DS 598
Introduction to RL

Xuezhou Zhang

Team Presentations

	Team 1	Team Members			Team 2	Team Members		
03/26	Team Zero	Mao Mao	Haotian Shangguan	Ziye Chen	Team Go	Zijian Guo	Yichen Song	Zhengyang Shan
03/28	Team Lux	Seunghwan Hyun	Osama Dabbousi	Zou(Zoey) Yang	Team Carbon	Bargav Jagatha	Akshat G	Mounika
04/02	Team Gamma	Wai Yuen Cheng	Andy Yang	Tariq Georges	Team Star (milesliiii)		Chenjia Li	YuCheng
04/04	Team S	Sahana Kowshik	Srishti Jain	Ruoxi Jin	Team Best	Minfeng Qian	Han Li	Qiji Zheng
04/09	Team Q (3/26)	Xavier Thomas	Shivacharan oruganti		Team Terrier Threat	Jack Campbell	Carmen Pelayo	Wilson Zhang
04/11	Team ZGL	Jasmine Dong	Yu Liang	Shuhan Wang	Team Rocket	Tejaswini S	Shreyas S	Abhaya Shukla

PPO with Action Masking

Maskable PPO

Implementation of invalid action masking for the Proximal Policy Optimization (PPO) algorithm. Other than adding support for action masking, the behavior is the same as in SB3's core PPO algorithm.

A Closer Look at Invalid Action Masking in Policy Gradient Algorithms

Shengyi Huang and **Santiago Ontañón** *
College of Computing & Informatics, Drexel University
Philadelphia, PA 19104
{sh3397,so367}@drexel.edu

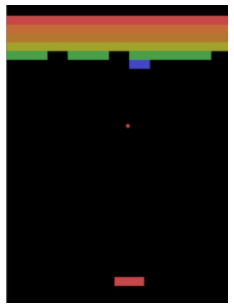
Recap: Exploration in online RL

- The **Optimism in the Face of Uncertainty (OFU)** Principle.

- Heuristic methods that attempt to mimic UCB in deep RL:
  1. Psuedo-count based reward bonus: distribution estimation, hashmap-based counts
  2. Uncertainty estimation-based reward bonus: RND
  3. Direct exploration: Go-Explore

Chapter 9: Offline RL

Offline RL -- No online Exploration

$$\left(\text{  }, \text{  } \right)_{1:M}$$

- Given a dataset of transition $D = \{(s_t, a_t, s'_t, r_t)\}_{t=1:T}$.
- Find the “best possible” policy π_θ.

Why offline RL?

1. Online RL sucks at the moment: very bad data efficiency in practice. Some people just gave up.

Human Pro Player:

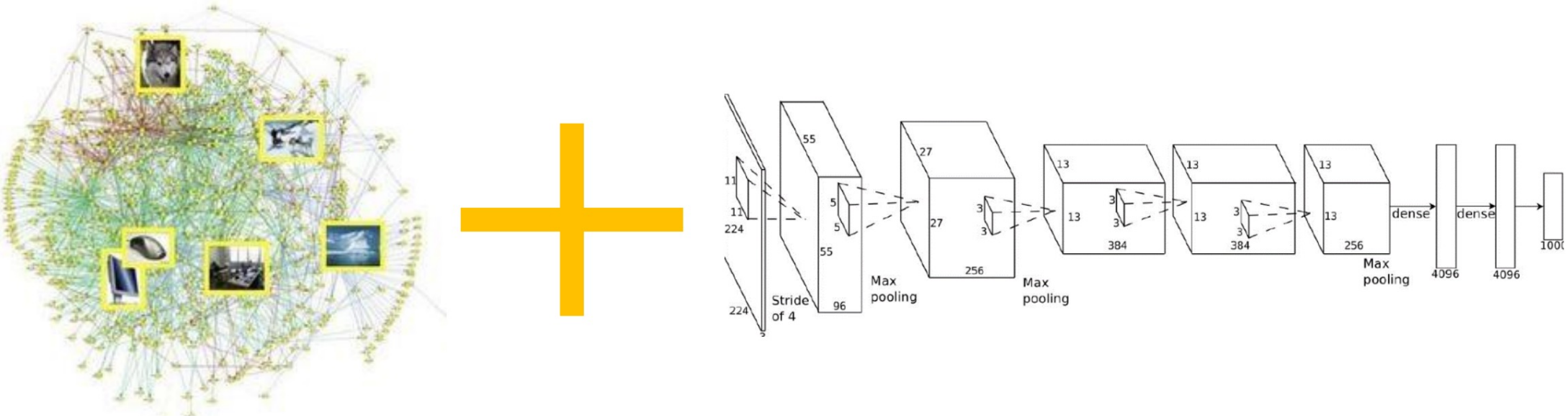
~ 50,000 games

AlphaZero:

44,000,000 games

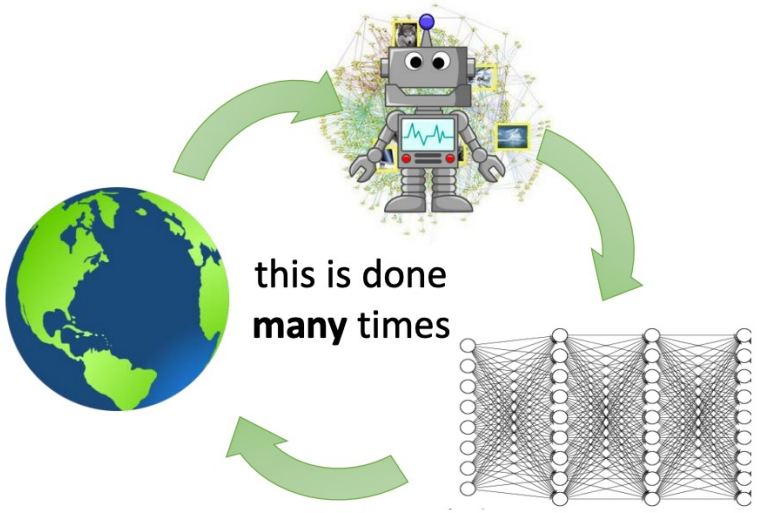
Why offline RL?

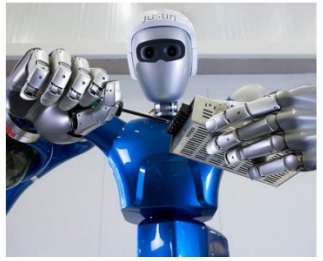
2. The success story in supervised learning: big data + big model. Maybe we can replicate that success in RL.



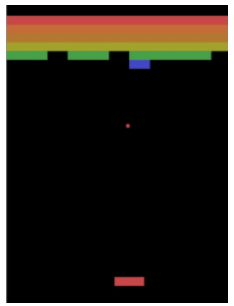
Why offline RL?

