How I Learned To Stop Worrying And Love Offline RL
An Optimistic Perspective on Offline Reinforcement Learning
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?

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

this is done many times
An Optimistic Perspective on Offline Reinforcement Learning

RL for Real-World: RL with Large Datasets

RL for Real-World: RL with Large Datasets

An Optimistic Perspective on Offline Reinforcement Learning

RL for Real-World: RL with Large Datasets

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/](https://bair.berkeley.edu/blog/2019/12/05/bear/)
Offline RL: A Data-Driven RL Paradigm

Offline RL can help:

- Pretrain agents on existing logged data.

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

Image Source: Data-Driven Deep Reinforcement Learning, BAIR Blog. [https://bair.berkeley.edu/blog/2019/12/05/bear/](https://bair.berkeley.edu/blog/2019/12/05/bear/)
Offline RL: A Data-Driven RL Paradigm

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.

Image Source: Data-Driven Deep Reinforcement Learning, BAIR Blog. https://bair.berkeley.edu/blog/2019/12/05/bear/
But .. Offline RL is Hard!

You Shall Not

Explore

NO new corrective feedback!
But .. Offline RL is Hard!

Requires Counterfactual Generalization
But .. Offline RL is Hard!

**Fully Off-Policy**

- **Bootstrapping**
  - (Learning guess from a guess)

- **Function Approximation**
Standard RL fails in Offline setting..

**Off-Policy Deep Reinforcement Learning without Exploration**

Scott Fujimoto, David Meger, Doina Precup

**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 standard RL fails in the offline setting.

**Behavior Regularized Offline Reinforcement Learning**

Yifan Wu, George Tucker, Ofir Nachum

**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.

**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
(siegeln1@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.

**Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction**

Aviral Kumar, Justin Fu

**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 actor-critic algorithms suffer from instability issues.

An Optimistic Perspective on Offline Reinforcement Learning
Standard RL fails in Offline setting..

Off-Policy Deep Reinforcement Learning without Exploration

Scott Fujimoto \textsuperscript{12} David Meger \textsuperscript{12} Doina Precup \textsuperscript{12}

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 the requirement on high-quality data, off-policy reinforcement learning algorithms often fail in Offline setting.

Behavior Regularized Offline Reinforcement Learning

Yifan Wu
Carnegie Mellon University
yw40@cs.cmu.edu

George Tucker
Google Research
gjt@google.com

Ofir Nachum
Google Research
ofir@nachum@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.

Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction

Aviral Kumar
UC Berkeley
aviralx@berkeley.edu

Justin Fu
UC Berkeley
justinfu@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 actor-critic suffer from high variance in estimated values.

OPTIMISTIC PERSPECTIVE ON OFFLINE REINFORCEMENT LEARNING

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

DeepMind
(siegelny@deepmind.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 for large-scale systems, but so far has been unexplored in practice. In this paper, we develop offline reinforcement learning algorithms that are provably sample-efficient and are shown to converge to optimal policies. We validate our method on a number of continuous control tasks, where we demonstrate significant savings in the required number of samples compared to state-of-the-art methods.
Standard RL fails in Offline setting ..
Can standard off-policy RL succeed in the offline Setting?
Offline RL on Atari 2600

Train 5 DQN (Nature) agents on each Atari game using sticky actions (stochasticity)

200 million frames (standard protocol)
Offline RL on Atari 2600

Save all of the tuples of (observation, action, next observation, reward) encountered to DQN-replay dataset(s)
Offline RL on Atari 2600

Train off-policy agents using DQN-replay dataset(s) without any further environment interaction
Does Offline DQN work?
Let's try recent off-policy algorithms!

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

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

\[
Q(s, a; \theta) := \mathbb{E}[Z] = \frac{1}{K} \sum_{i=1}^{K} \theta_i(s, a)
\]
An Optimistic Perspective on Offline Reinforcement Learning

Does Offline QR-DQN work?
An Optimistic Perspective on Offline Reinforcement Learning

Does Offline DQN work?

![Graph showing the performance of Offline DQN and DQN in various games. The x-axis represents different games, and the y-axis represents the percentage improvement on a log scale. The graph compares Offline DQN (Nature) and DQN, with Offline DQN generally performing better than DQN.](image-url)
Offine DQN (Nature) vs Offline C51

Average online scores of C51 and DQN (Nature) agents trained offline on DQN replay dataset for the same number of gradient steps as online DQN. The horizontal line shows the performance of fully-trained DQN.
An Optimistic Perspective on Offline Reinforcement Learning

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.
An Optimistic Perspective on Offline Reinforcement Learning

Ensemble-DQN

Train multiple (linear) $Q$-heads with different random initialization.
Does Offline Ensemble-DQN work?

![Graph showing improvement in game performance for Offline Ensemble-DQN compared to DQN and Random strategies.](image)

An Optimistic Perspective on Offline Reinforcement Learning
Does Offline DQN work?
Developing Robust Offline RL algorithms

➢ 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)

Minimize TD error on random (per minibatch) convex combination of multiple Q-estimates.
REM vs QR-DQN

REM

\[
\sum_{i} \alpha_i Q_i
\]

\[
\sum \alpha_i = 1
\]

\[
\alpha_{1:K} \sim P_{\Delta}
\]

QR-DQN

Shared Neural Network

\[
Z_{(1/K)} \quad Z_{(2/K)} \quad Z_{(K/K)}
\]
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.
Offline REM vs. Baselines

- Offline REM
- Offline Averaged Ensemble-DQN
- Offline Ensemble-DQN
- Offline DQN (Adam)

An Optimistic Perspective on Offline Reinforcement Learning
Reviewers asked: 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.
An Optimistic Perspective on Offline Reinforcement Learning

Key Factor in Success: Offline Dataset Size

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

Divergence with 1% of data for prolonged training!
Key Factor in Success: Offline Dataset Composition

Subsample first 10% of total frames (20 million) for offline training -- much lower quality data.
Choice of Algorithm: Offline Continuous Control

Offline agents trained using full experience replay of DDPG on MuJoCo environments.
Offline RL: 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 eventually degrade performance :(}

An Optimistic Perspective on Offline Reinforcement Learning
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 Žohna, 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 rigid or deformable objects. Moreover, they display a significant
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

- Rigorous characterization of role of generalization in offline RL
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 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
  - Currently, online evaluation used for early stopping. “True” offline RL requires offline policy evaluation.
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
  - Currently, online evaluation used for early stopping. “True” offline RL require offline policy evaluation.

- Model-based RL approaches
TL;DR

- Robust 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!

For code, DQN-replay dataset(s) and previous version of paper, refer to offline-rl.github.io