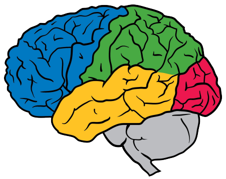
Rishabh Agarwal, Dale Schuurmans, Mohammad Norouzi

How I Learned To Stop Worrying And Love Offline RL

An Optimistic Perspective on Offline Reinforcement Learning

Google Research

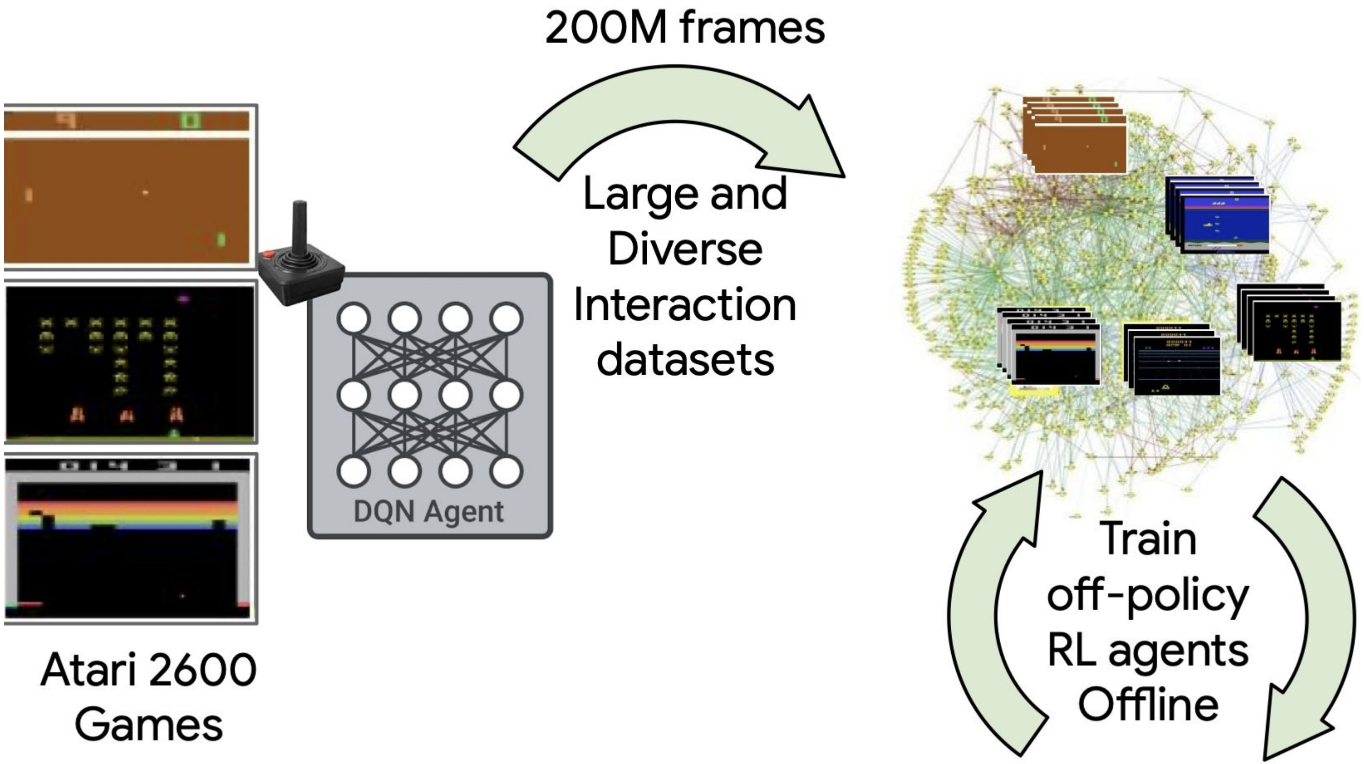


Offline RL: A Data-Driven RL Paradigm

Reinforcement Learning with Online Interactions

Offline Reinforcement Learning

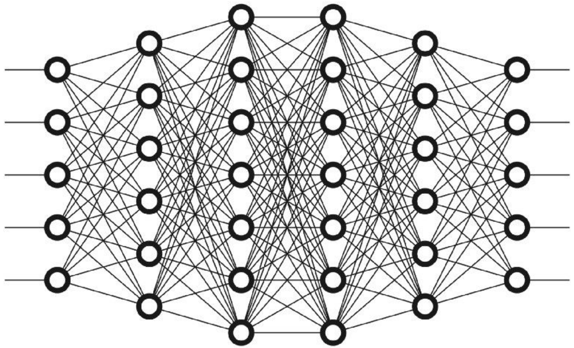
Offline RL on Atari 2600



Full Talk

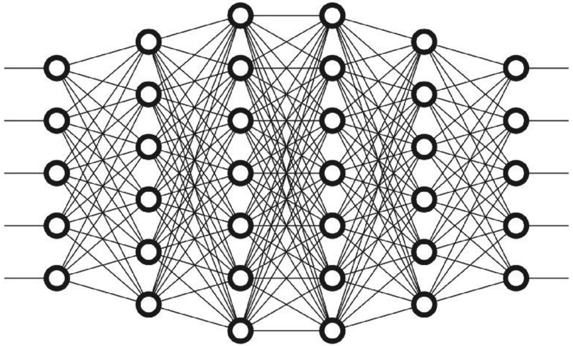
What makes Deep Learning Successful?

Expressive function approximators



What makes Deep Learning Successful?

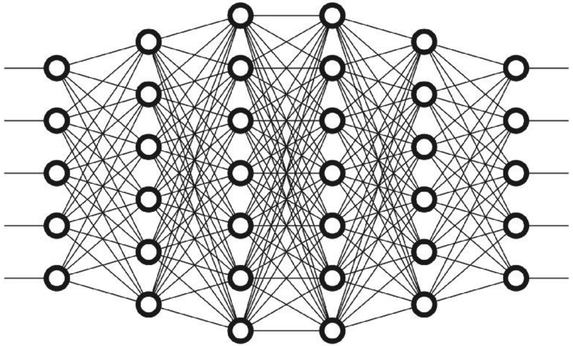
Expressive function approximators



Powerful learning algorithms

What makes Deep Learning Successful?

Expressive function approximators

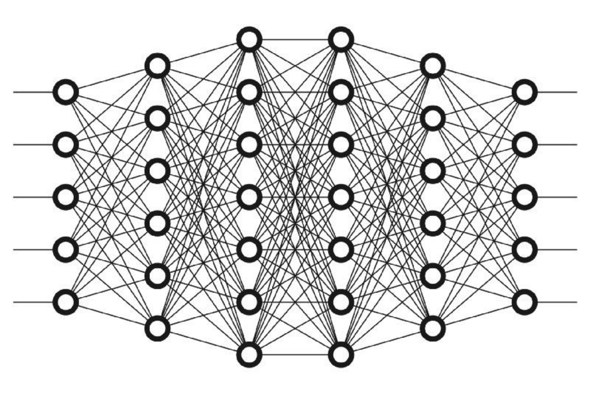


Powerful learning algorithms

Large and Diverse Datasets

How to make Deep RL similarly successful?

Expressive function approximators

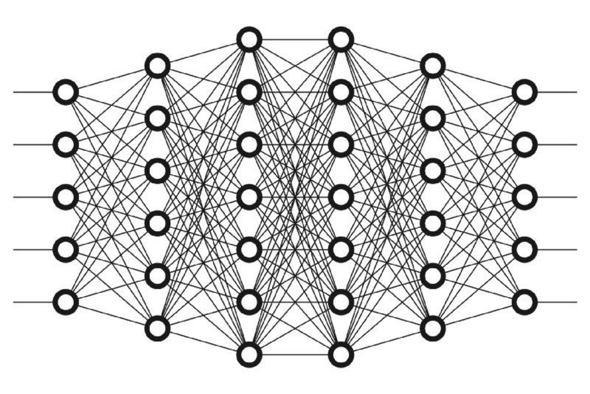


Good learning algorithms e.g., actor-critic, approx DP

How to make Deep RL similarly successful?

Google Research

Expressive function approximators

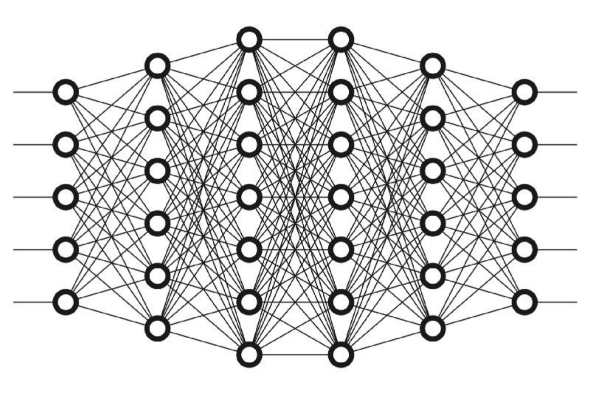


Good learning algorithms e.g., actor-critic, approx DP

Large and Diverse Datasets

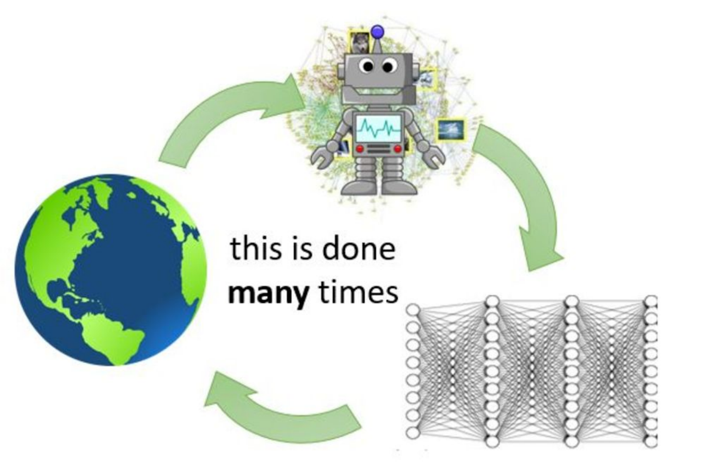
How to make Deep RL similarly successful?

