DeeepAveragers: Offline Reinforcement Learning By Solving Derived Non-Parametric MDPs

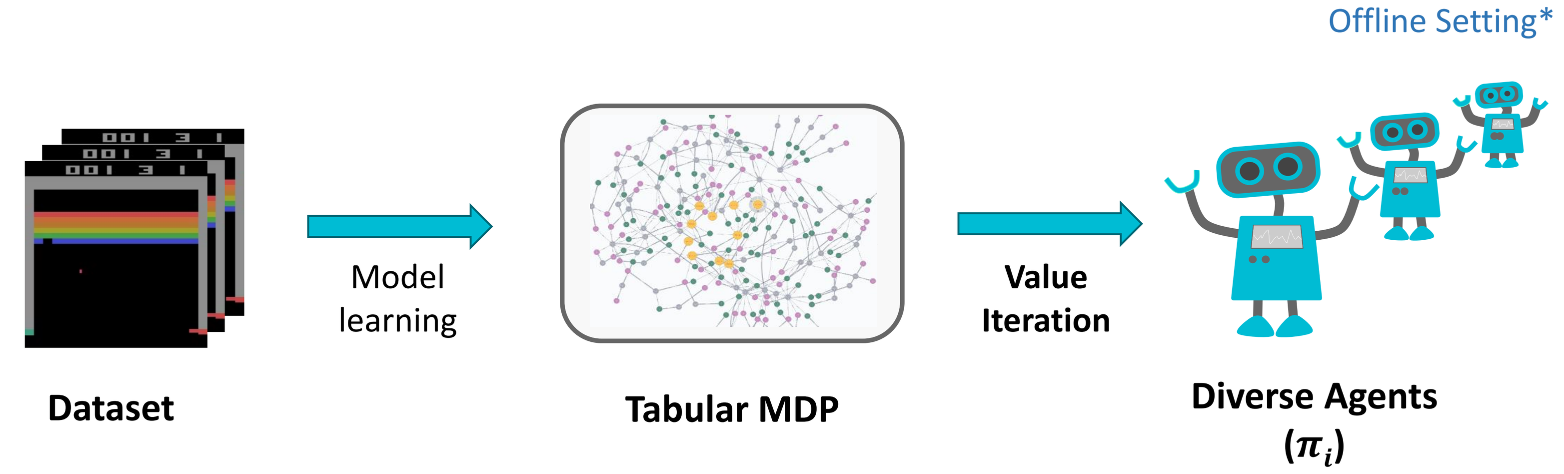
Oregon State University

Aayam Shrestha, Stefan Lee, Prasad Tadepalli, Alan Fern

1. Motivation and Proposal:

The promise of Model Based Reinforcement Learning:

- Learn an environment model once. (Learn)
- Optimize for different goals and behaviors. (Plan)

