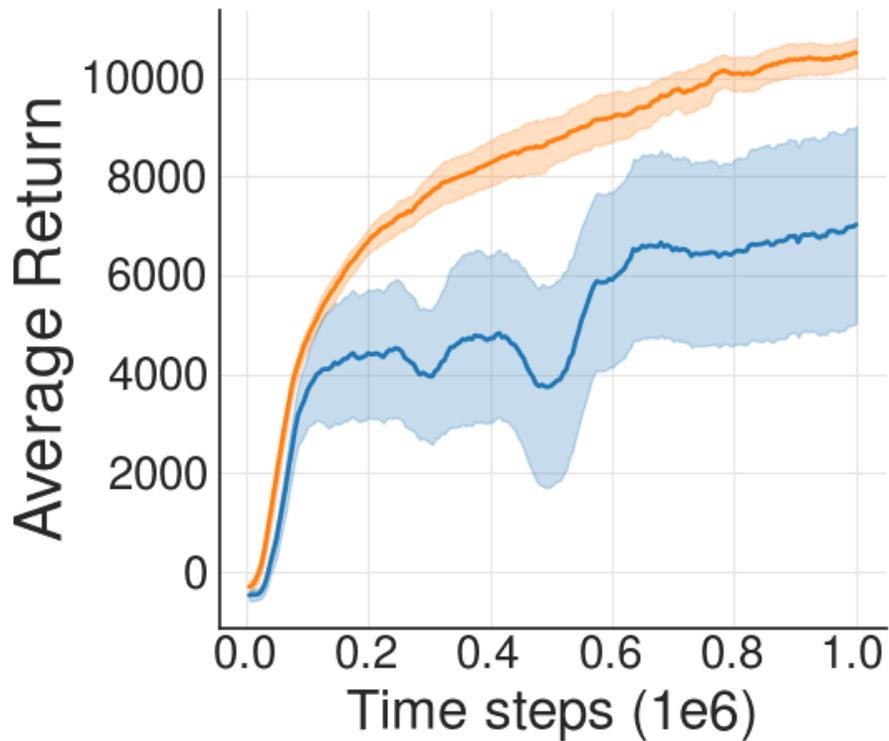


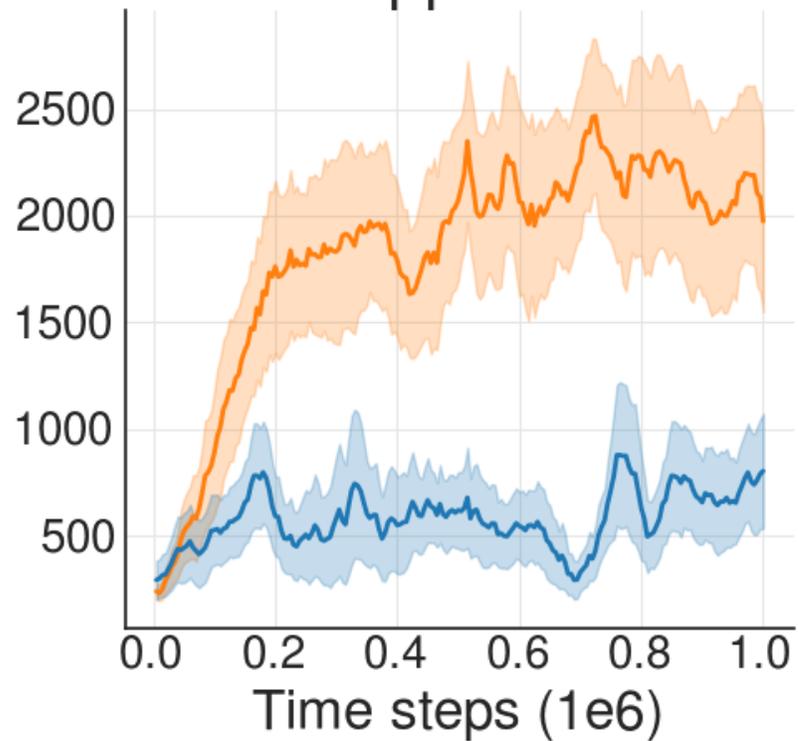
Off-Policy Deep Reinforcement Learning without Exploration