Expressive function approximators



Good learning algorithms e.g., actor-critic, approx DP

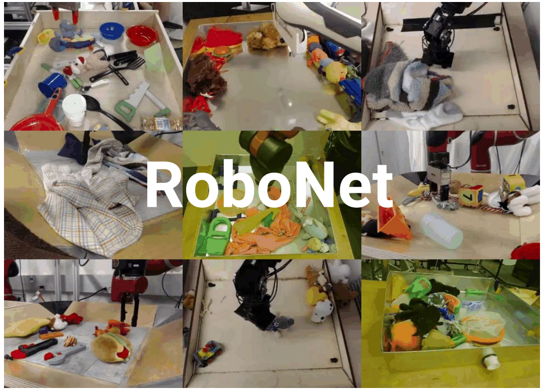
Interactive Environments



Active Data Collection

Google Research

RL for Real-World: RL with Large Datasets

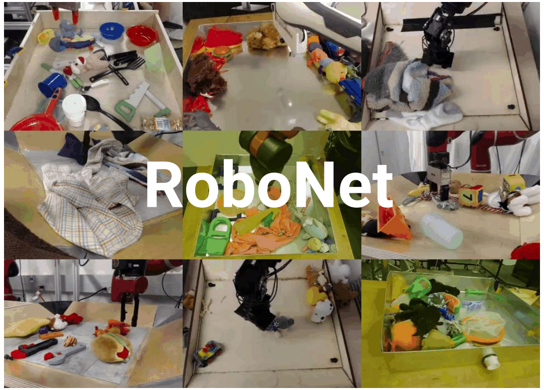


Robotics

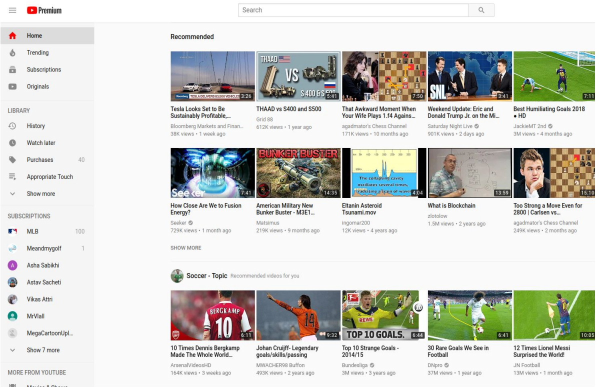
[1] Dasari, Ebert, Tian, Nair, Bucher, Schmeckpeper, .. Finn. RoboNet: Large-Scale Multi-Robot Learning.
[2] Yu, Xian, Chen, Liu, Liao, Madhavan, Darrell. BDD100K: A Large-scale Diverse Driving Video Database.

Google Research

RL for Real-World: RL with Large Datasets



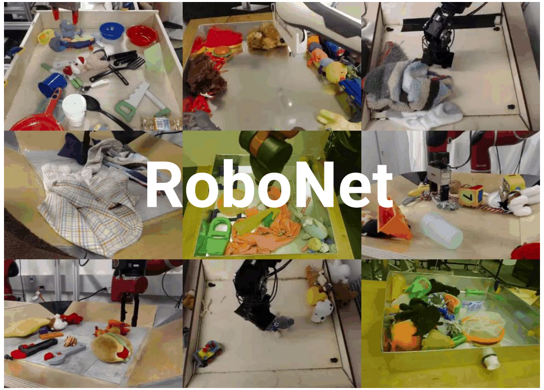
Robotics



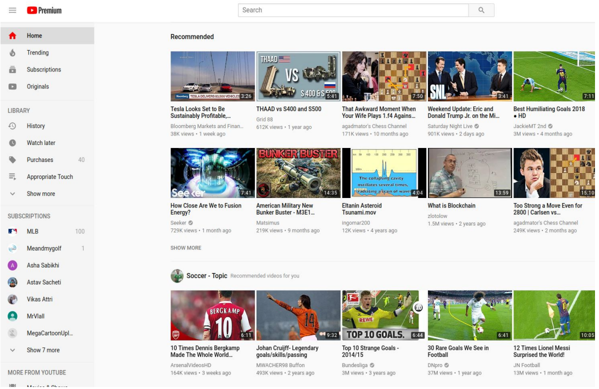
Recommender Systems

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

RL for Real-World: RL with Large Datasets



Robotics



Recommender Systems

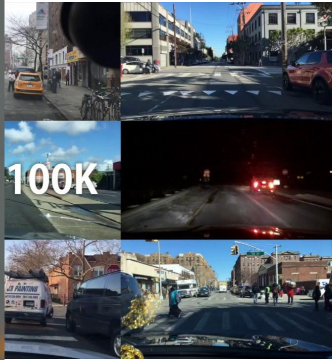
Autonomous Driving

[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.

Google Research

RL for Real-World: RL with Large Datasets

Roboti



ing Cars

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

Offline RL: A Data-Driven RL Paradigm

Reinforcement Learning with Online Interactions

Offline Reinforcement Learning

Offline RL: A Data-Driven RL Paradigm

Offline RL can help:

- **Pretrain** agents on existing logged data.

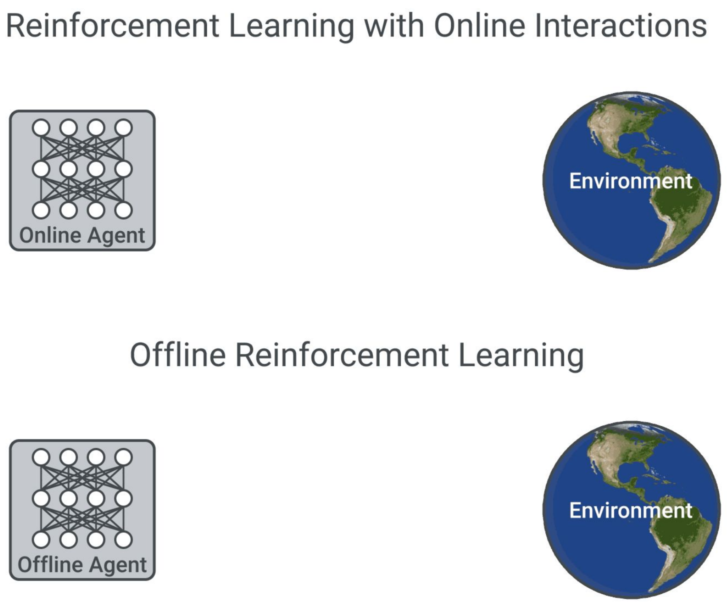
Reinforcement Learning with Online Interactions

Offline Reinforcement Learning

Offline RL: A Data-Driven RL Paradigm

Offline RL can help:

- Pretrain agents on existing logged data.
- **Evaluate** RL algorithms on the basis of exploitation alone on common datasets.

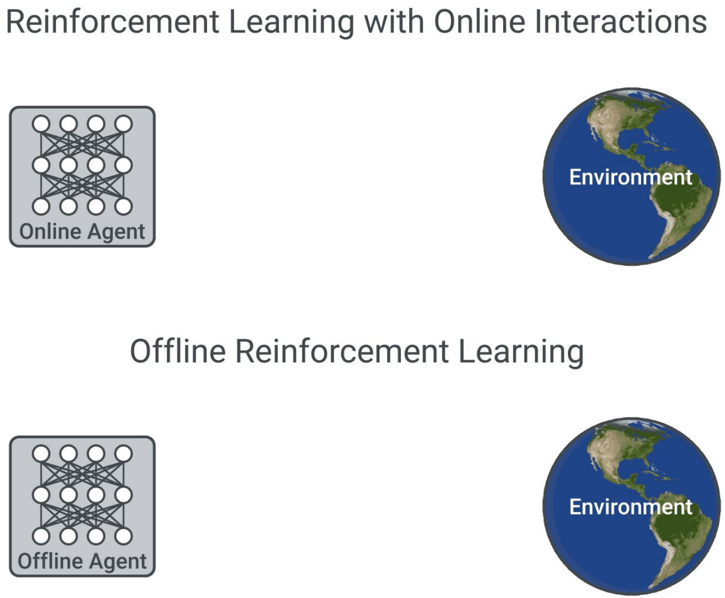


Offline RL: A Data-Driven RL Paradigm

Google Research

Offline RL can help:

- Pretrain the agents on existing logged data.
- Evaluate RL algorithms on the basis of exploitation alone on common datasets.
- Deliver real-world **impact**.



But .. Offline RL is Challenging!

Distribution mismatch

But .. Offline RL is Challenging!

No New Corrective Feedback

But .. Offline RL is Challenging!

Fully Off-Policy

Bootstrapping
(Learning guess from a guess)

Function
Approximation

Standard RL fails in the Offline setting?

Off-Policy Deep Reinforcement Learning without Exploration

Scott Fujimoto [1 2] David Meger [1 2] Doina Precup [1 2]

Abstract

Many practical applications of reinforcement learning constrain agents to learn from a fixed batch of data which has already been gathered, without offering further possibility for data collection. In this paper, we demonstrate that due to

require further interactions with the environment to compensate (Hester et al., 2017; Sun et al., 2018; Cheng et al., 2018). On the other hand, batch reinforcement learning offers a mechanism for learning from a fixed dataset without restrictions on the quality of the data.

