknown* data composition?

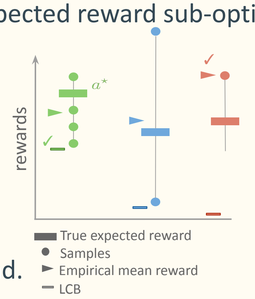
- Yes, we consider a pessimism-based algorithm that constructs lower confidence bounds on policy quality.
- We prove that given finite data, the algorithm is near-optimal in information-theoretic minimax sense.

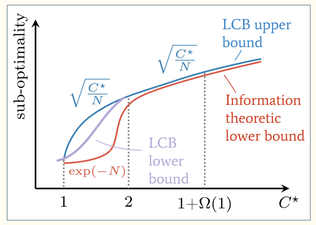
Warm-Up: Multi-Armed Bandits

Setting. Receive stochastic rewards on arms
Goal. Maximize the expected reward sub-optimality

Challenge. Arm with the largest expected reward fails.

Solution. Pick the arm with maximum lower confidence bound.



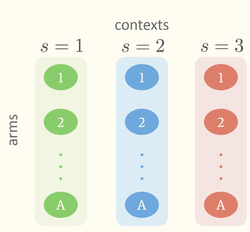


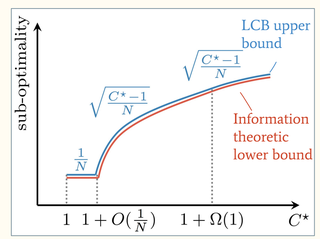
Contextual Bandits

Setting. Several contexts each with multiple arms
Goal. Minimize expected reward sub-optimality

Challenge. Arm with the largest expected reward fails.

Solution. Pick the arm with maximum LCB in each context.





Markov Decision Processes

Setting. A stochastic dynamic environment with dataset $\mathcal{D} = \{(s, a, r, s')\}$
Goal. Maximize the value (expected cumulative rewards) of the learned policy.

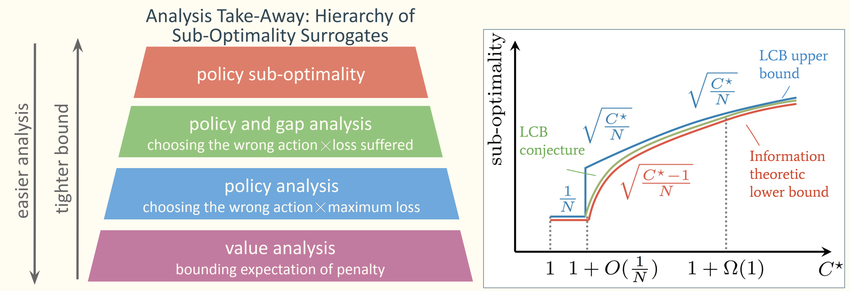
Approach. We construct lower confidence bounds (LCB) of values by subtracting a penalty from rewards, which captures confidence intervals.

Value-iteration update:

$$Q(s,a) \leftarrow r(s,a) + \gamma P_{s,a} \cdot V, \qquad \text{for all } (s,a),$$
$$V(s) \leftarrow \max_a Q(s,a), \qquad \text{for all } s.$$

Similar to LCB in bandits, subtract penalty to account for the fluctuations in reward and transition estimates:

$$Q(s,a) \leftarrow \hat{r}(s,a) - b(s,a) + \gamma \hat{P}_{s,a} \cdot V, \qquad \text{for all } (s,a),$$
$$V(s) \leftarrow \max_a Q(s,a), \qquad \text{for all } s.$$



Summary

- Proposed a new framework for studying offline learning problems that captures the entire data composition range.
- Proved that regardless of data composition, *pessimism* achieves near-optimal performance.
- Proved that in case the dataset covers expert actions, pessimism achieves fast convergence rate analogous to that of imitation learning, without having any knowledge of dataset composition.

See paper: https://arxiv.org/abs/2103.12021