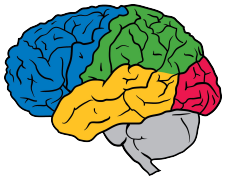
Rishabh Agarwal, Dale Schuurmans, Mohammad Norouzi

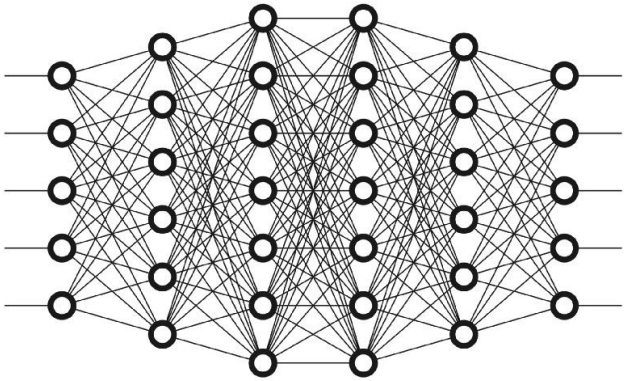
How I Learned To Stop Worrying And Love Offline RL

Or .. Striving for Simplicity in Off-Policy Deep RL



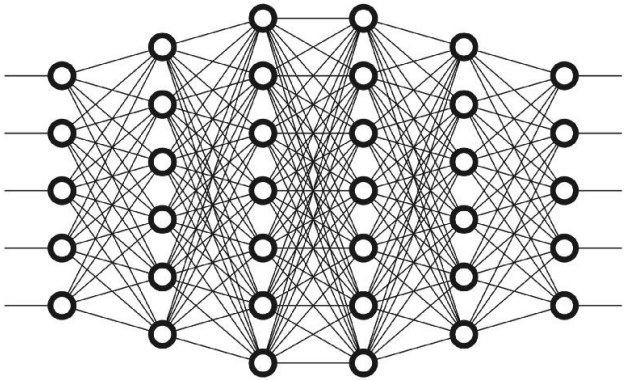
Google Research

A Lesson from Deep Learning



Powerful Function Approximators

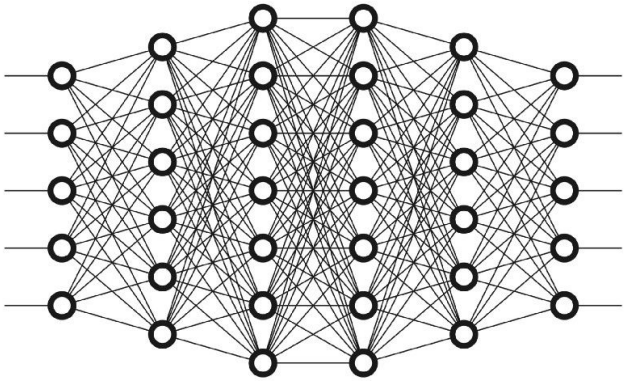
A Lesson from Deep Learning



Powerful Function Approximators

Large & Diverse Datasets

A Lesson from Deep Learning

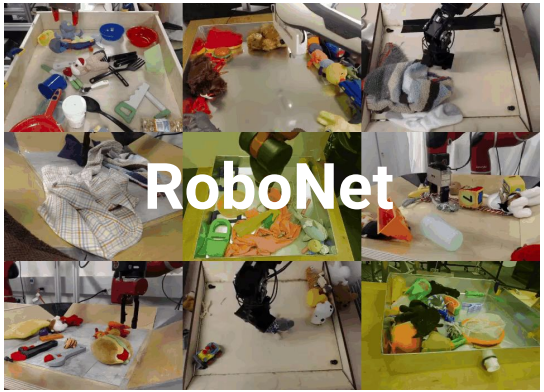


Powerful Function Approximators

Large & Diverse Datasets

Success

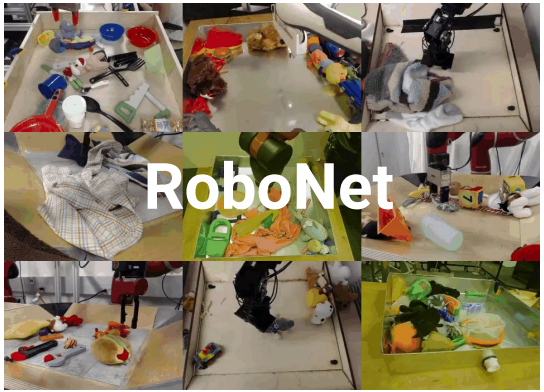
RL for Real-World: RL with Large Datasets



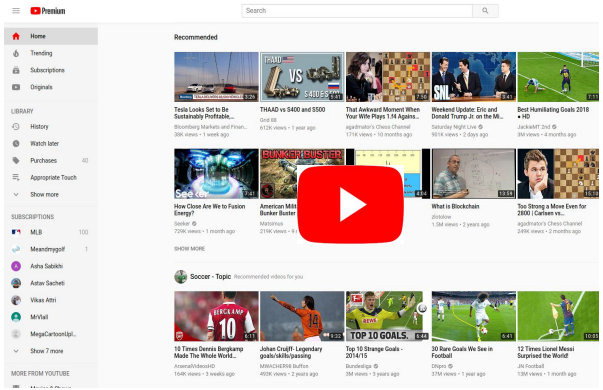
Robotics

[1] Dasari, Ebert, Tian, Nair, Bucher, Schmeckpeper, .. Finn. RoboNet: Large-Scale Multi-Robot Learning.