Most modern off-policy deep reinforcement learning al-

KEEP DOING WHAT WORKED: BEHAVIOR MODELLING PRIORS FOR OFFLINE REINFORCEMENT LEARNING

Noah Y. Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, Martin Riedmiller

DeepMind
{siegeln}@google.com

ABSTRACT

Off-policy reinforcement learning algorithms promise to be applicable in settings where only a fixed data-set (batch) of environment interactions is available and no new experience can be acquired. This property makes these algorithms appealing

Behavior Regularized Offline Reinforcement Learning

Yifan Wu* Carnegie Mellon University yw4@cs.cmu.edu

George Tucker Google Research gjt@google.com

Ofir Nachum Google Research ofirnachum@google.com

Abstract

In reinforcement learning (RL) research, it is common to assume access to direct *online* interactions with the environment. However in many real-world applications, access to the environment is limited to a fixed *offline* dataset of logged experience. In such settings, standard RL algorithms have been shown to diverge or otherwise yield poor performance. Accordingly, recent work has suggested a number of remedies to these issues. In this work, we introduce a general framework, *behavior regularized actor critic* (BRAC), to empirically evaluate recently proposed methods as well as a number of simple baselines across a variety of offline continuous control tasks. Surprisingly, we find that many of the technical complexities introduced in recent methods are unnecessary to achieve strong performance. Additional ablations provide insights into which design choices matter most in the offline RL setting.[1]

Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction

Aviral Kumar* UC Berkeley aviralk@berkeley.edu

Justin Fu* UC Berkeley justinjfu@eecs.berkeley.edu

George Tucker Google Brain gjt@google.com

Sergey Levine UC Berkeley, Google Brain svlevine@eecs.berkeley.edu

Abstract

Off-policy reinforcement learning aims to leverage experience collected from prior policies for sample-efficient learning. However, in practice, commonly used off-policy approximate dynamic programming methods based on Q-learning and

Standard RL fails in the Offline setting ..

Off-Policy Deep Reinforcement Learning without Exploration

Scott Fujimoto [1 2] David Meger [1 2] Doina Precup [1 2]

Abstract

Many practical applications of reinforcement learning constrain agents to learn from a fixed batch of data which has already been gathered, without offering further possibility for data collection. In this paper, we demonstrate that due to

require further interactions with the environment to compensate (Hester et al., 2017; Sun et al., 2018; Cheng et al., 2018). On the other hand, batch reinforcement learning of- fer- res- Mo-

KEEP DOING WHAT WORKED: BEHAVIOR MODELLING PRIORS FOR OFFLINE REINFORCEMENT LEARNING

Noah Y. Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, Martin Riedmiller

DeepMind

{siegeln}@google.com

n settings
ble and no
appealing

A Deeper Look at Experience Replay

Shangtong Zhang, Richard S. Sutton
Dept. of Computing Science
University of Alberta
{shangtong.zhang, rsutton}@ualberta.ca

Abstract

Recently experience replay is widely used in various deep reinforcement learning (RL) algorithms, in this paper we rethink the utility of

et al. 2016), which is a desired property for many RL algorithms as they are often pretty hungry for data. Although algorithms in pre-deep-RL era do not need to care about how to stabilize a neural network, they do care data efficiency. If experience replay is a perfect idea,

Behavior Regularized Offlin

Yifan Wu*
Carnegie Mellon University
yw4@cs.cmu.edu

George
Google I
gjt@goo

Abstr

In reinforcement learning (RL) research *online* interactions with the environment. access to the environment is limited to a fix such settings, standard RL algorithms have poor performance. Accordingly, recent wor these issues. In this work, we introduce a ge *critic* (BRAC), to empirically evaluate recent simple baselines across a variety of offline co that many of the technical complexities introduced in recent methods are unnecessary to achieve strong performance. Additional ablations provide insights into which design choices matter most in the offline RL setting.[1]

earning via Bootstrapping eduction

Justin Fu*
UC Berkeley
justinjfu@eecs.berkeley.edu

Sergey Levine
UC Berkeley, Google Brain
svlevine@eecs.berkeley.edu

tract

Off-policy reinforcement learning aims to leverage experience collected from prior policies for sample-efficient learning. However, in practice, commonly used off-policy approximate dynamic programming methods based on Q-learning and

Can standard off-policy RL succeed in the offline setting?

Offline RL on Atari 2600

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

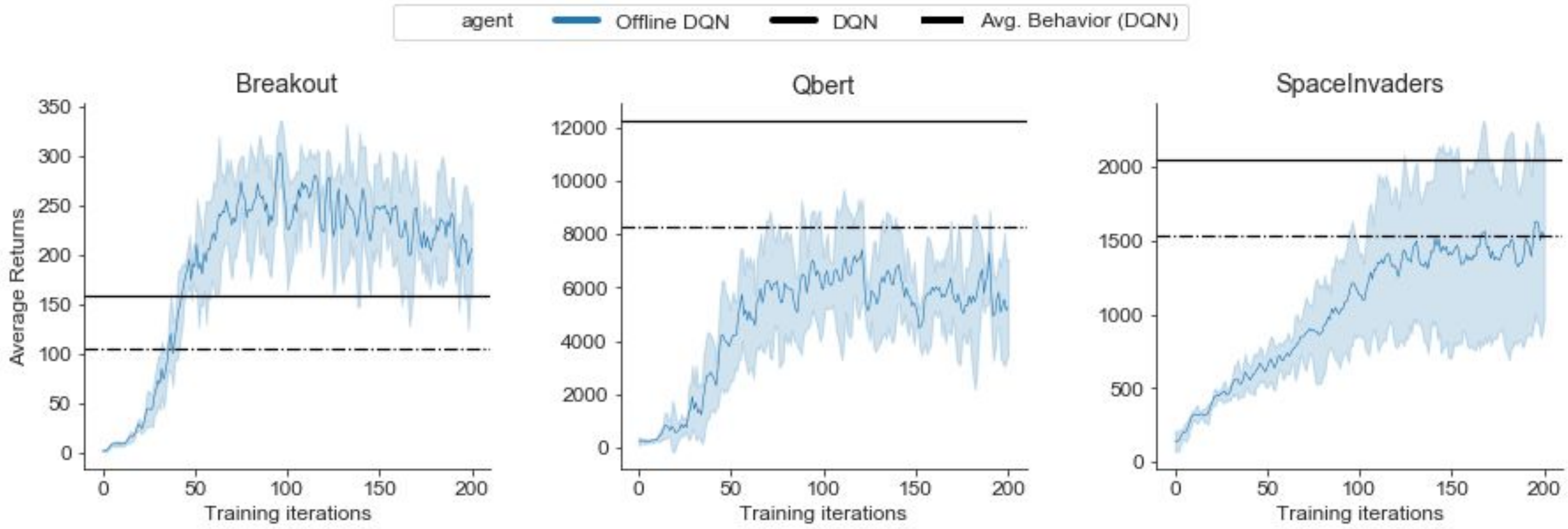
Offline RL on Atari 2600

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

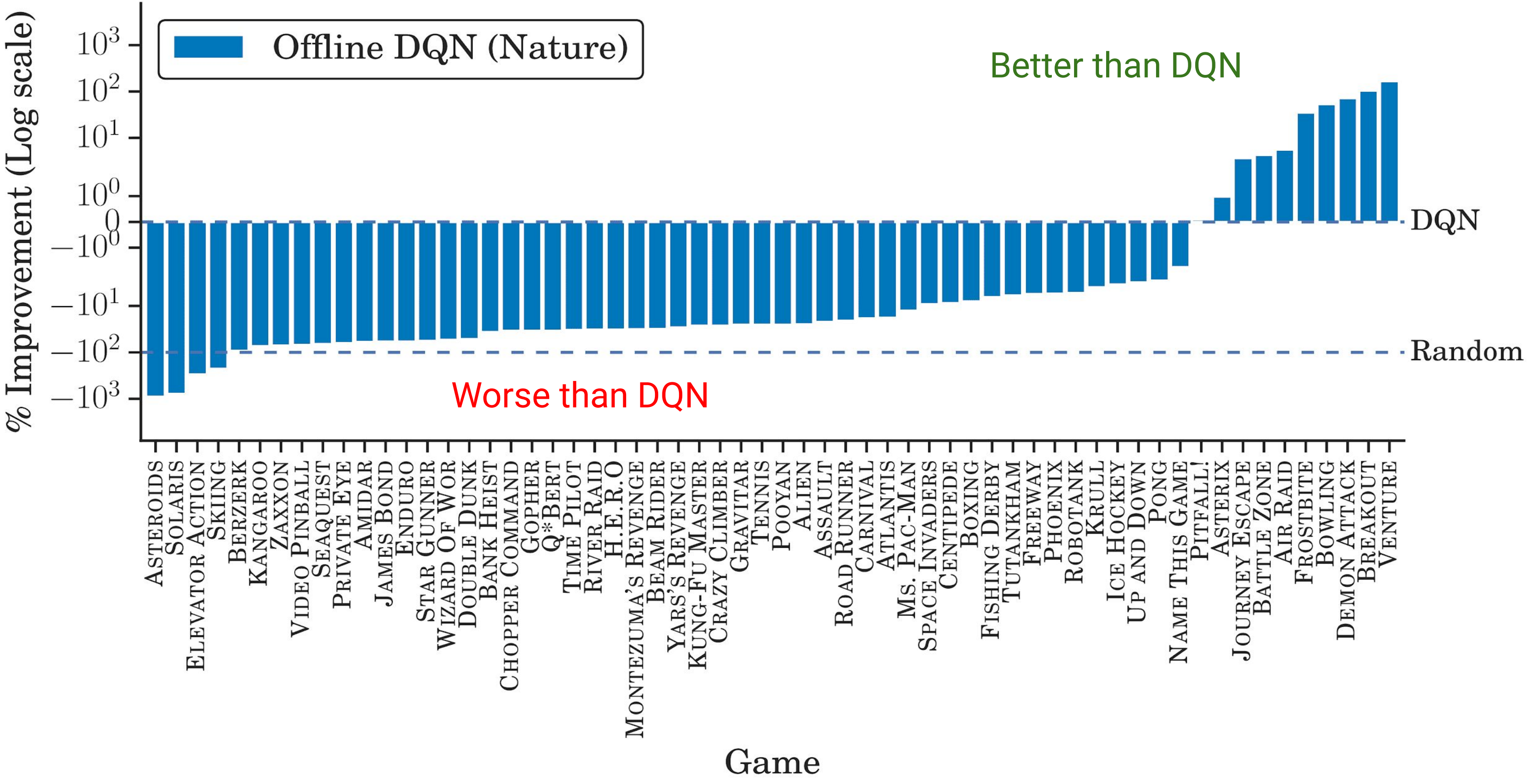
Offline RL on Atari 2600

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

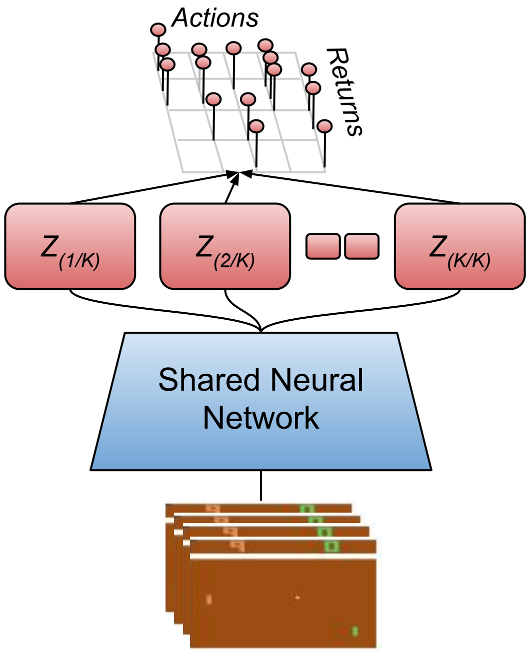
Offline DQN on DQN Replay Dataset



Does Offline DQN work?