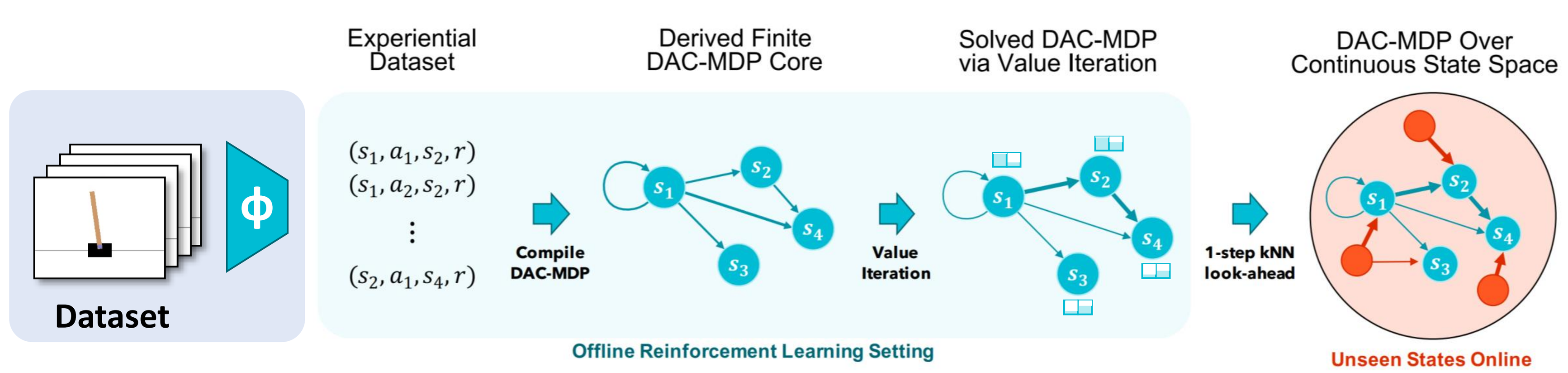


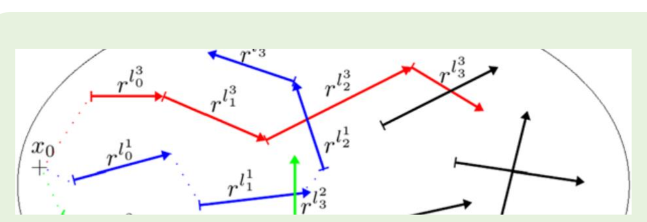
- Fast to adapt and Easy to Debug.
- Theoretical Guarantees.
- Different reward structures
- Safety constraints.

2. Approach:

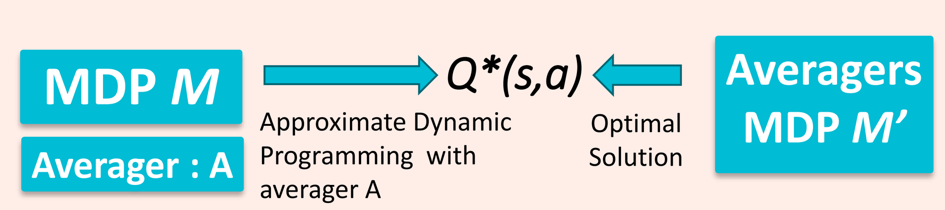
- Compile a finite MDP *M*.
- Solve MDP *M* using value iteration (optimized for GPU)
- Calculate Q values for unseen state actions via one step lookup.

Deep Averagers with Costs (DAC) - MDPs





Fonteneau F. et al. : **MFMC**



Gordon, G. (1995): **Averagers**

3. DAC MDP Formulation:

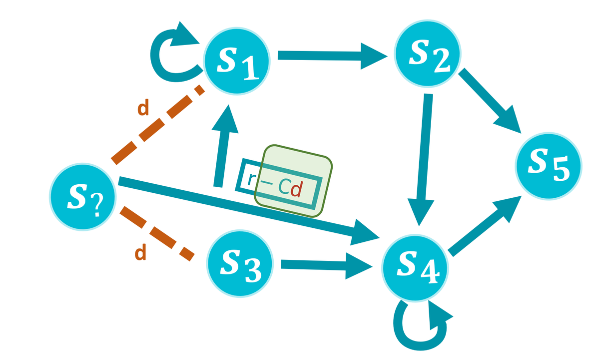
- Model transition and reward by a simple k-nearest neighbor regression
- Additional cost to ensure pessimism in sparse data regions.

$$\tilde{T}(s,a,s') = \frac{1}{k}\sum_{i\in kNN(s,a)} I[s'=s'_i]$$

$$\tilde{R}(s,a) = \frac{1}{k}\sum_{i\in kNN(s,a)} r_i - C\, d(s,a,s_i,a_i)$$

$kNN(s,a) \rightarrow$ Set of indices of the k nearest neighbors to (s,a) in the Dataset

$d(s,a,s_i,a_i) \rightarrow$ The distance between state action tuples (s,a) and (s_i, a_i, s'_i, r_i)