RL for Real-World: RL with Large Datasets



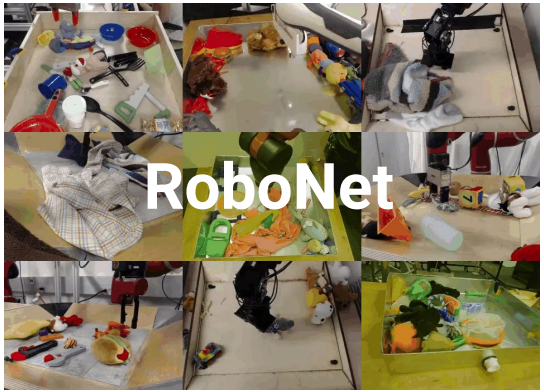
Robotics



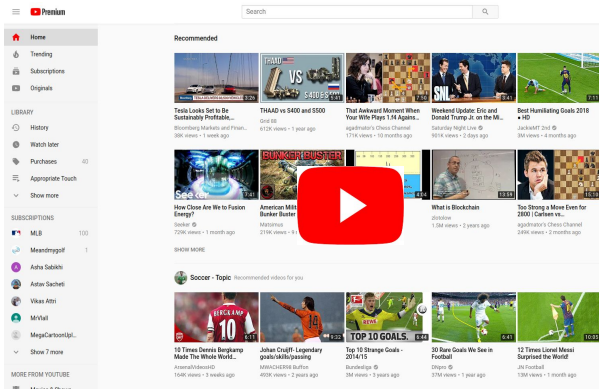
Recommender Systems

[1] Dasari, Ebert, Tian, Nair, Bucher, Schmeckpeper, .. Finn. RoboNet: Large-Scale Multi-Robot Learning.

RL for Real-World: RL with Large Datasets



Robotics



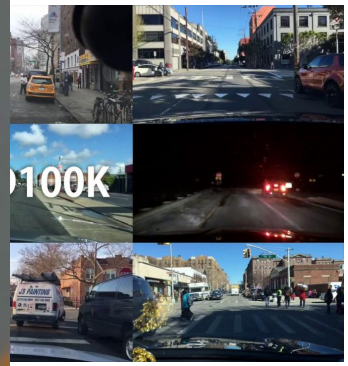
Recommender Systems

Self-Driving Cars

[1] Dasari, Ebert, Tian, Nair, Bucher, Schmeckpeper, .. Finn. RoboNet: Large-Scale Multi-Robot Learning.
[2] Yu, Xian, Chen, Liu, Liao, Madhavan, Darrell. BDD100K: A Large-scale Diverse Driving Video Database.

RL for Real-World: RL with Large Datasets

Robotic



ving Cars

[1] Dasari, Ebert, Tian g.
[2] Yu, Xian, Chen, Liu, Liao, Madhavan, Darrell. BDD100K: A Large-scale Diverse Driving Video Database.

Offline RL: A Data-Driven RL Paradigm

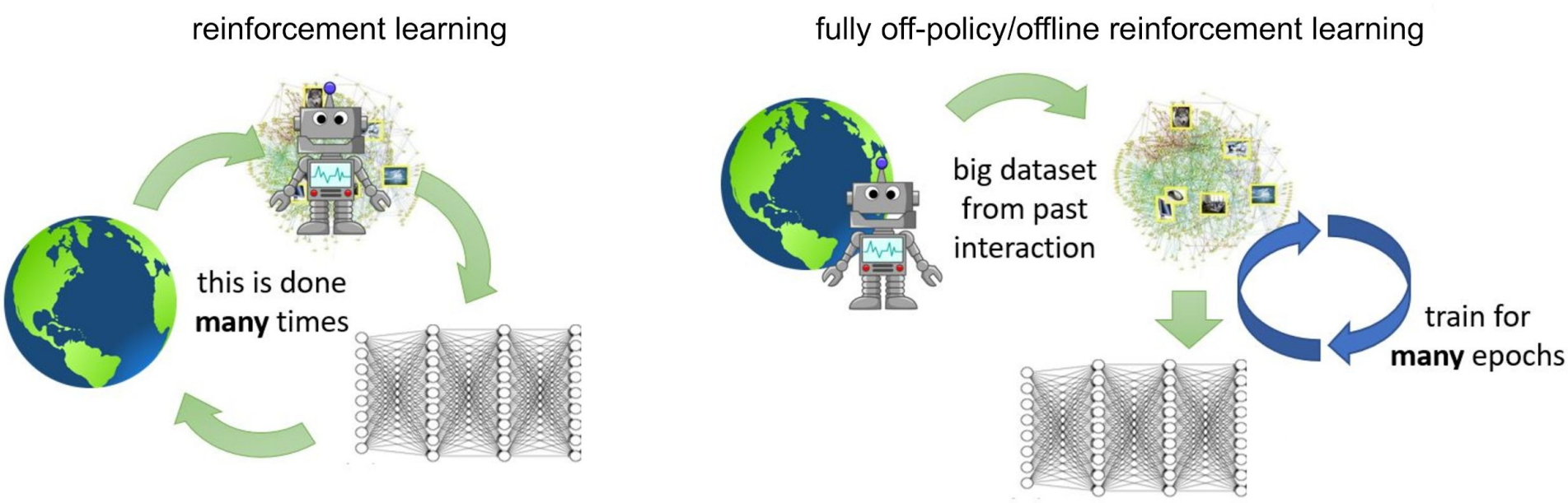


Image Source: Data-Driven Deep Reinforcement Learning, BAIR Blog. https://bair.berkeley.edu/blog/2019/12/05/bear/

Can Offline RL with Large & Diverse Datasets Succeed?

- Fujimoto et al. [2019] present a pessimistic view that standard off-policy methods don't work in the offline setting even with large diverse datasets.

[1] Fujimoto, Meger, Precup. Off-Policy Deep Reinforcement Learning without Exploration.

Can Offline RL with Large & Diverse Datasets Succeed?

- Fujimoto et al. [2019] present a pessimistic view that standard off-policy methods don't work in the offline setting even with large diverse datasets.
- Zhang and Sutton [2017] claim that large replay buffer deteriorates performance of simple DQN variants

[1] Fujimoto, Meger, Precup. Off-Policy Deep Reinforcement Learning without Exploration.
[2] Zhang, Sutton. A Deeper Look at Experience Replay.