3. In many applications, offline data are abundant, while online experimentation is risky and maybe even illegal.





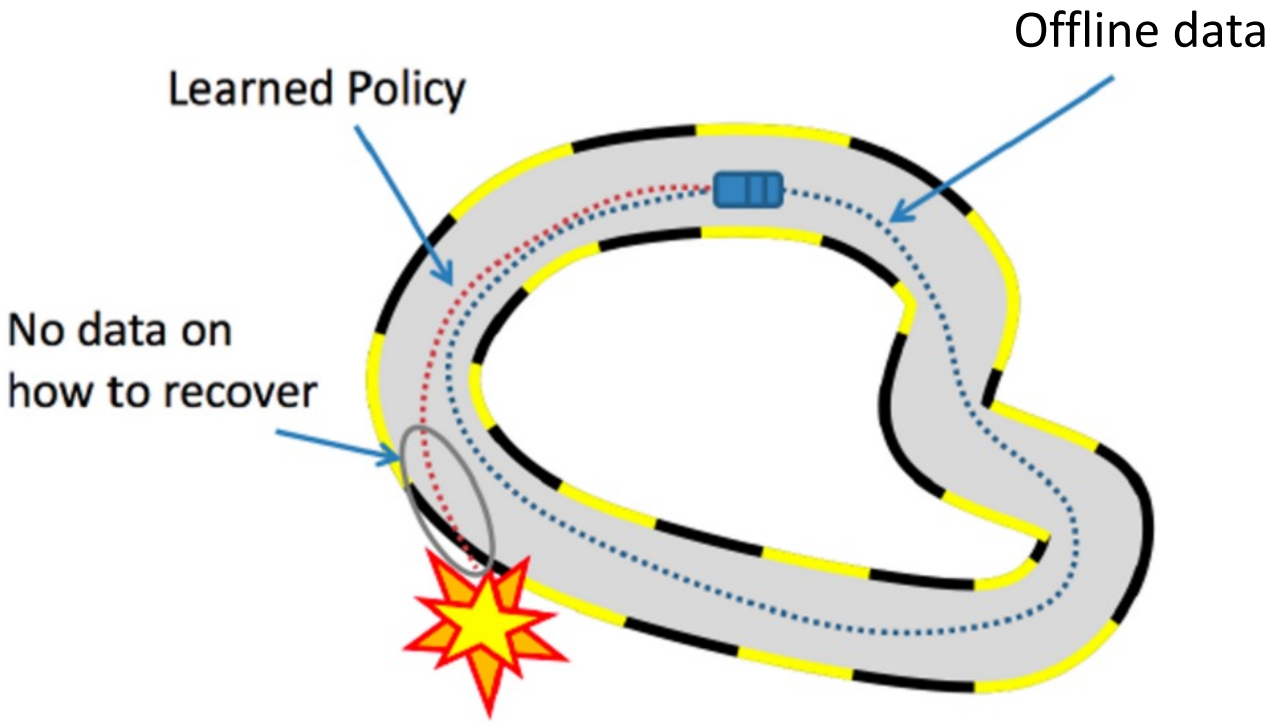
Offline RL -- No online Exploration

$$\left(\text{  }, \text{  } \right)_{1:M}$$

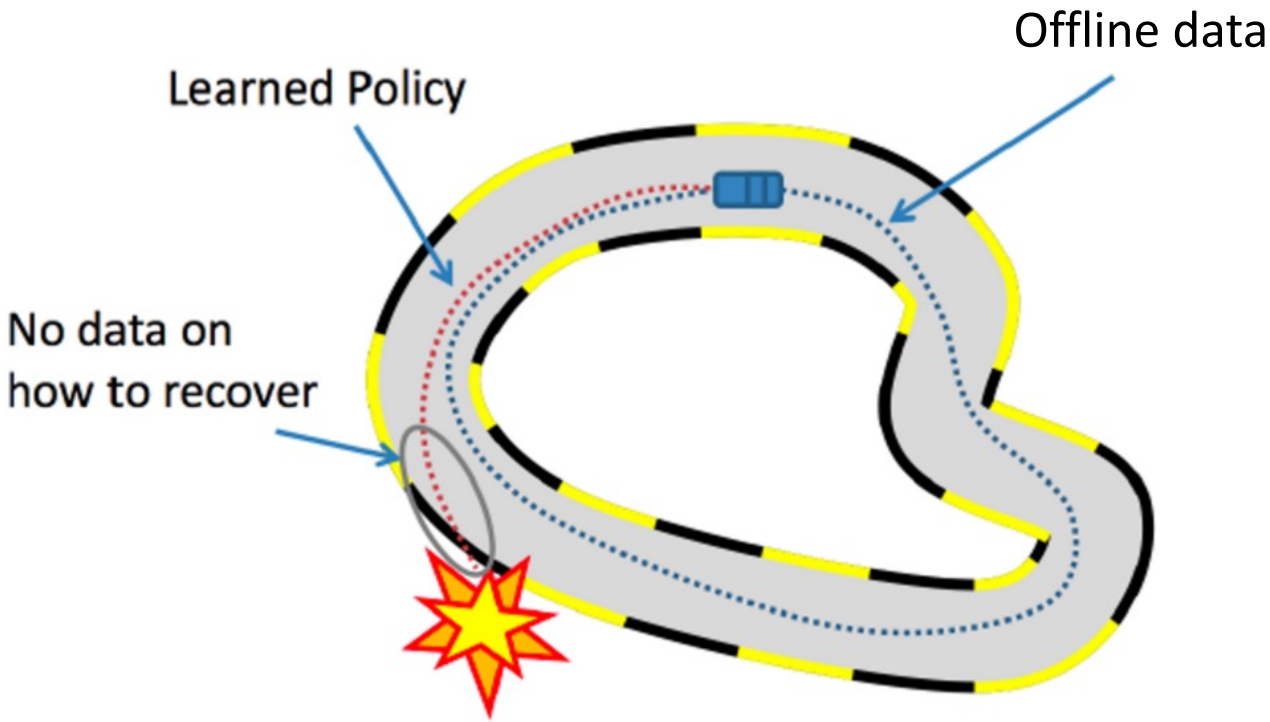
- Given a dataset of transition $D = \{(s_t, a_t, s'_t, r_t)\}_{t=1:T}$.
- Find the "best possible" policy π_θ.