Let's try recent off-policy methods!



QR-DQN

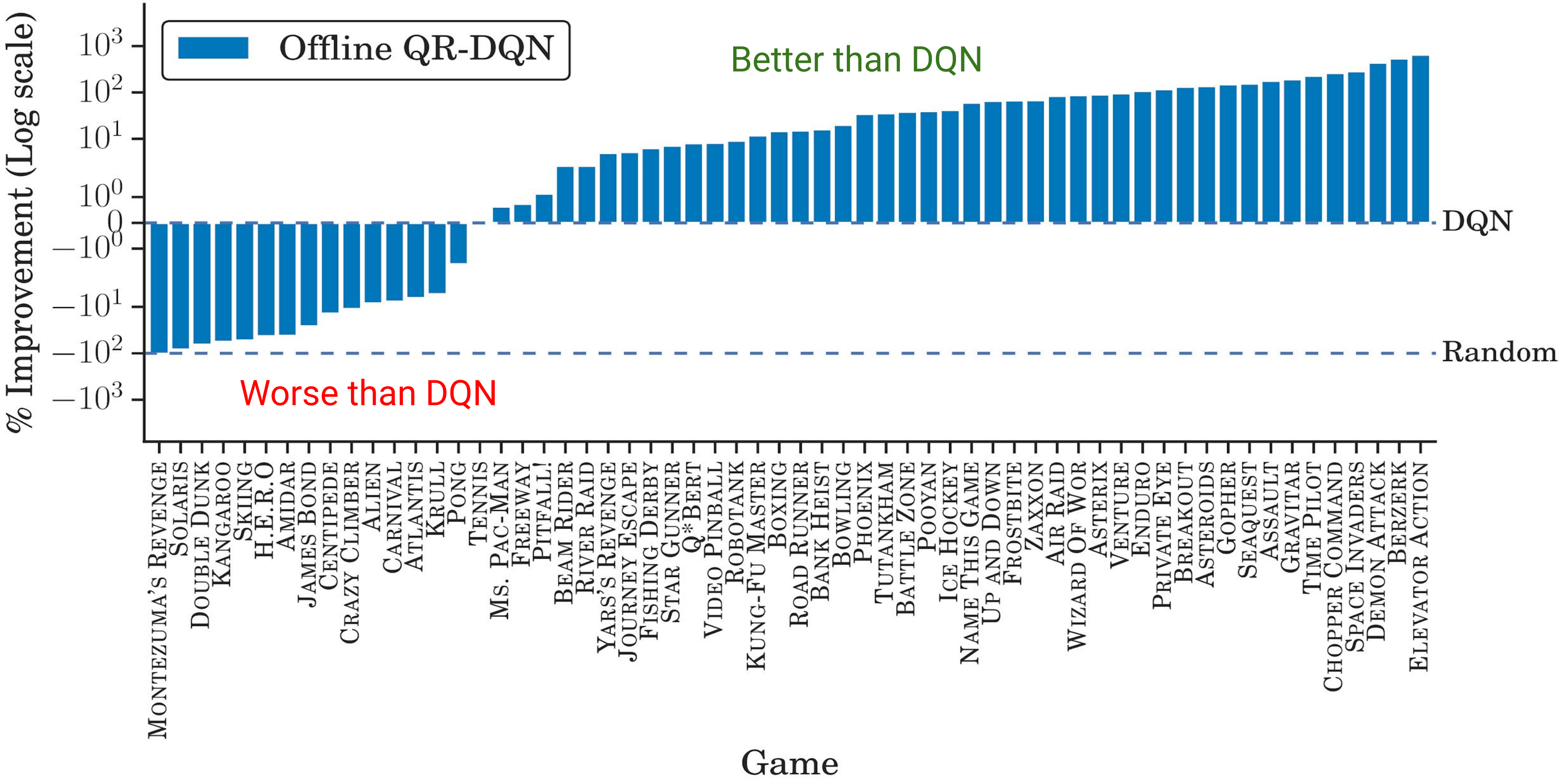
Distributional RL uses Z(s, a), a distribution over returns, instead of the Q-function.

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

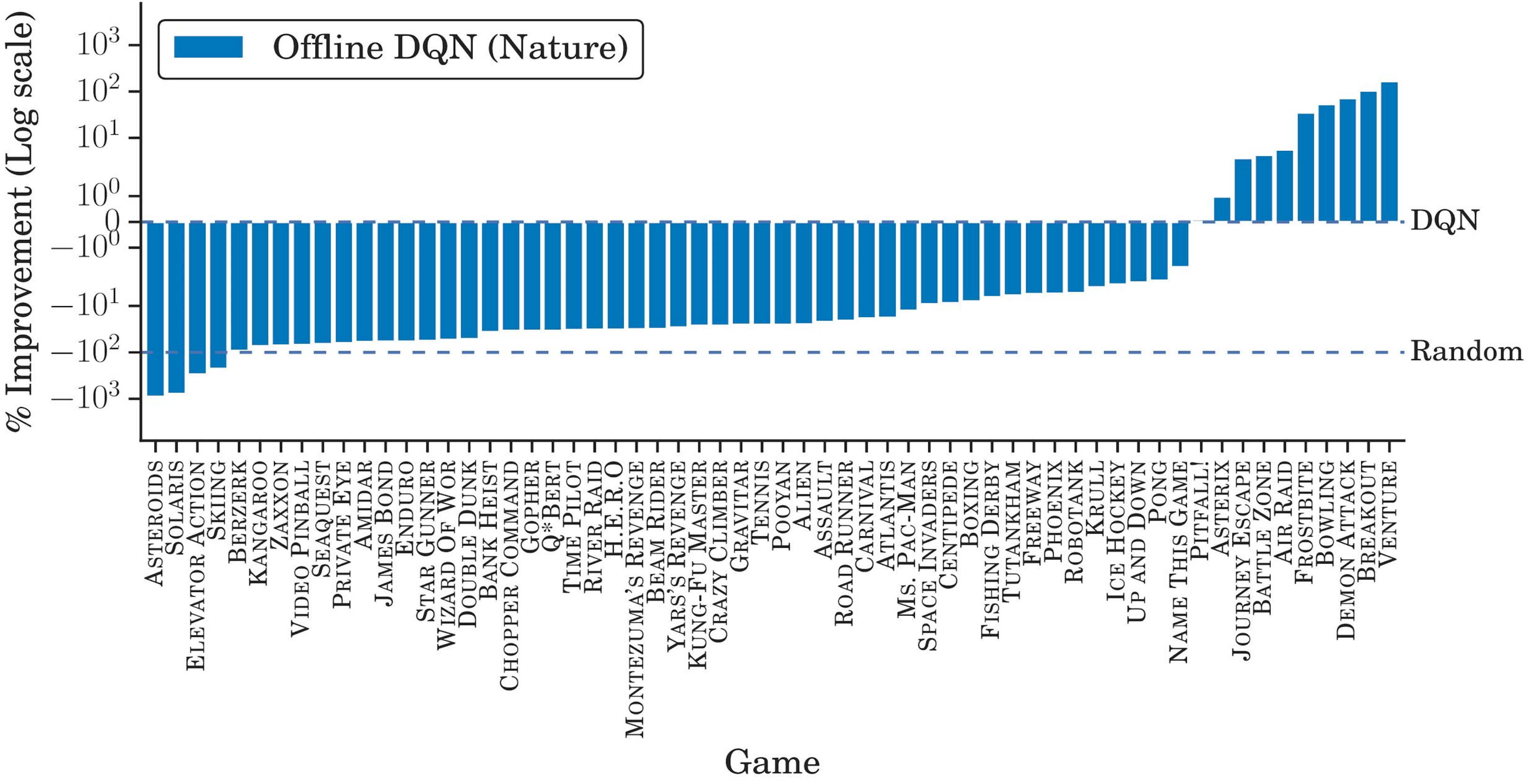
$$Q(s, a; \theta) := \mathbb{E}[Z] = \frac{1}{K} \sum_{i=1}^{K} \theta_i(s, a)$$