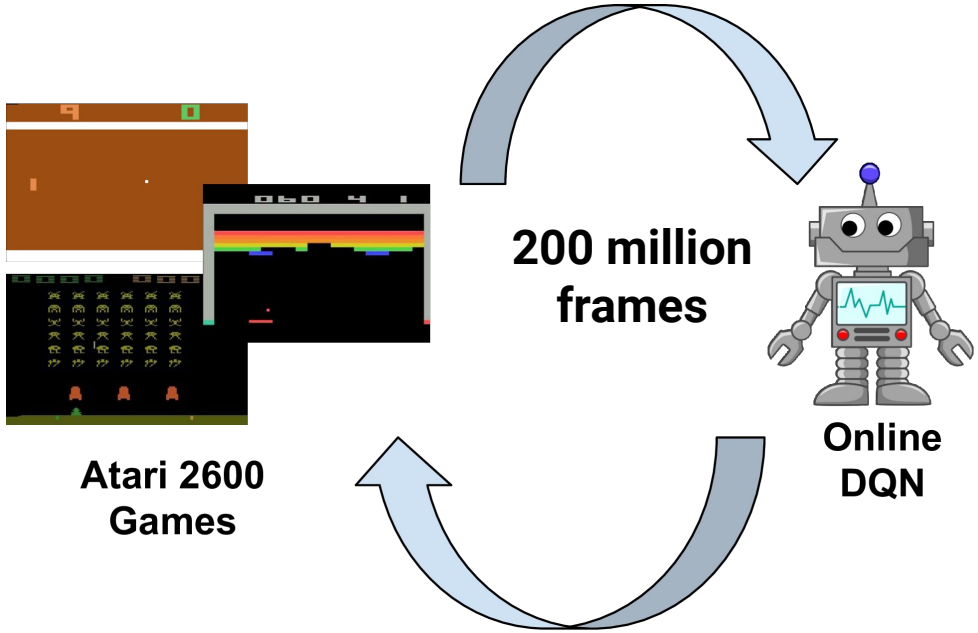
Can Offline RL with Large & Diverse Datasets Succeed?

- Deadly Triad: Off-Policy, Bootstrapping, and Function Approximation
  - Current “off-policy” algorithms use near on-policy exploratory policies

Can Offline RL with Large & Diverse Datasets Succeed?

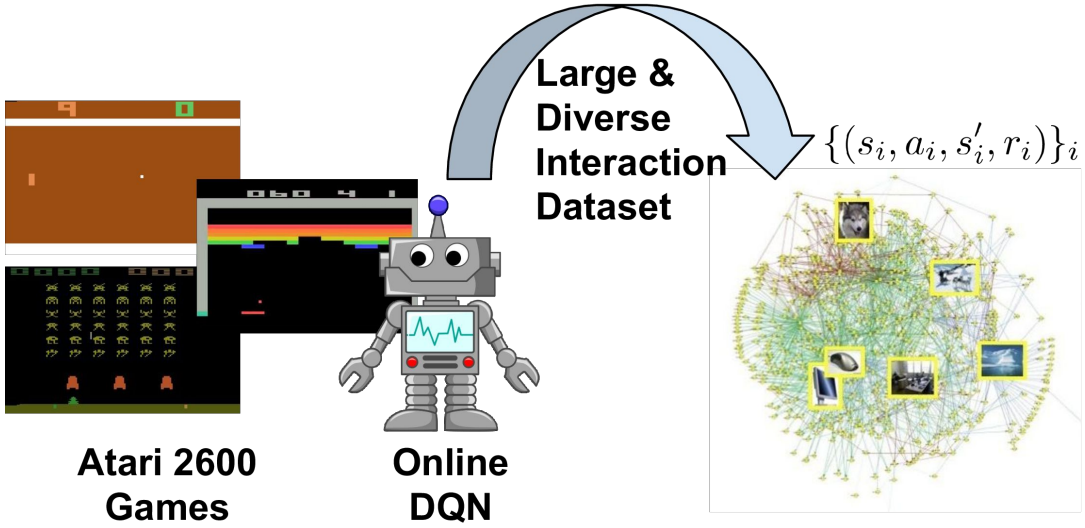
- Deadly Triad: Off-Policy, Bootstrapping, and Function Approximation
  - Current “off-policy” algorithms use near on-policy exploratory policies
- No new corrective feedback in offline RL

Offline Deep RL on Atari 2600 Games



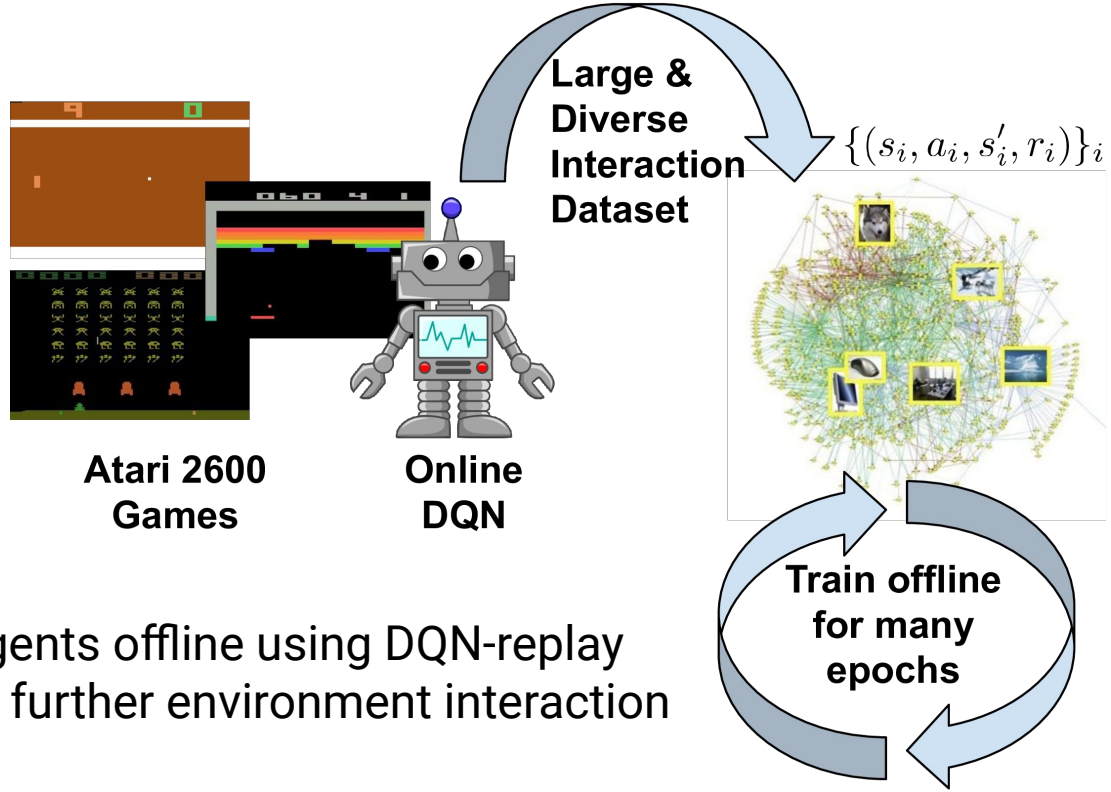
Train a Nature DQN agent on 60 Atari 2600 games with sticky actions for 200 million frames (standard protocol)