- What's the difficulty?

1. The Distribution Shift problem in Offline RL

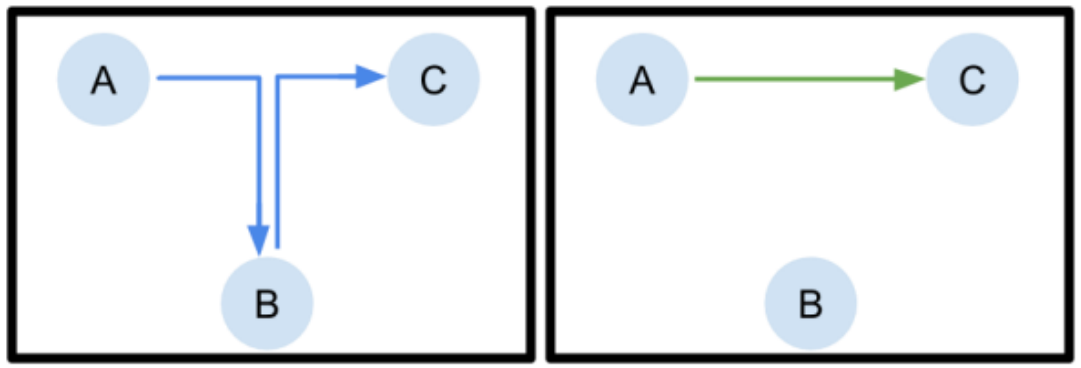


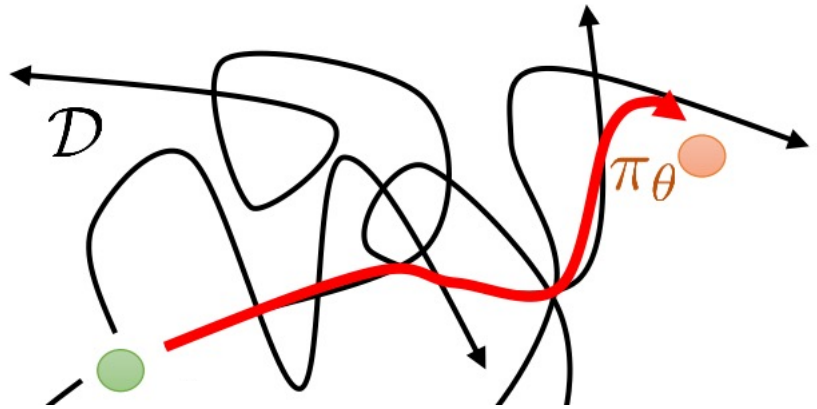
2. Unable to verify the quality of a policy



3. Demonstration can be suboptimal

Why can it still work? “Stitching” partially good trajectories.