Google Research

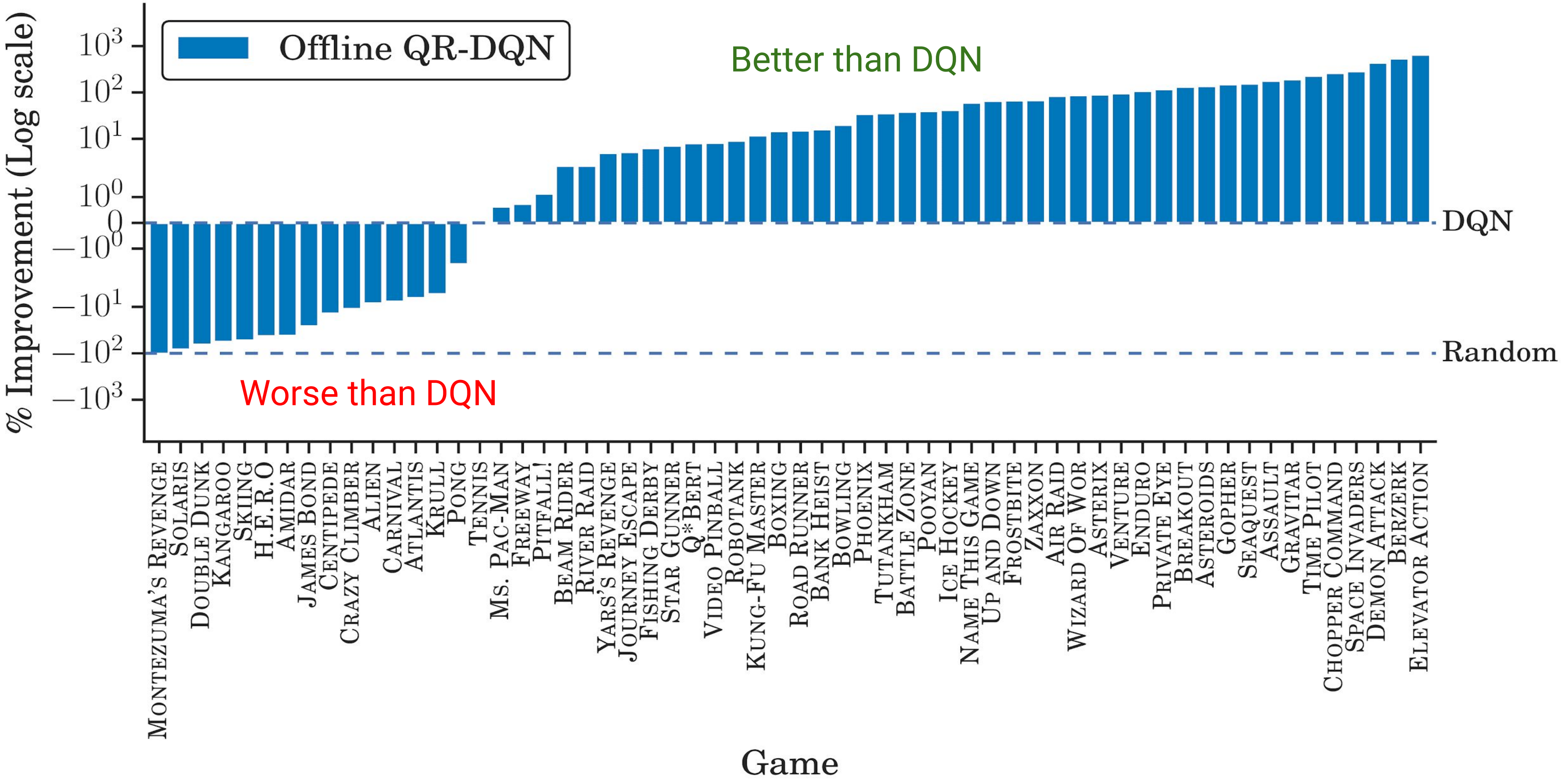
Does Offline QR-DQN work?



Does Offline DQN work?



Does Offline QR-DQN work?



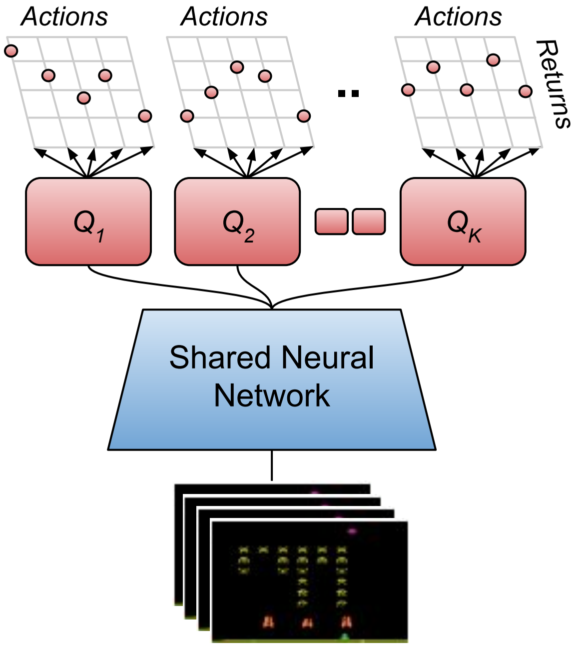
Developing Robust Offline RL algorithms

- Emphasis on Generalization
  - Given a fixed dataset, generalize to unseen states during evaluation.

Developing Robust Offline RL algorithms

- ➢ Emphasis on Generalization
  - Given a fixed dataset, generalize to unseen states during evaluation.
- ➢ Ensemble of Q-estimates:
  - Ensembling, Dropout widely used for improving generalization.