Offline Deep RL on Atari 2600 Games



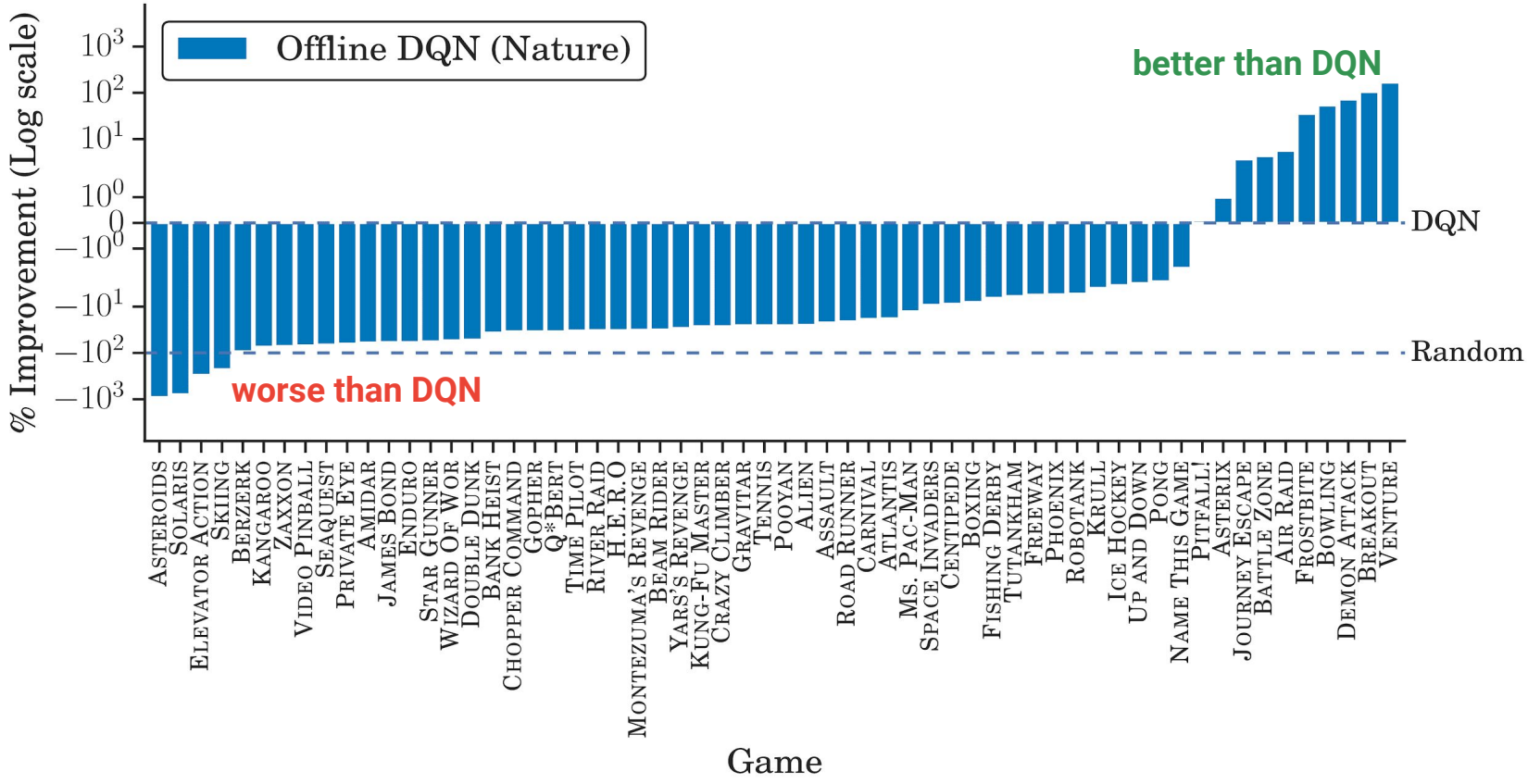
For each game, save all of the tuples of
(observation, action, next observation, reward)
encountered to DQN-replay dataset(s)

Offline Deep RL on Atari 2600 Games

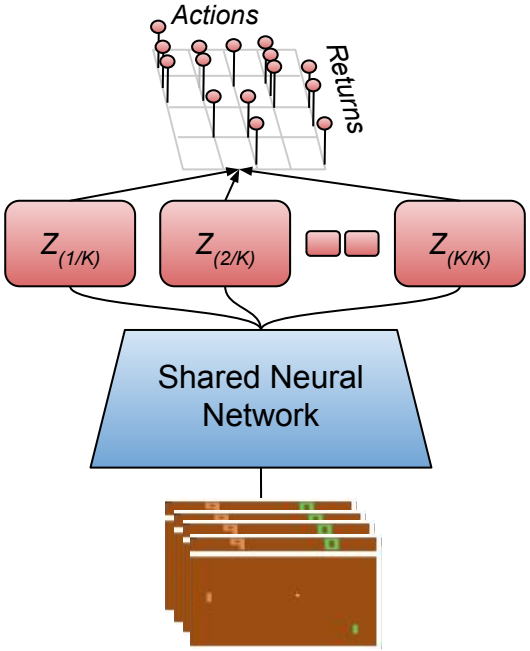


Train off-policy agents offline using DQN-replay dataset without any further environment interaction

Does Offline DQN work?



Let's try sophisticated off-policy algorithms!