Warmup: Offline RL with full data coverage

- Setting:

1. Infinite horizon Discounted MDPs $\gamma \in (0,1)$

2. A given offline distribution $\nu \in \Delta(S \times A)$ from which we sample offline data

3. Function class $\mathcal{F} = \{f : S \times A \mapsto [0,1/(1-\gamma)]\}$

Warmup: Offline RL with full data coverage

- Recall the Bellman Operator:

$$(\mathcal{T}f)(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s'|s,a)}\left[\max_{a'} f(s',a')\right]$$

Warmup: Offline RL with full data coverage

- Assumptions:

1. offline distribution ν has full coverage (i.e., diverse):

$$\max_{\pi} \max_{s,a} \frac{d^{\pi}(s,a)}{\nu(s,a)} \leq C < \infty$$

2. Small inherent Bellman error, i.e., near Bellman Completion (note it's averaged over ν):

$$\max_{g \in \mathcal{F}} \min_{f \in \mathcal{F}} \mathbb{E}_{s,a \sim \nu} \left(f(s,a) - \mathcal{T} g(s,a) \right)^2 \leq \epsilon_{approx,\nu}$$

The Fitted Q Iteration algorithm (FQI)

1. offline data points obtained from ν:

$$\mathcal{D} = \{s, a, r, s'\}, \quad (s, a) \sim \nu, r = r(s, a), s' \sim P(\cdot \,|\, s, a)$$

2. Initialize $f_0 \in \mathcal{F}$, and iterate:

$$f_{t+1} = \arg\min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s, a) - r - \gamma \max_{a'} f_t(s', a') \right)^2$$

3. After K iterations, return $\pi(s) = \arg\max_a f_K(s, a), \forall s$

(Note: the algorithmic idea here is similar to DQNs [Deepmind 15]

Theoretical Guarantee

Theorem

Theorem: Fix iteration number K, w/ probability at least $1-\delta$,

$$V^{\star} - V^{\pi} \leq O\left(\frac{1}{(1-\gamma)^4}\sqrt{\frac{C\ln(|\mathcal{F}|K/\delta)}{N}} + \frac{1}{(1-\gamma)^3}\sqrt{C\epsilon_{approx,\nu}}\right) + \frac{2\gamma^K}{(1-\gamma)^2}$$

Statistical error related to regression

Inherent Bellman error

VI-style Convergence rate

A proof sketch

$$f_{t+1} = \arg\min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

$$y := r(s,a) + \gamma \max_{a'} f_t(s',a')$$

Bayes optimal: $\underbrace{r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a} f_t(s',a')}_{(\mathcal{T} f_t)(s,a)}$

1. Near Bellman completion means regression target $\mathcal{T} f_t$ nearly belongs to $\mathcal{F}$

$$\mathbb{E}_{s,a \sim \nu} \left(f_{t+1}(s,a) - \mathcal{T} f_t(s,a) \right)^2 \approx \frac{1}{N} + \epsilon_{approx,\nu}$$

2. $f_{t+1} \approx \mathcal{T} f_t$ (under the diverse ν), i.e., it's like Value Iteration, we could hope for a convergence

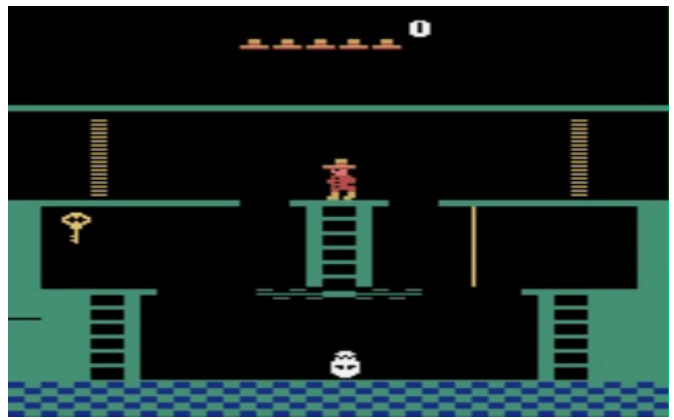
Deep RL Implementation

- Exactly same as DQN, except no more collection of new experience!