Scott Fujimoto, David Meger, Doina Precup
Mila, McGill University

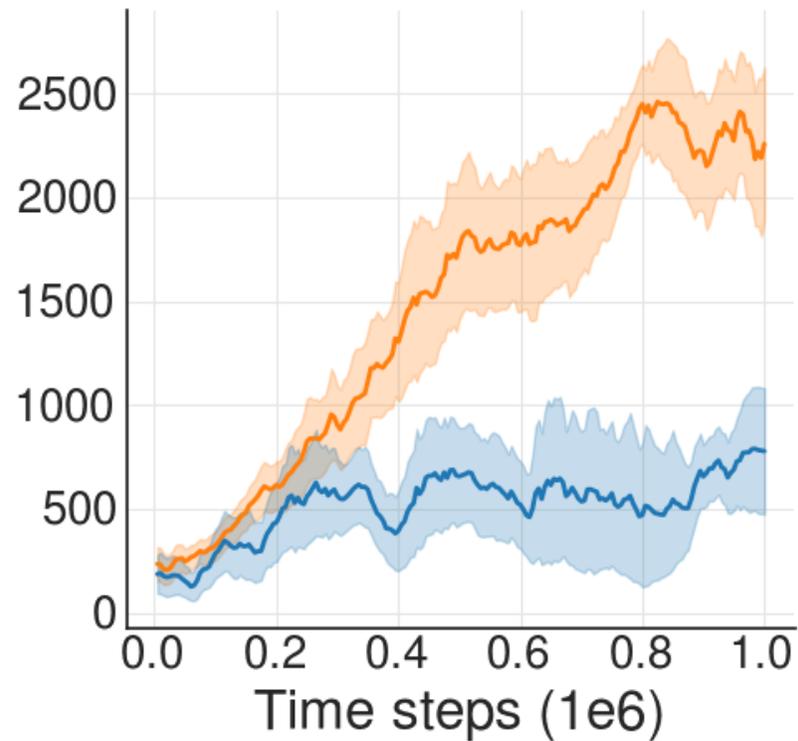
HalfCheetah-v1



Hopper-v1



Walker2d-v1



Surprise!

Agent orange and agent blue are trained with...

