Striving for Simplicity in Off-Policy Deep Reinforcement Learning

Rishabh Agarwal, Dale Schuurmans, Mohammad Norouzi

Google AI

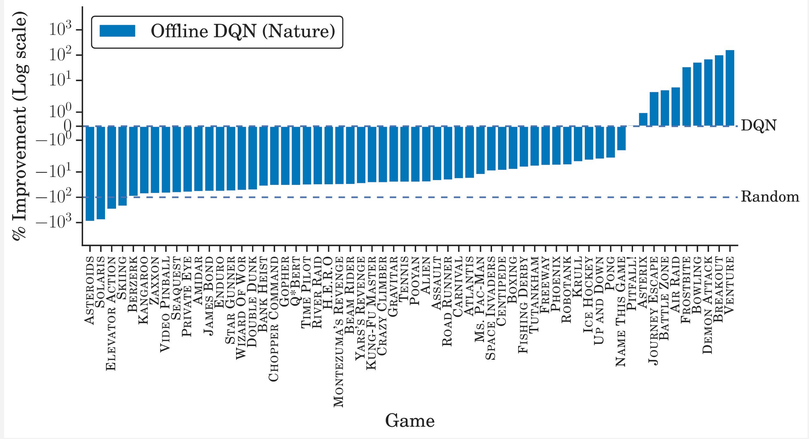
Motivation

1. Is it possible to train successful RL agents **solely based on** diverse **offline data**?
2. Can one design **simple alternatives** to **intricate RL** algorithms?
3. Are the insights gained from the offline setting **beneficial** in an **online** setting?

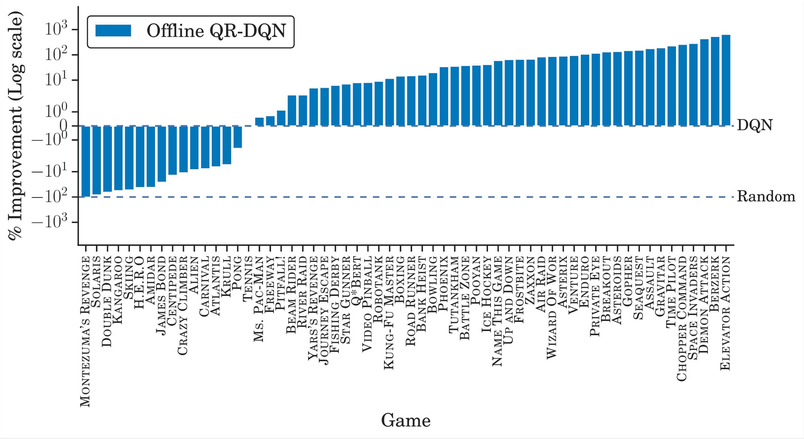
Offline RL on Atari games

1. Train a **Nature DQN** agent on Atari 2600 games with sticky actions for 200 million frames (standard protocol)
2. Save all *(observation, action, reward, next observation)* tuples encountered during training to fixed dataset $\boldsymbol{B}_{\text{DQN}}$
3. Train *off-policy* agents (*e.g.*, DQN) **offline** using $\boldsymbol{B}_{\text{DQN}}$ without interacting with the environment

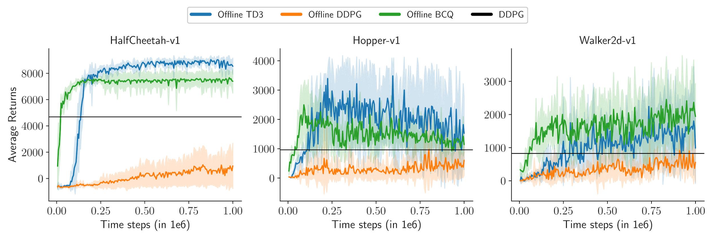
Does offline DQN work?



Let's try offline Distributional RL!