- Does it work in practice?

Google Research

Offline RL on Atari 2600



Train 5 DQN (Nature) agents on 60 Atari games
with sticky actions for 200 million frames.

Google Research

Offline RL on Atari 2600

Save all *(observation, action, next observation, reward) tuples* encountered to **DQN Replay Dataset**. Total of 300 datasets, 5 per game.

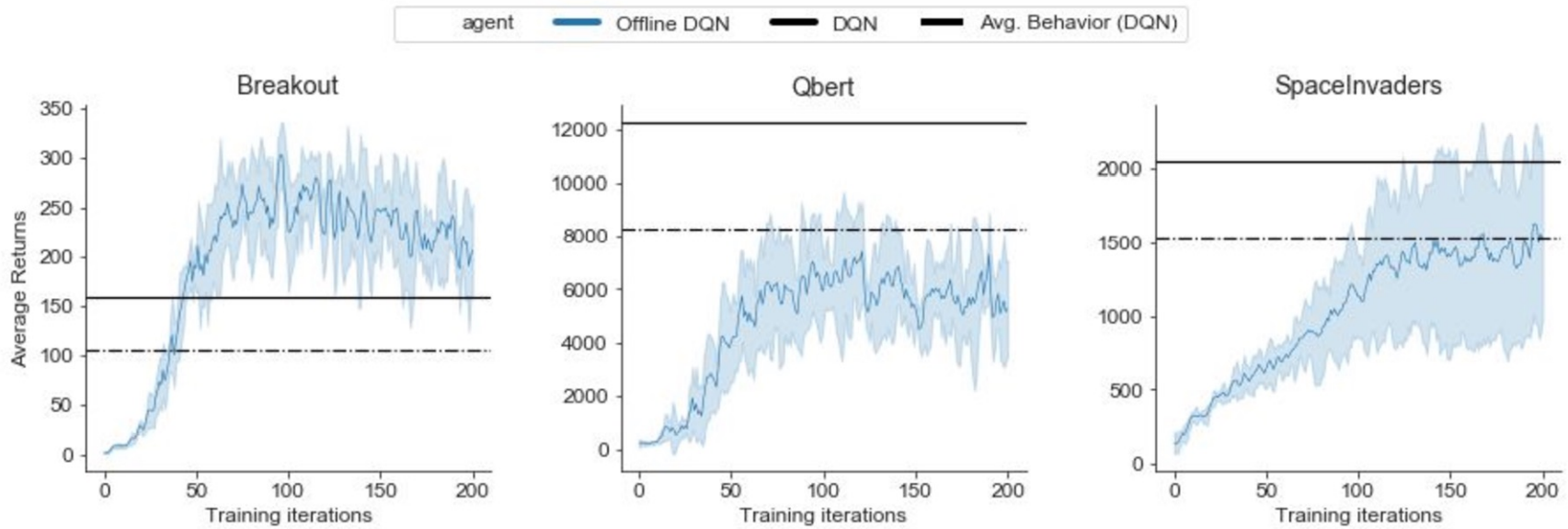
Google Research

Offline RL on Atari 2600

Train offline agents using DQN Replay Dataset
without any further environment interactions.