Ensemble-DQN

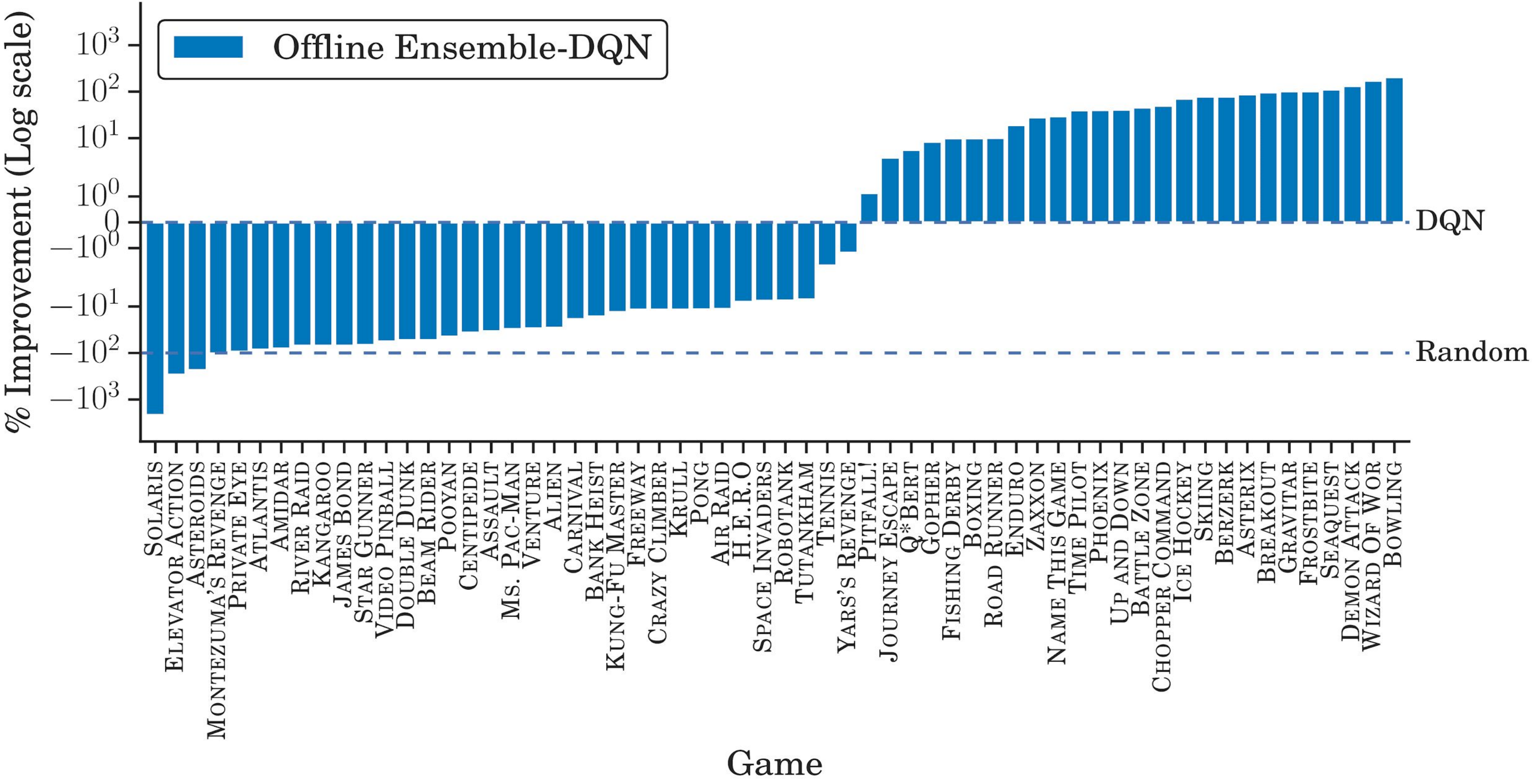


Ensemble-DQN

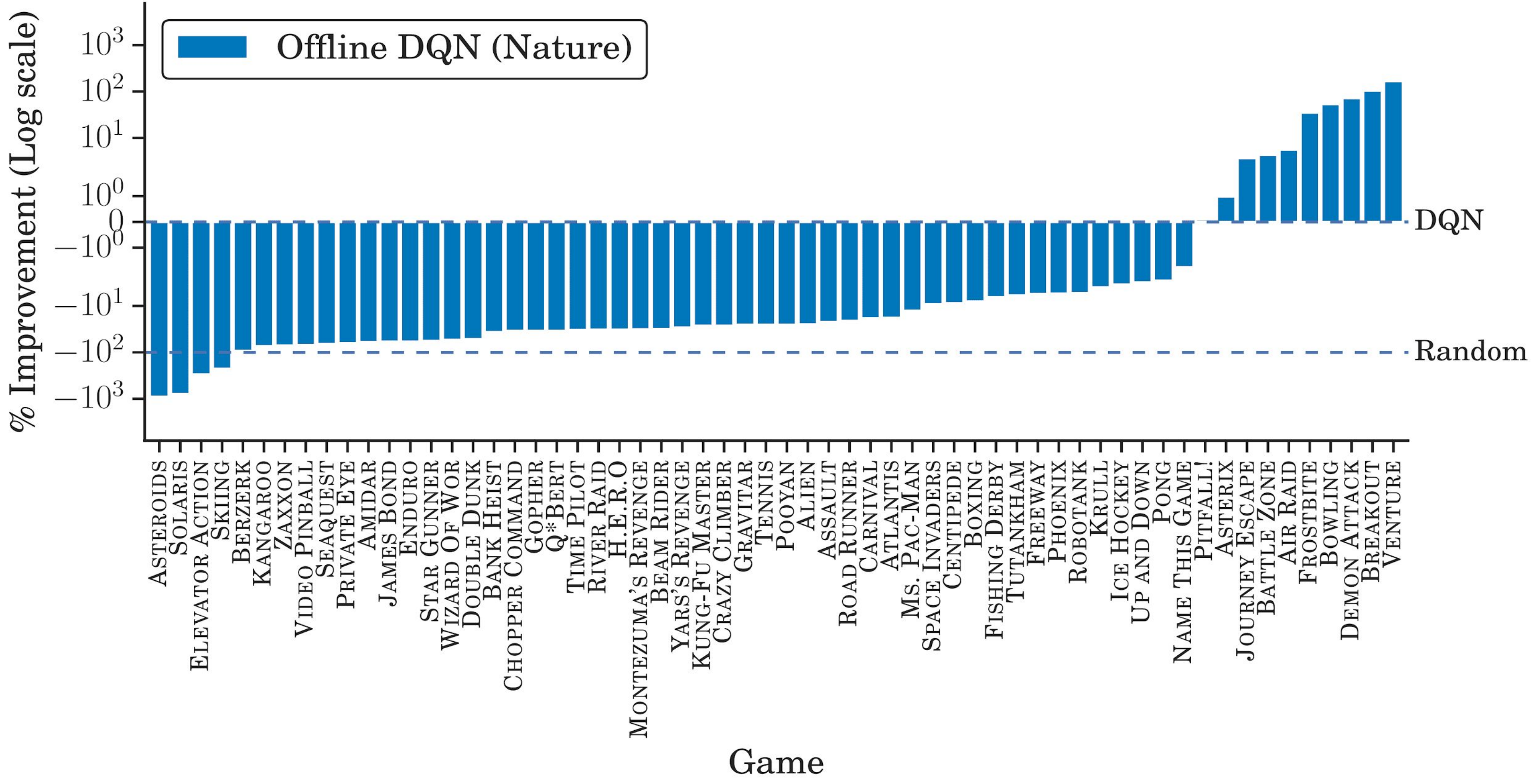
Train multiple (linear) Q-estimates with different random initialization.

Does Offline Ensemble-DQN work?

Google Research



Offline DQN



Offline DQN (Nature)
% Improvement (Log scale)
10^3
10^2
10^1
10^0
0
-10^0
-10^1
-10^2
-10^3
DQN
Random
Asteroids
Solaris
Elevator Action
Skiing
Berzerk
Kangaroo
Zaxxon
Video Pinball
Seaquest
Private Eye
Amidar
James Bond
Enduro
Star Gunner
Wizard Of Wor
Double Dunk
Bank Heist
Chopper Command
Gopher
Q*Bert
Time Pilot
River Raid
H.E.R.O
Montezuma's Revenge
Beam Rider
Yars's Revenge
Kung-Fu Master
Crazy Climber
Gravitar
Tennis
Pooyan
Alien
Assault
Road Runner
Carnival
Atlantis
Ms. Pac-Man
Space Invaders
Centipede
Boxing
Fishing Derby
Tutankham
Freeway
Phoenix
Robotank
Krull
Ice Hockey
Up and Down
Pong
Name This Game
Pitfall!
Asterix
Journey Escape
Battle Zone
Air Raid
Frostbite
Bowling
Demon Attack
Breakout
Venture
Game

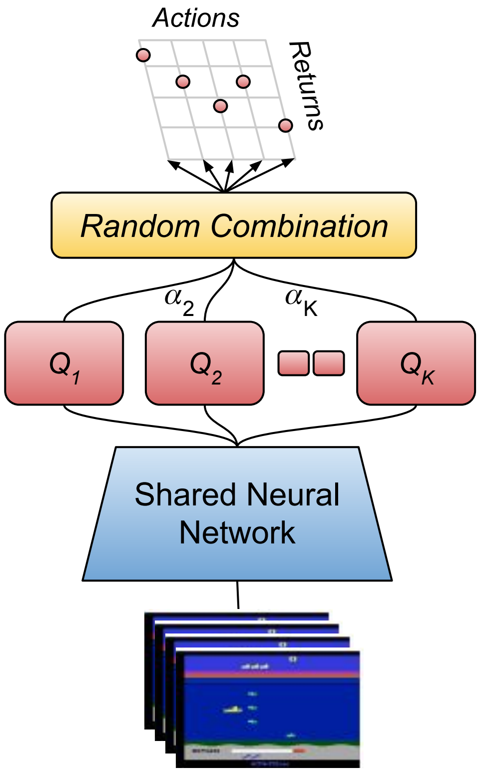
Developing Robust Offline RL algorithms

Google Research

- Emphasis on Generalization
  - Given a fixed dataset, generalize to unseen states during evaluation.
- Q-learning as constraint satisfaction:
  - $\forall\ (s, a, s', r) :\ Q^*(s, a) = r\ + max_{a'}\ Q^*(s', a')$

Random Ensemble Mixture (REM)

Google Research

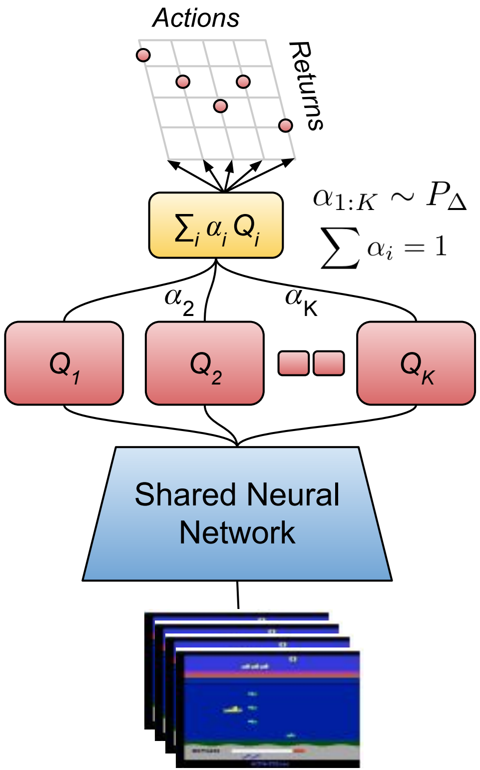


REM

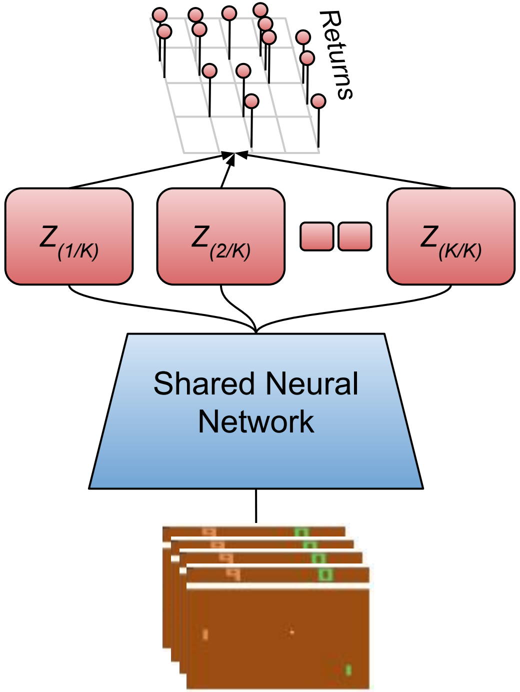
Minimize TD error on random (per minibatch) convex combination of multiple *Q*-estimates.