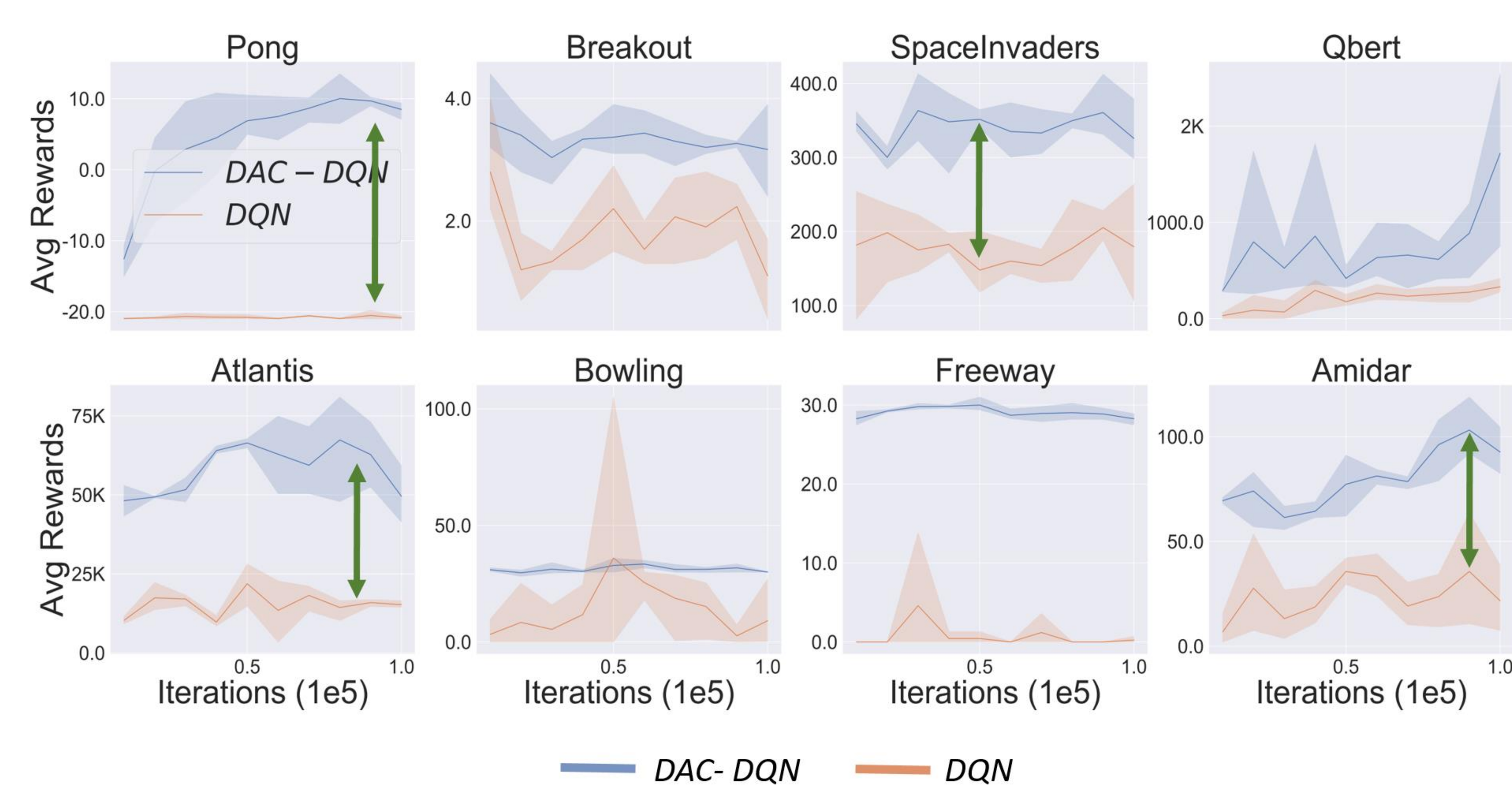


Fill in unknown state-action pairs (With discounted rewards)

4. Experimental Results:

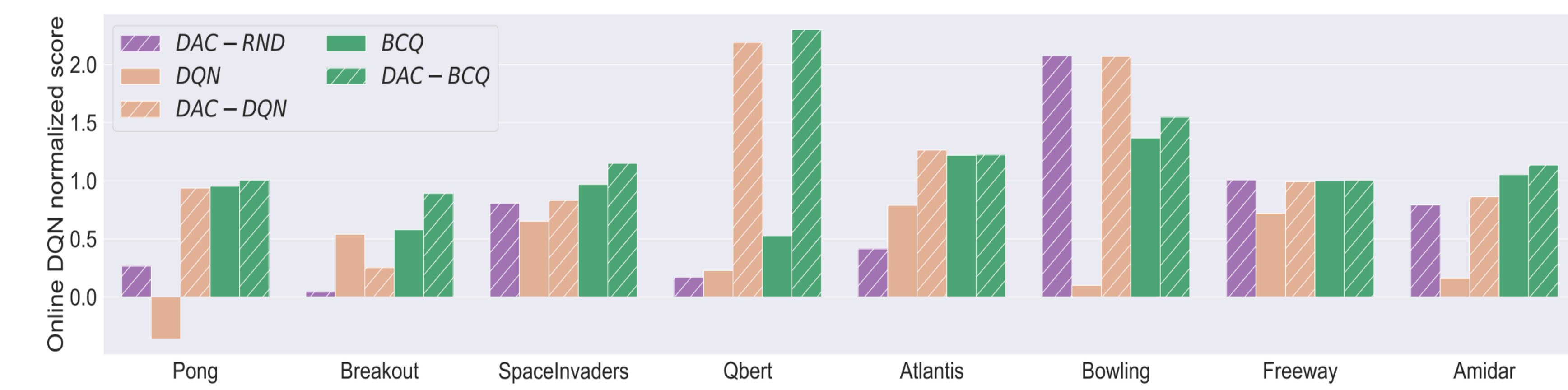
[E1] We test our approach on stochastic Atari Domain for small dataset size of 100k.