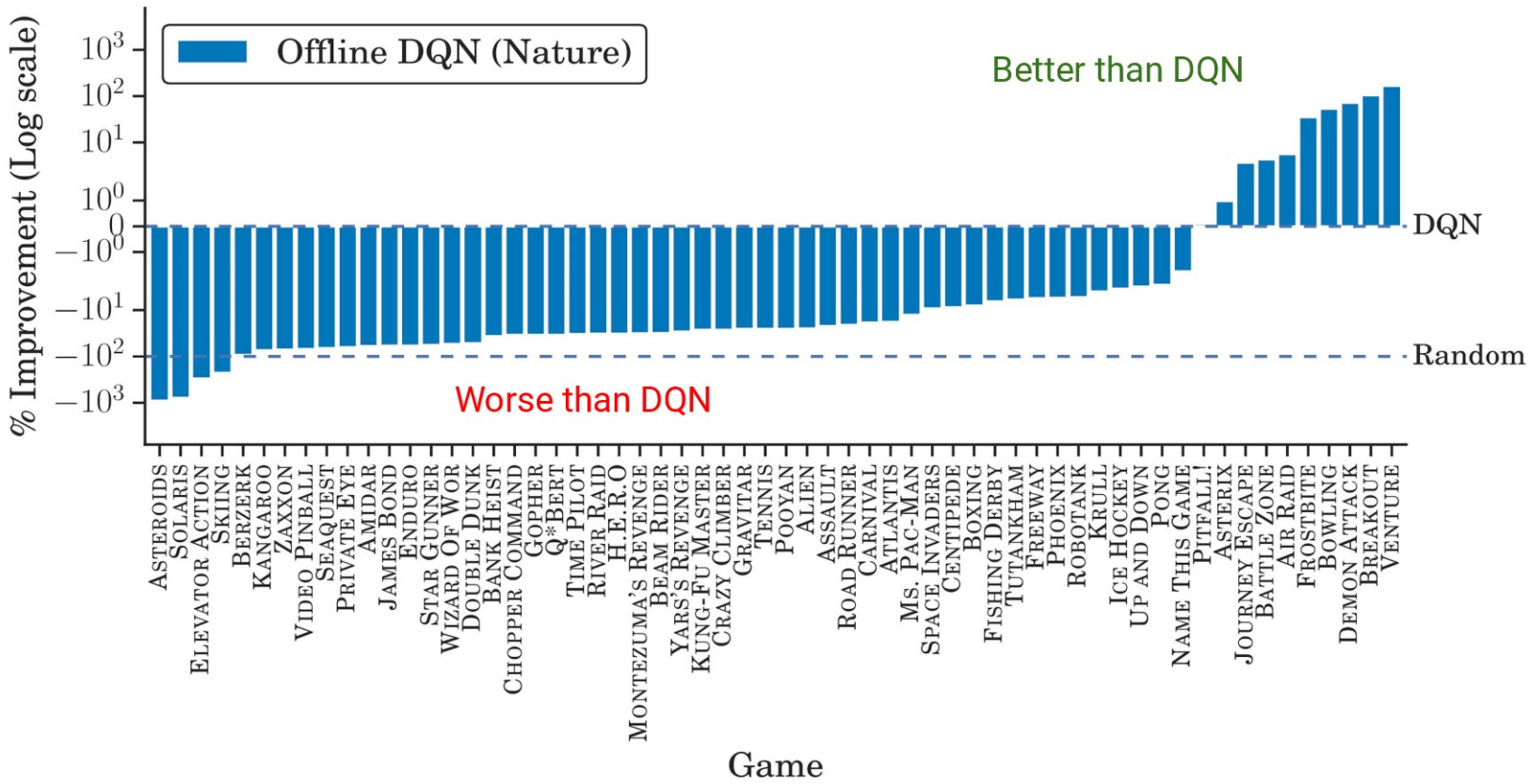
Offline DQN on DQN Replay Dataset

Google Research



Google Research

Does Offline DQN work?



Adapting Online RL algorithms to Offline

- Any off-policy RL algorithm (using replay buffer) can be adapted to the offline RL setting.
- Ironically, they don’t work well for offline RL for the same reason they don’t work well in online exploration!
- We will continue next time.