1. **The same off-policy algorithm (DDPG).**
2. **The same dataset.**

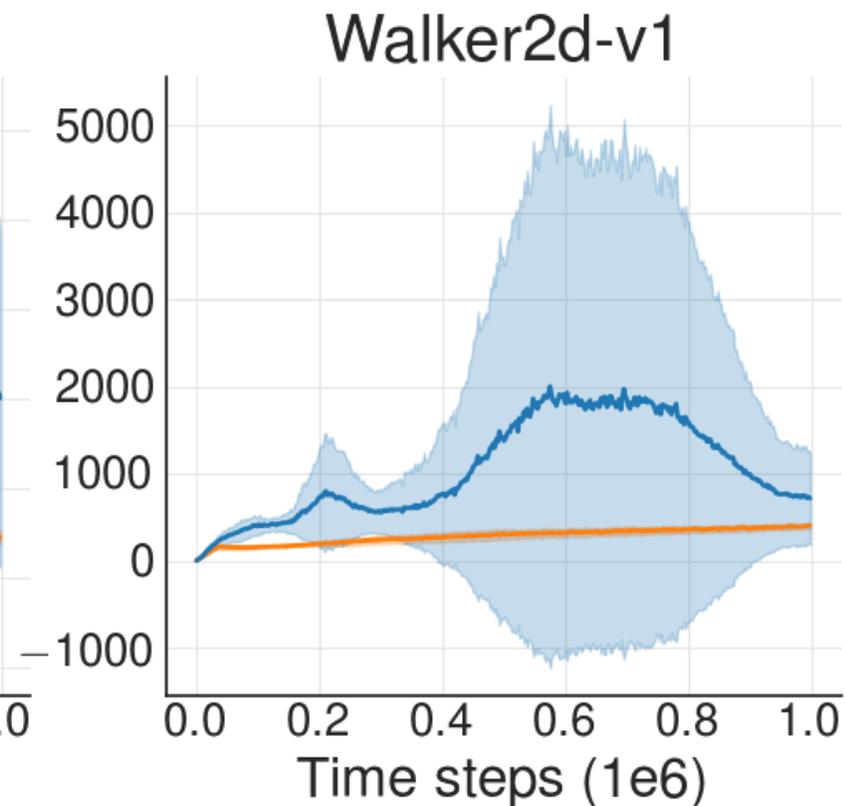
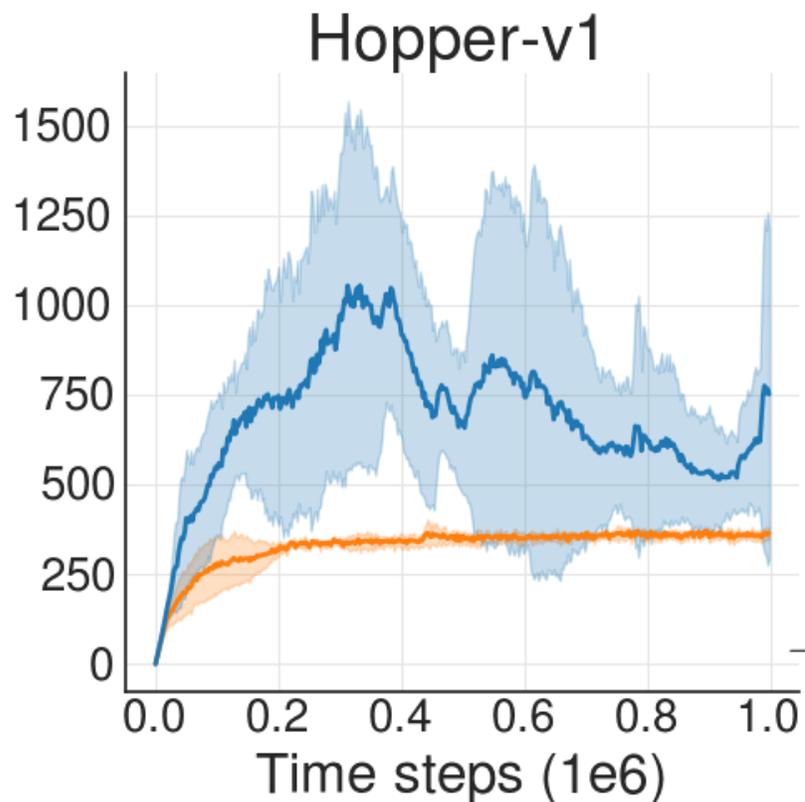
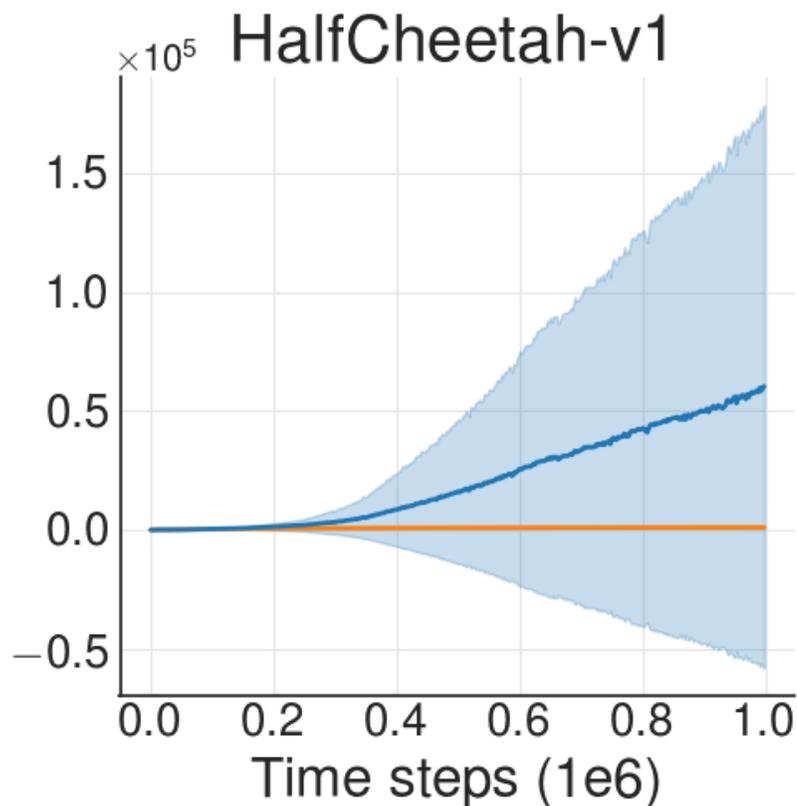
The Difference?