- DQN or BCQ is frozen at each evaluation point
- DAC-MDPs are derived from the same DQN representation.



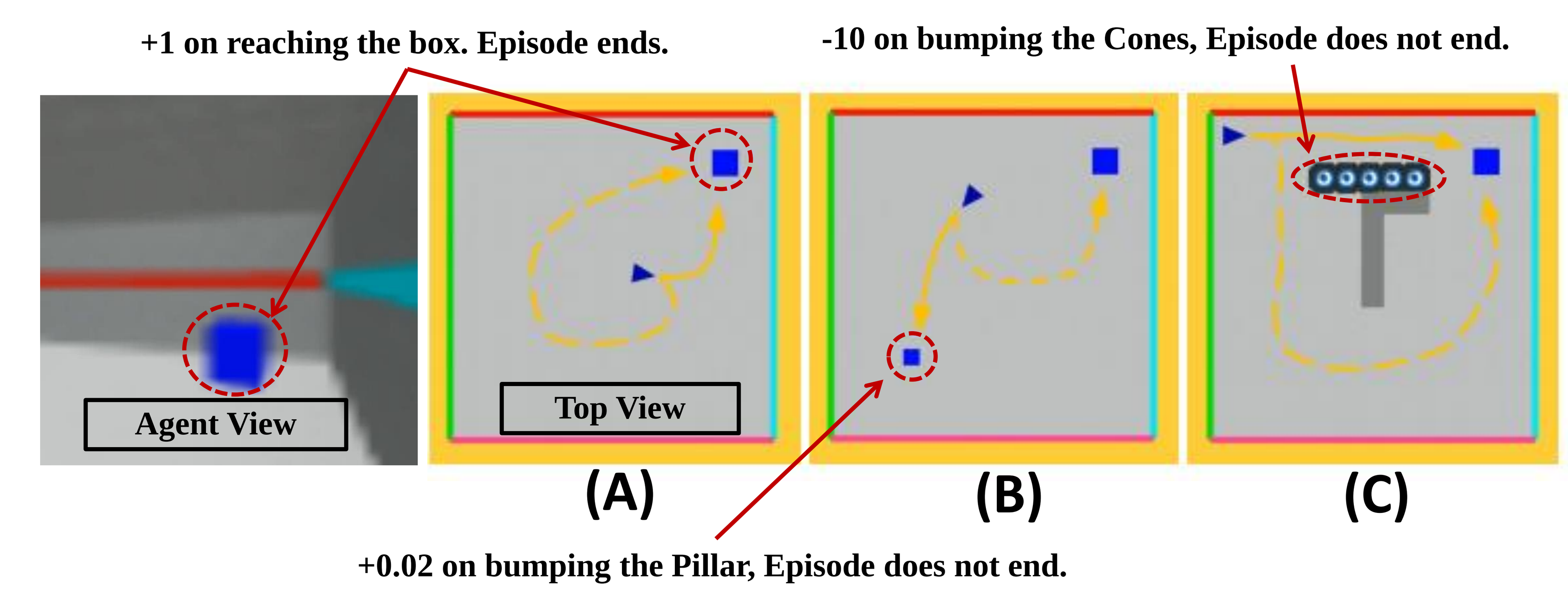
DAC-DQN clearly finds better policies than DQN. A similar trend though not as drastic was found BCQ algorithm as well.

[E2] We also perform similar experiments for 2.5M dataset across representations from random projection, DQN and BCQ. The approach can scale and outperform the baselines.



[E3] We show the flexibility of our approach on 3D navigation domain.

A. Adaptability: Optimal policy (solid), Left action penalized policy (dotted)
B. Planning Horizon: Short-term planning(solid), Long-term planning. (dotted)
C. Robustness: Optimal policy (solid), Safe policy (dotted)



5. Summary:

- Non-parametric MBRL for offline RL. (With theoretical guarantees)
- Scales to medium sized Atari games.
  - Uses GPU optimized VI Solver
- Flexibility for zero-shot learning on different auxiliary tasks.
  - Adaptability
  - Planning horizons
  - Robust Behavior

First to show scalability on

2.5 M Dataset | Stochastic Domain | Explicit Model | Exact Planning

Paper ! Code !