REM vs QR-DQN

Google Research

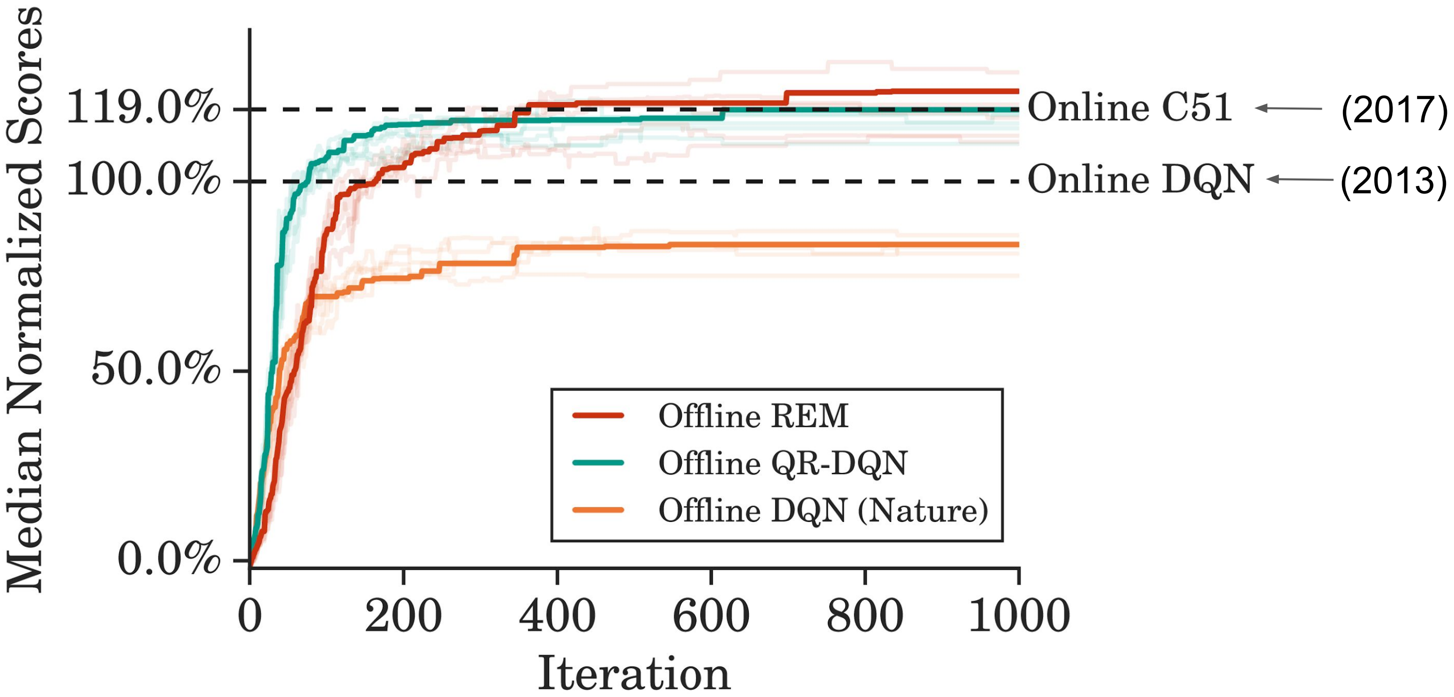


REM



QR-DQN

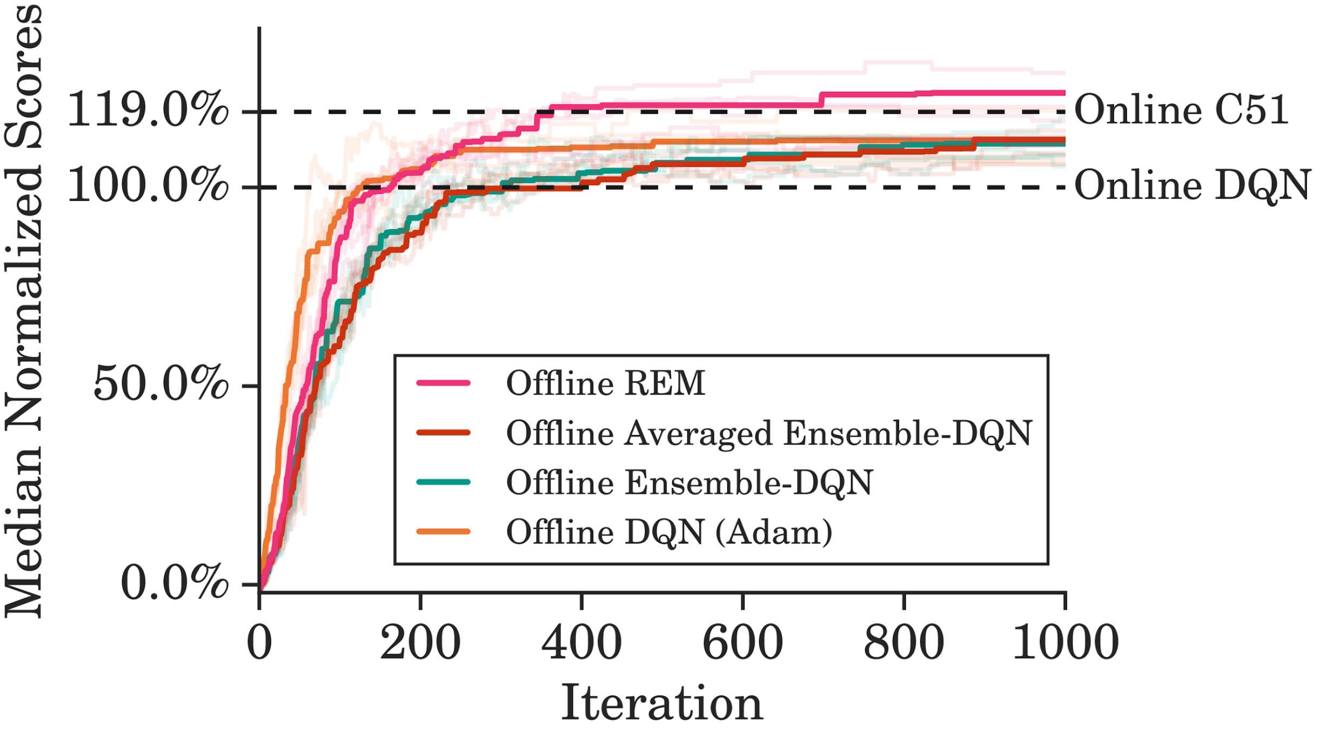
Offline Stochastic Atari Results



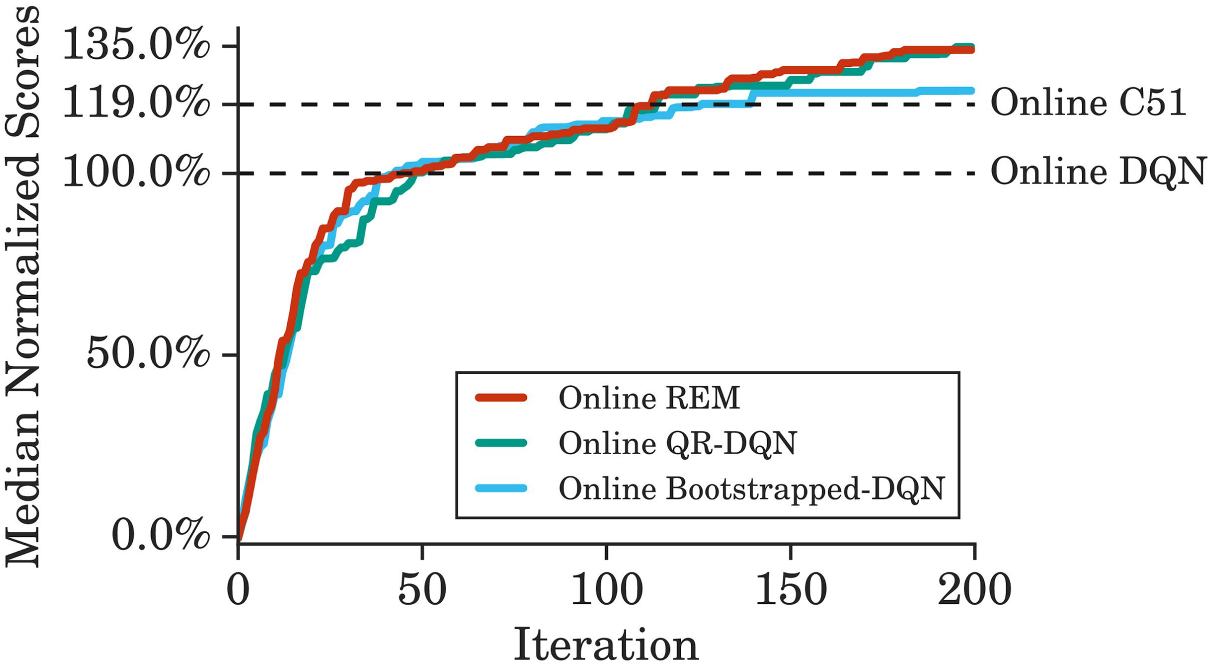
Scores averaged over 5 runs of offline agents trained using DQN replay data across 60 Atari games for 5X gradient steps. Offline REM surpasses gains from online C51 and offline QR-DQN.

Google Research

Offline REM *vs.* Baselines



Does Online REM work?



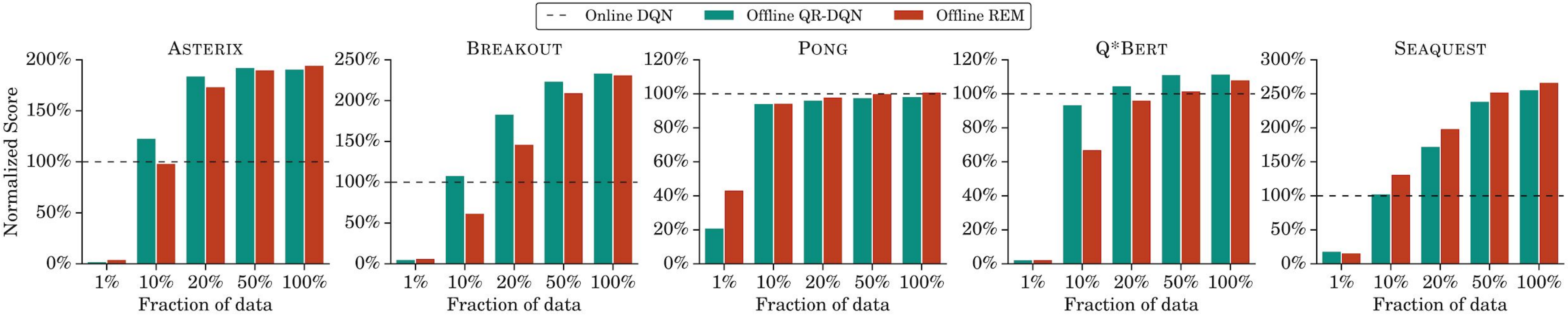
Average normalized scores of online agents trained for 200 million game frames. Multi-network REM with 4 Q-functions performs comparably to QR-DQN.

Google Research

Important Factors in Offline RL

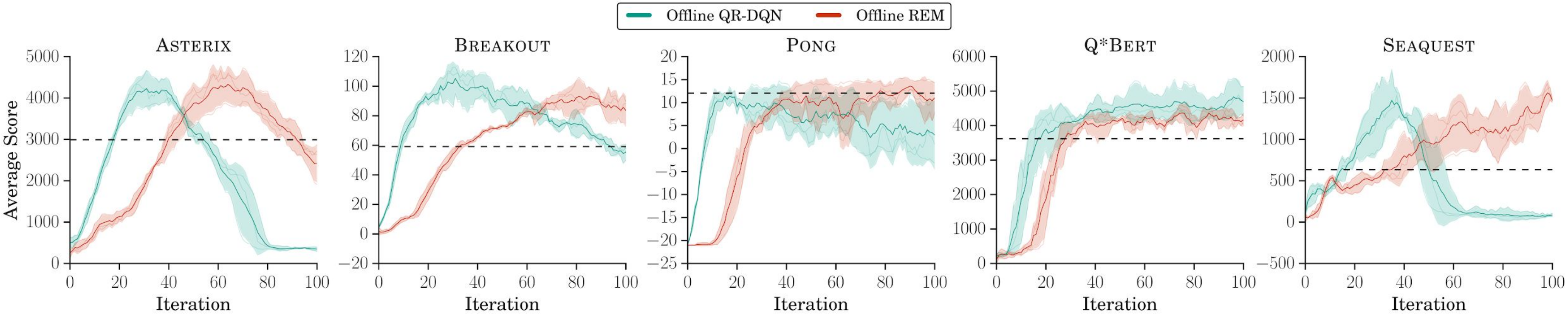
Key Factor in Success: Offline Dataset Size