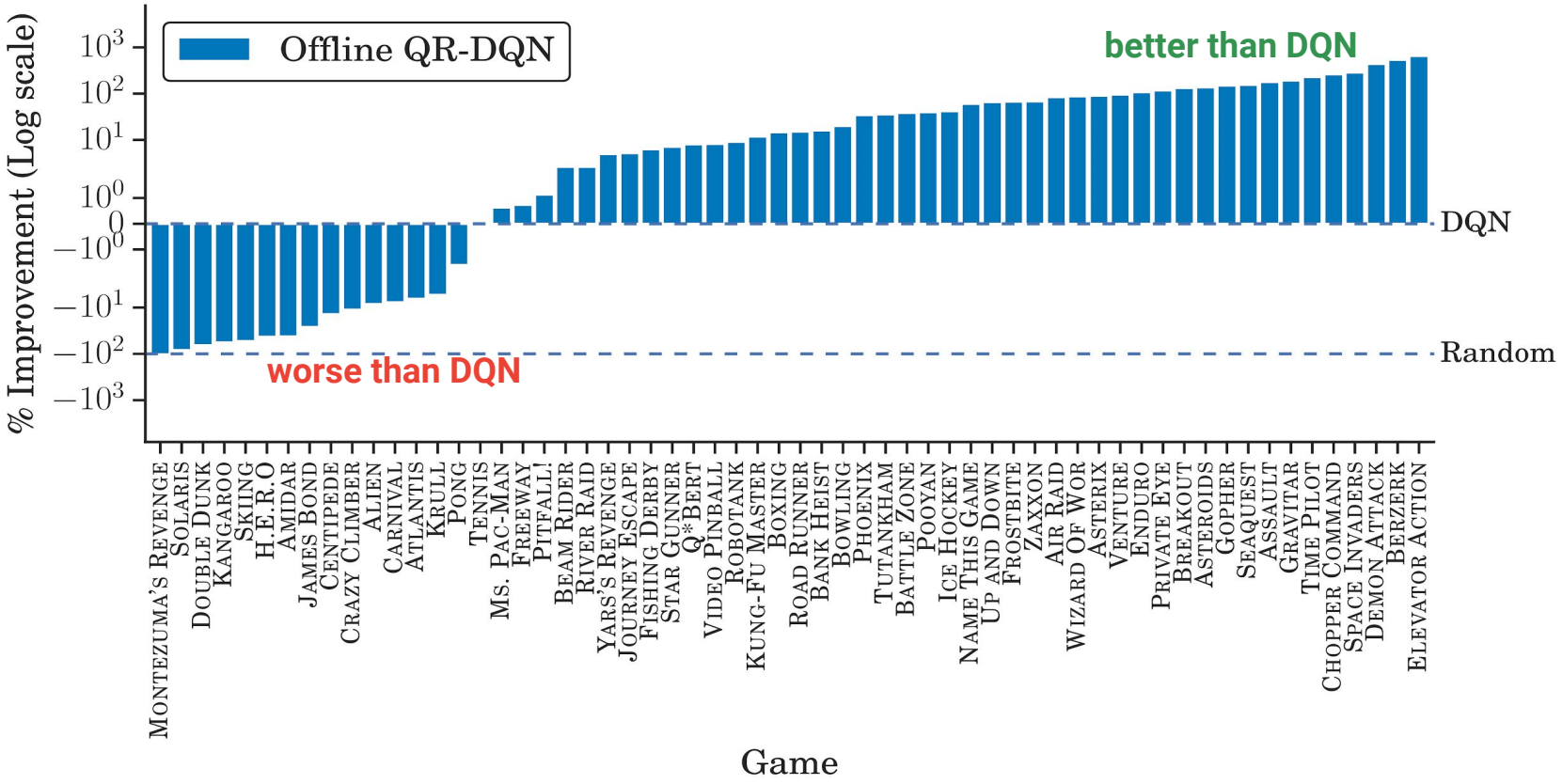
QR-DQN

Distributional RL makes use of a distribution over returns, denoted Z(s, a), instead of the scalar Q-function.

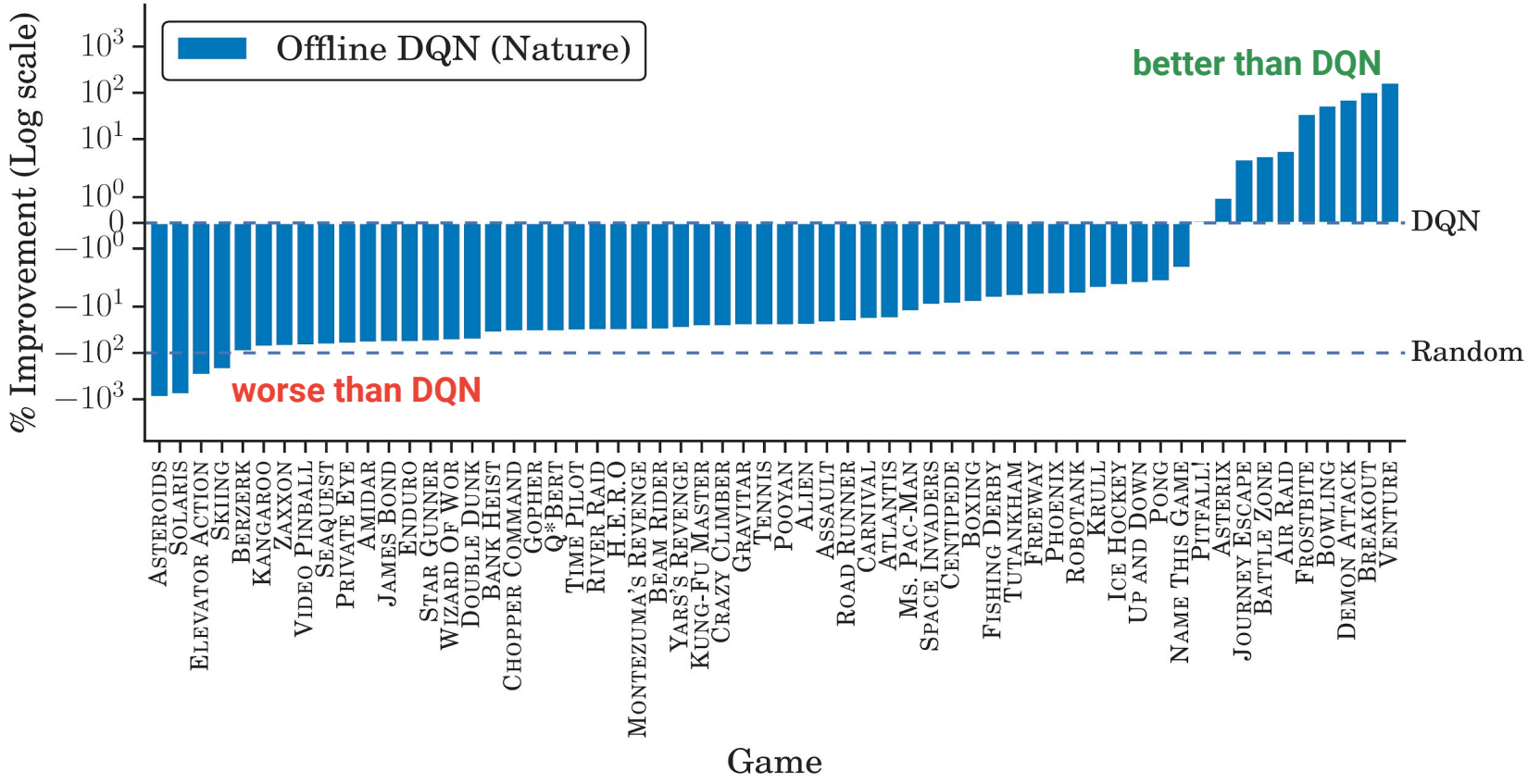
$$Z(s,a;\theta) := \frac{1}{K}\sum_{i=1}^{K}\delta_{\theta_i(s,a)}$$