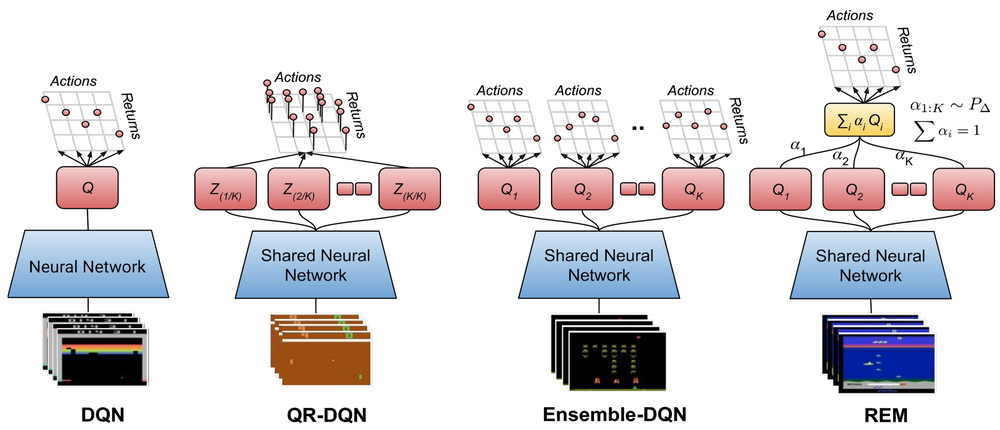
Offline Continuous Control Experiments



Offline TD3 significantly outperforms the data collecting DDPG agent as well as offline DDPG.

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

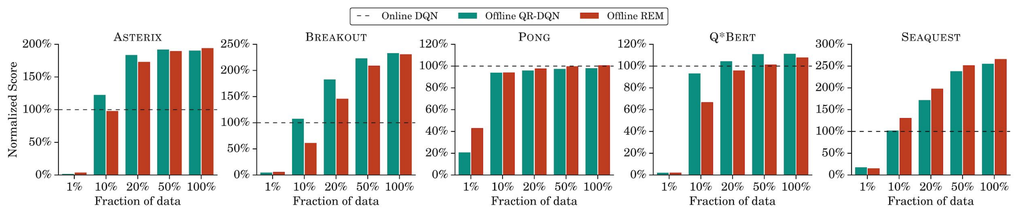
Seeking Simpler Off-Policy RL Algorithms



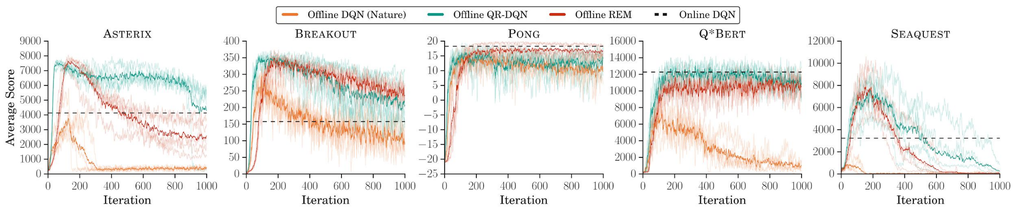
Neural network architectures for DQN, distributional QR-DQN and the proposed variants, i.e., Ensemble-DQN and REM, with the multi-head Q-network.

REM trains randomly sampled convex combination of multiple Q-value estimates.

Future Challenges: Sample Efficiency and Stability

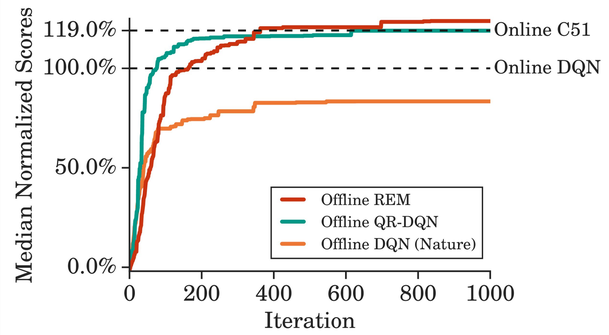


Normalized performance using only a fraction of 200 million frames in DQN replay data $\boldsymbol{B}_{\text{DQN}}$ obtained via random subsampling.

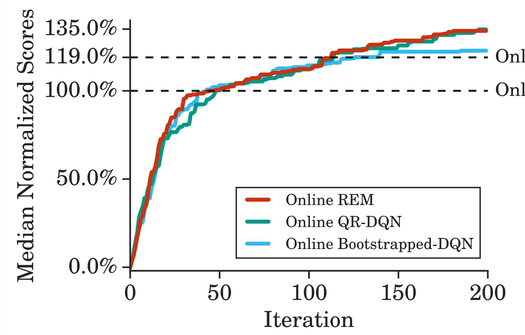


Average online scores of offline agents trained on 5 games using logged DQN replay data for 5X gradient steps compared to online DQN.

Atari Results



***Offline Results**. Normalized scores averaged over 5 runs of offline agents trained using DQN replay data across 60 Atari games for 5X gradient steps. Offline REM outperforms C51 and offline QR-DQN.*



***Online 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.*

Conclusion

In addition to being important for real-world applications, offline RL provides a simple and reproducible experimental setup for:

- Segregating *exploration* and *exploitation*
- Developing *simple* and *effective* off-policy algorithms (*e.g.*, REM)
- Studying and improving *sample efficiency* and *stability* of *off-policy* algorithms

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