Google Research



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

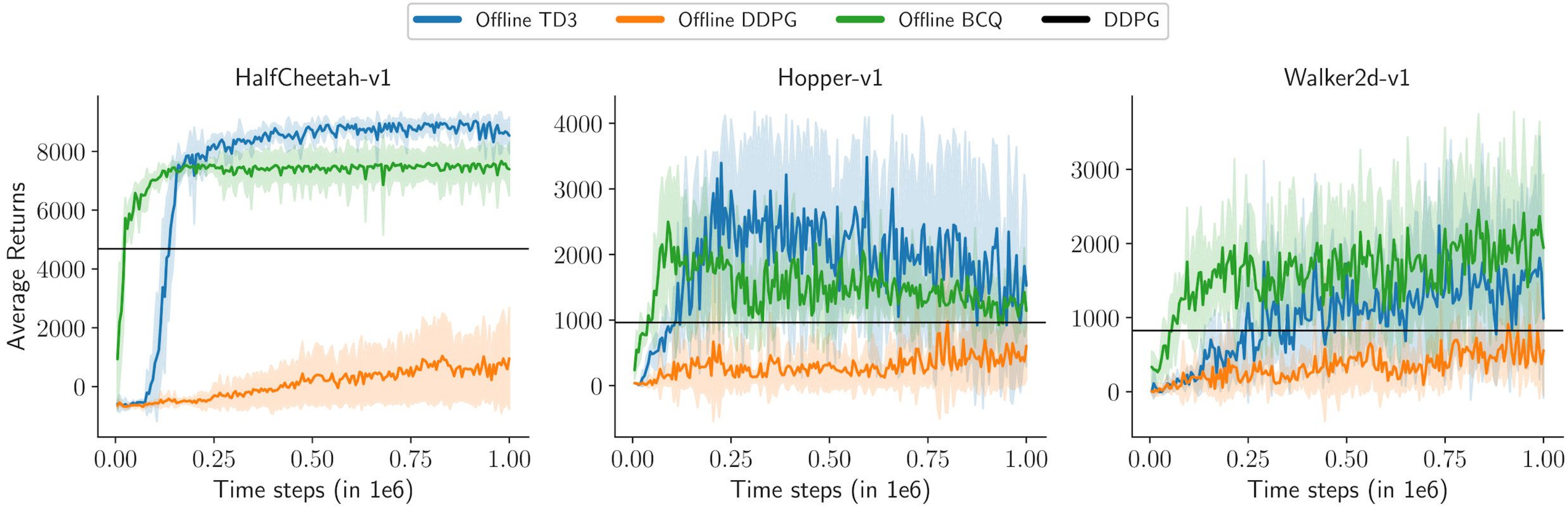
Key Factor in Success: Offline Dataset Diversity



Subsample first 10% of total frames (20 million) for offline training -- much lower quality data.

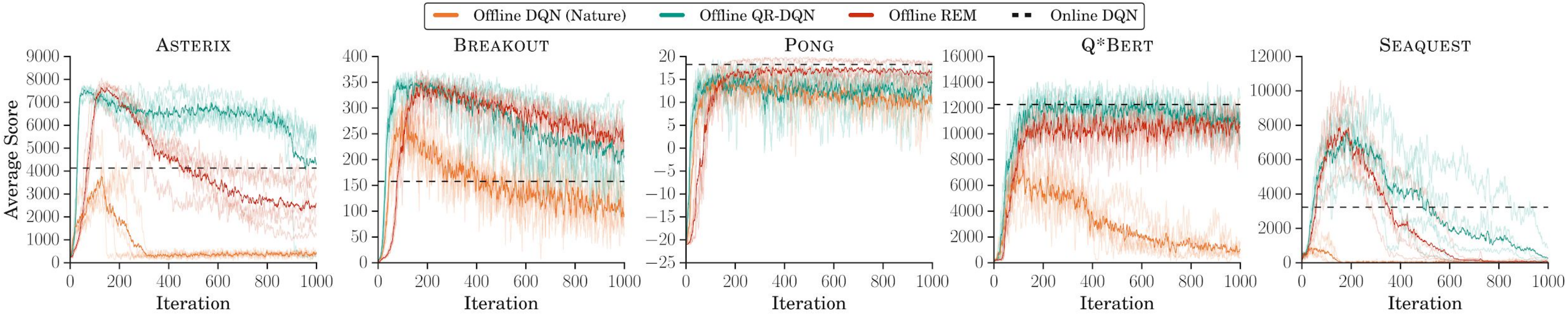
Choice of Algorithm: Offline Continuous Control

Google Research



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

Overfitting in Offline RL: Number of Gradient Updates



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

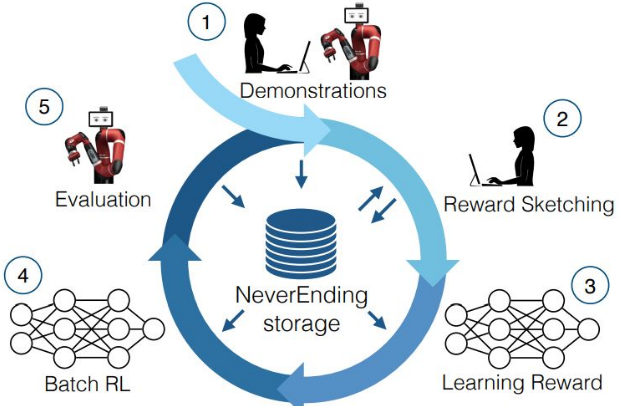
Need for early stopping / better regularization methods

Offline RL for Robotics

Scaling data-driven robotics with reward sketching and batch reinforcement learning

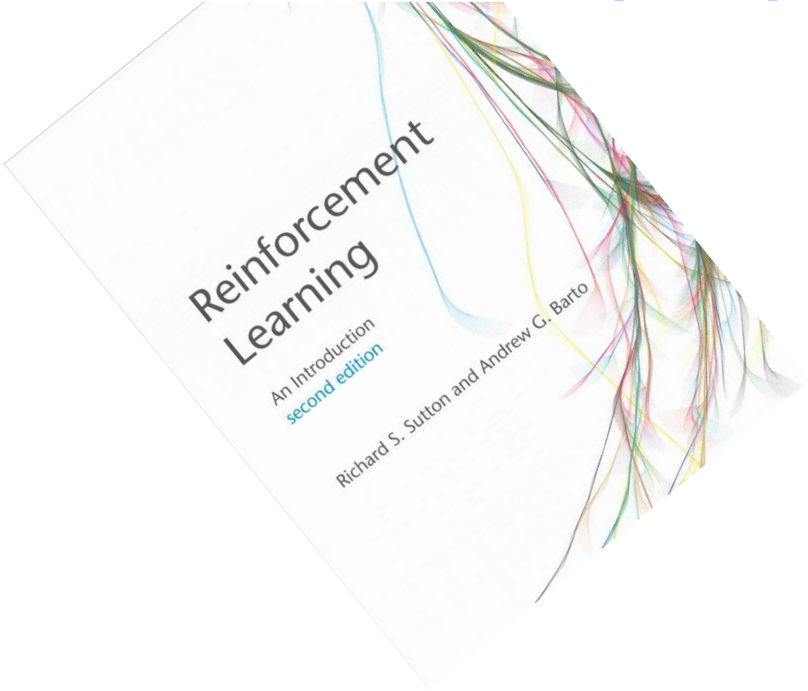
Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Żołna, Yusuf Aytar, David Budden, Mel Vecerik, Oleg Sushkov, David Barker, Jonathan Scholz, Misha Denil, Nando de Freitas, Ziyu Wang

***Abstract*—By harnessing a growing dataset of robot experience, we learn control policies for a diverse and increasing set of related manipulation tasks. To make this possible, we introduce reward sketching: an effective way of eliciting human preferences to learn the reward function for a new task. This reward function is then used to retrospectively annotate all historical data, collected for different tasks, with predicted rewards for the new task. The resulting massive annotated dataset can then be used to learn manipulation policies with batch reinforcement learning (RL) from visual input in a completely off-line way, *i*.e. without interaction with the real robot. This approach makes it possible to scale up RL in robotics, as we no longer need to run the robot for each step of learning. We show that the trained batch RL agents, when deployed in real robots, can perform a variety of challenging tasks involving multiple interactions among**



Future Work

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



Offline RL: Future Work

Google Research

- Rigorous characterization of role of generalization in offline RL

Offline RL: Future Work

Google Research

- Rigorous characterization of role of generalization in offline RL
- Benchmarking with various data collection strategies
  - Subsampling **DQN Replay Dataset** (*e.g.*, first / last *k* million frames)

Offline RL: Future Work

- Rigorous characterization of role of generalization in offline RL
- Benchmarking with various data collection strategies
  - Subsampling DQN-replay datasets (*e.g.*, first / last *k* million frames)
- **Offline Evaluation / Hyperparameter Tuning**

Google Research

Offline RL: Future Work

- Rigorous characterization of role of generalization in offline RL
- Benchmarking with various data collection strategies
  - Subsampling DQN-replay datasets (*e.g.*, first / last *k* million frames)
- Offline Evaluation / Hyperparameter Tuning
- Self-supervised / Model-based RL approaches

Offline RL: Future Work

- Rigorous characterization of role of generalization in offline RL
- Benchmarking with various data collection strategies
  - Subsampling DQN-replay datasets (*e.g.*, first / last *k* million frames)
- Offline Evaluation / Hyperparameter Tuning
- Self-supervised / Model-based RL approaches
- Combining REM with behavior regularization (BCQ, SPIBB, CQL *etc.*)

Google Research

TL;DR

- Standard RL algorithms (*e.g.* REM, QR-DQN), trained on sufficiently large and diverse datasets, perform quite well in the offline setting.

- Offline RL provides a standardized setup for:
  - Isolating *exploitation* from exploration
  - Developing *sample efficient* and *stable algorithms*
  - Pretrain RL agents on logged data

Thank you!

Code, dataset, blog and paper at

offline-rl.github.io