$$Q(s,a;\theta) := \mathbb{E}[Z] = \frac{1}{K}\sum_{i=1}^{K}\theta_i(s,a)$$

Does Offline QR-DQN work?



Does Offline DQN work?



Rethinking Offline RL Algorithms

- Emphasis on Generalization
  - Given a fixed training dataset, generalize to unseen states that will be seen when we run the agent in the environment.

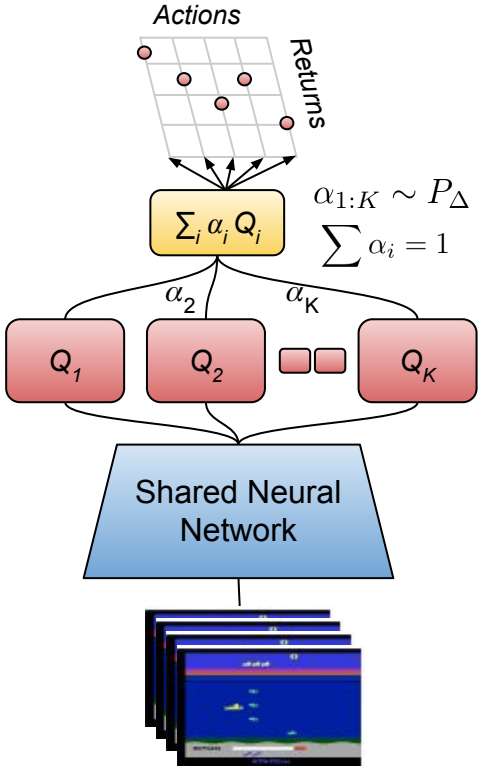
Rethinking Offline RL Algorithms

- Emphasis on Generalization
  - Given a fixed training dataset, generalize to unseen states that will be seen when we run the agent in the environment.
- Ensemble of *Q*-estimates:
  - Ensembling and Dropout are widely used in supervised learning for improving generalization.

Rethinking Offline RL Algorithms

- Emphasis on Generalization
  - Given a fixed training dataset, generalize to unseen states that will be seen when we run the agent in the environment.
- *Q*-learning as constraint satisfaction:
  - $\forall\ (s, a, s', r):\ \ Q^*(s, a) = r\ + max_{a'}\ Q^*(s', a')$

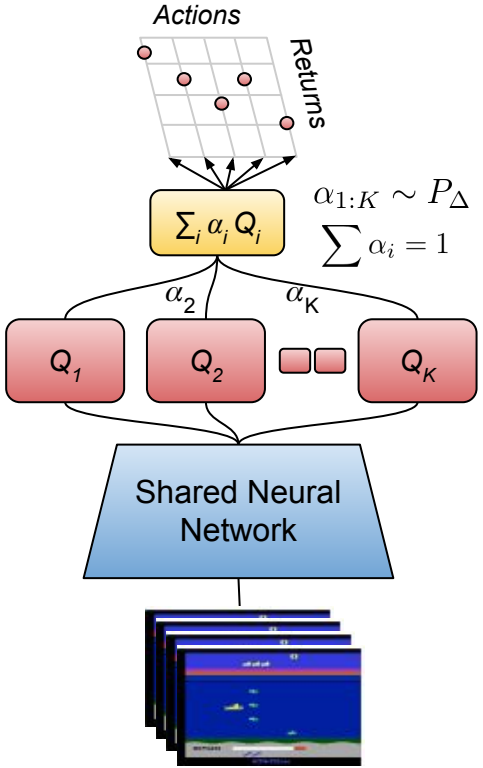
Random Ensemble Mixture (REM)



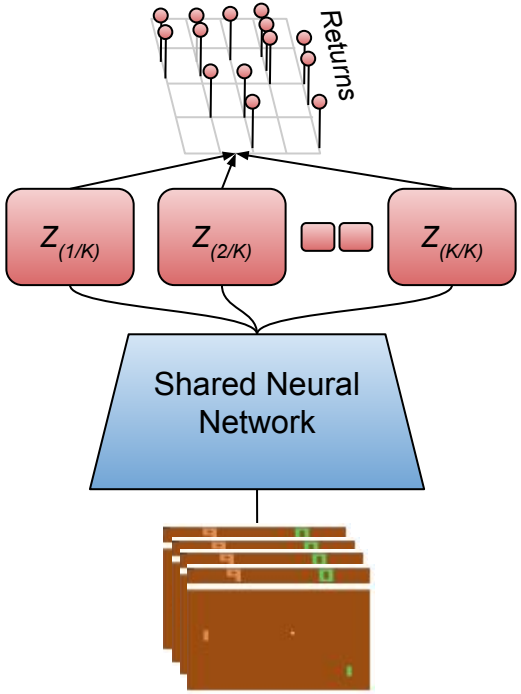
REM

Minimize TD error on randomly sampled convex combination of multiple *Q*-estimates.

REM vs QR-DQN

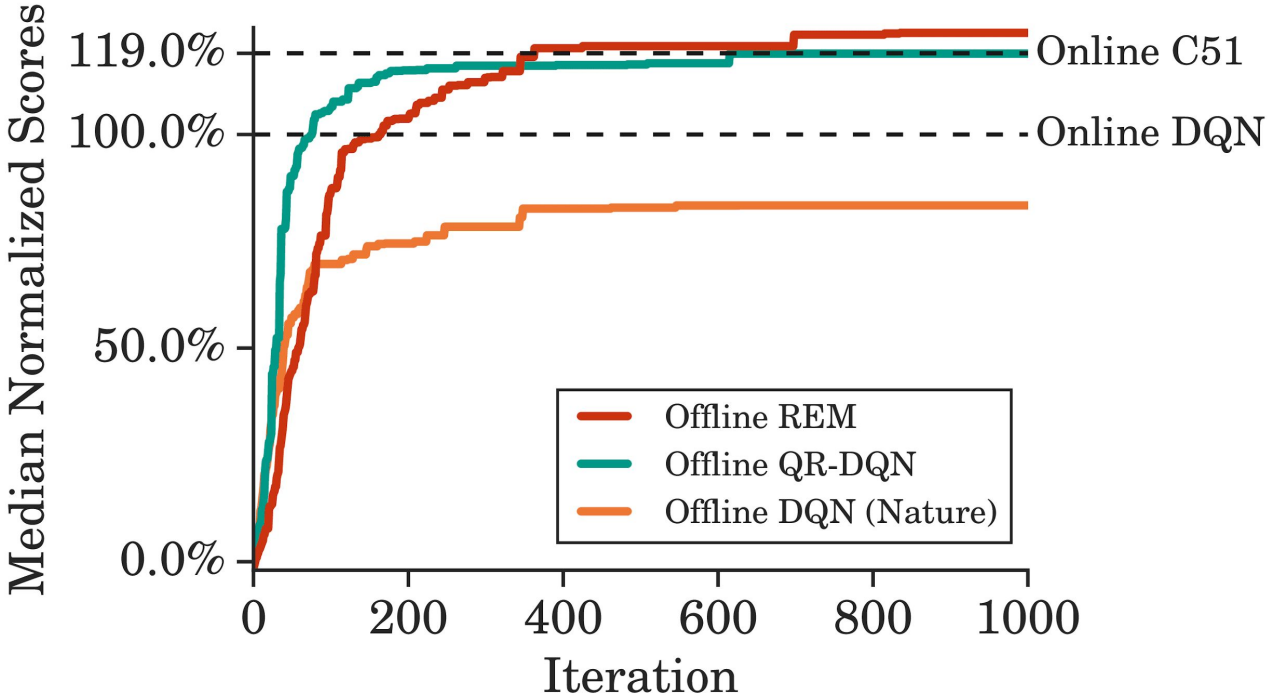


REM



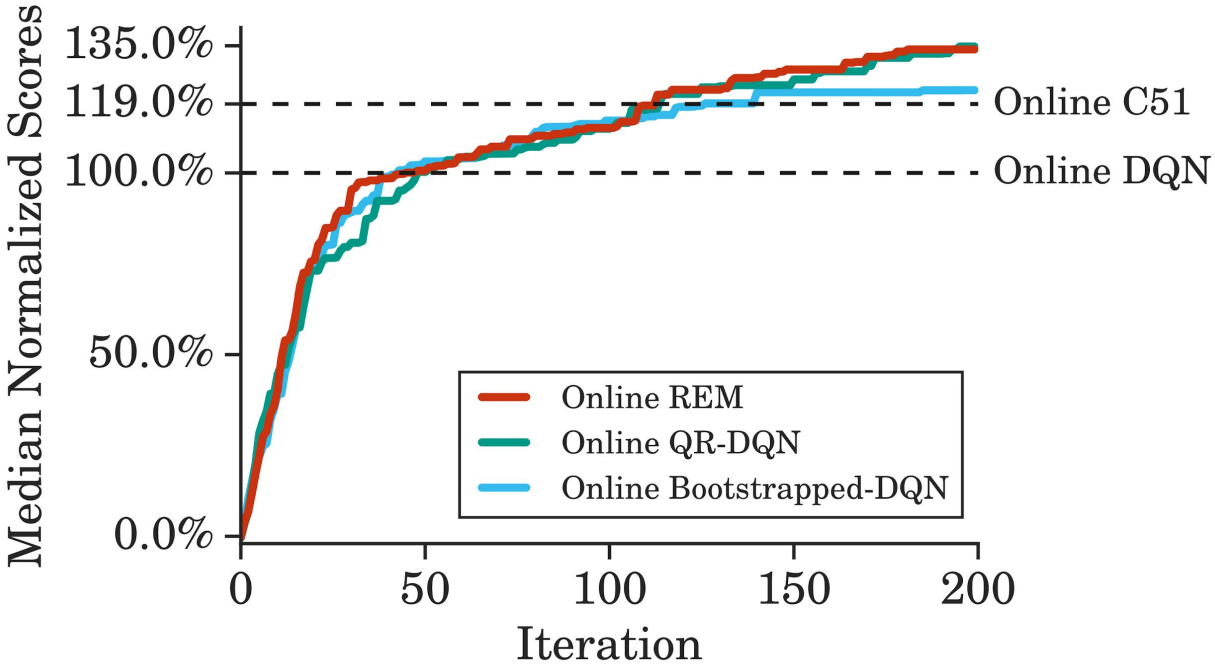
QR-DQN

Offline Stochastic Atari Results



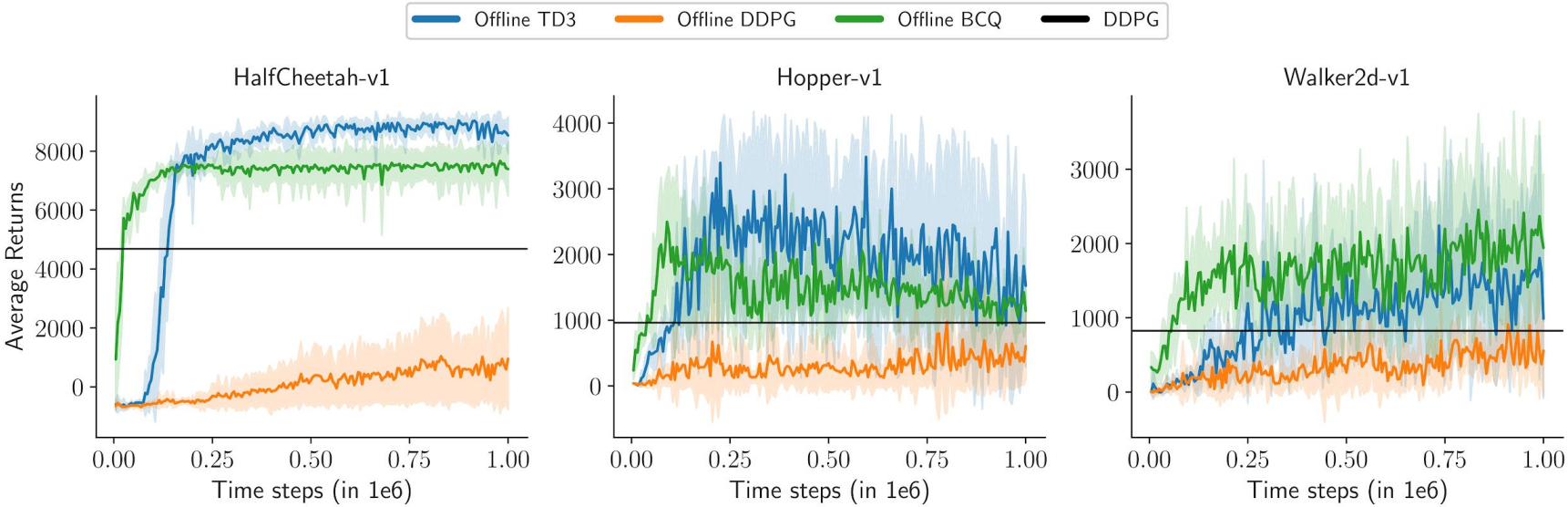
Scores averaged over 5 runs of offline agents trained using DQN replay data across 60 Atari games for 5X gradient steps. Offline REM surpasses gains from online C51 and offline QR-DQN.

Online Stochastic Atari results



Average normalized scores of online agents trained for 200 million game frames. Multi-network REM with 4 Q-functions performs comparably to QR-DQN.

Offline Continuous Control Experiments

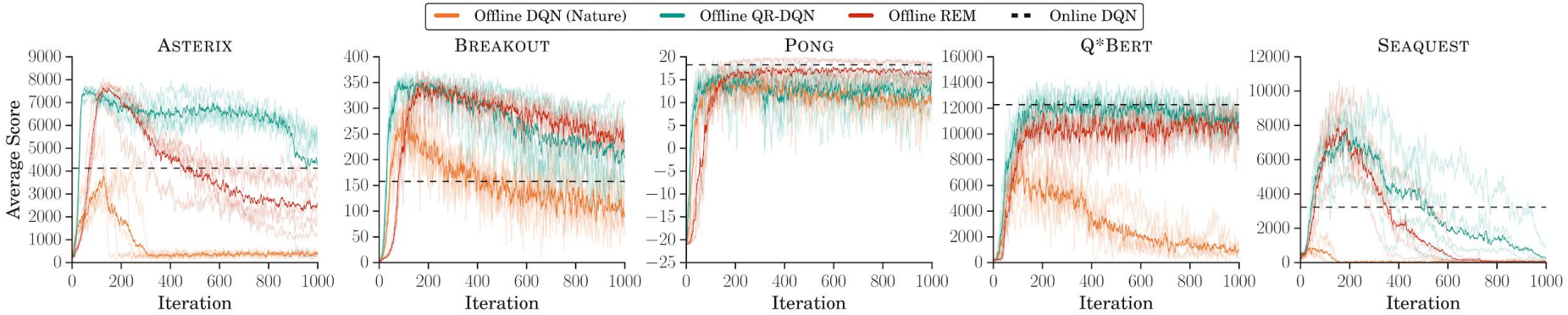


Offline agents trained using full experience replay of DDPG on MuJoCo environments.

Future Challenges

"The potential for off-policy learning remains tantalizing, the best way to achieve it still a mystery." - Sutton & Barto

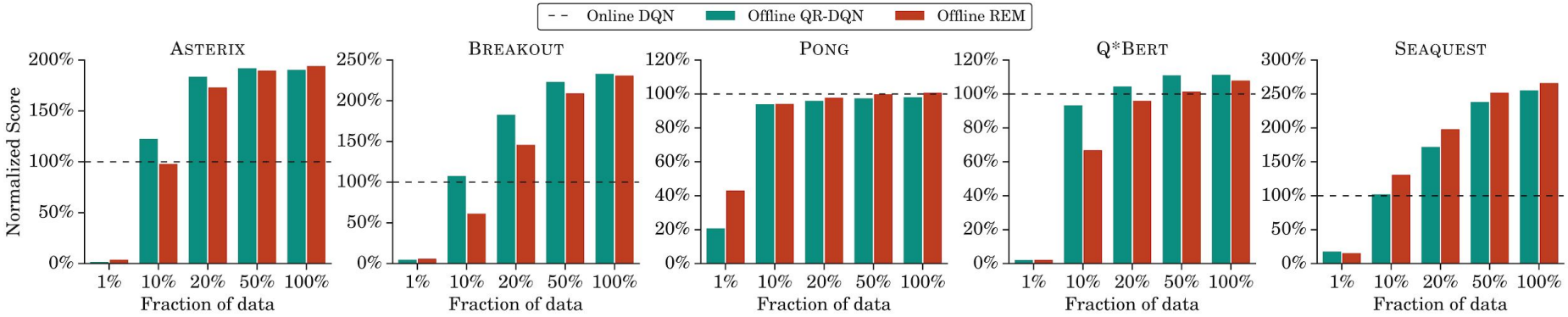
Future Challenges: (1) Stability/Overfitting



Average online scores of offline agents trained on 5 games using logged DQN replay data for 5X gradient steps compared to online DQN.

- More gradient updates on fixed data eventually lead to degradation in performance for most environments :(
- Need for early stopping for *off-policy* RL

Future Challenges: (2) Sample Efficiency



Randomly subsample N% of frames from 200 million frames for offline training.

Divergence with 1% of data for prolonged training!

TL;DR

In addition to real-world implications, offline RL provides a reproducible experimental setup for:

- Isolating *exploitation* from exploration
- Developing *simple* and *effective* off-policy algorithms (*e.g.*, REM)
- Improving *sample efficiency* and *stability* of *off-policy* algorithms

Thank you!

For code, DQN-replay dataset and paper, refer to

offline-rl.github.io