1. **Agent orange:** Interacted with the environment.
 - Standard RL loop.
 - Collect data, store data in buffer, train, repeat.
2. **Agent blue:** Never interacted with the environment.
 - Trained with data collected by agent orange concurrently.

1. Trained with the same off-policy algorithm.
2. Trained with the same dataset.
3. One interacts with the environment. One doesn't.

Off-policy deep RL fails when **truly off-policy**.

Value Predictions



Extrapolation Error

$$Q(s, a) \leftarrow r + \gamma Q(s', a')$$

Extrapolation Error

$$Q(s, a) \leftarrow r + \gamma Q(s', a')$$

The diagram illustrates the Bellman optimality equation $Q(s, a) \leftarrow r + \gamma Q(s', a')$ with annotations. The word "GIVEN" is written in red below the state-action pair (s, a) . Two red arrows point upwards from "GIVEN" to s and a . Another red arrow points from the reward r to the left-hand side of the equation. A second red arrow points from the term $Q(s', a')$ to the left-hand side. The word "GENERATED" is written in blue below the next state-action pair (s', a') . A blue arrow points upwards from "GENERATED" to a' .

Extrapolation Error

$$Q(s, a) \leftarrow r + \gamma Q(s', a')$$

1. $(s, a, r, s') \sim \text{Dataset}$
2. $a' \sim \pi(s')$

Extrapolation Error

$$Q(s, a) \leftarrow r + \gamma Q(s', a')$$

$(s', a') \notin \text{Dataset} \rightarrow Q(s', a') = \mathbf{bad}$

$\rightarrow Q(s, a) = \mathbf{bad}$

Extrapolation Error

$$Q(s, a) \leftarrow r + \gamma Q(s', a')$$

$(s', a') \notin \text{Dataset} \rightarrow Q(s', a') = \mathbf{bad}$

$\rightarrow Q(s, a) = \mathbf{bad}$

Extrapolation Error

$$Q(s, a) \leftarrow r + \gamma Q(s', a')$$

$(s', a') \notin \text{Dataset} \rightarrow Q(s', a') = \mathbf{bad}$

$\rightarrow Q(s, a) = \mathbf{bad}$

Extrapolation Error

Attempting to evaluate π without (sufficient) access to the (s, a) pairs π visits.

Batch-Constrained Reinforcement Learning

Only choose π such that we have access to the (s, a) pairs π visits.

Batch-Constrained Reinforcement Learning

1. $a \sim \pi(s)$ such that $(s, a) \in \text{Dataset}$.
2. $a \sim \pi(s)$ such that $(s', \pi(s')) \in \text{Dataset}$.
3. $a \sim \pi(s)$ such that $Q(s, a)$ is maxed.

Batch-Constrained Deep Q-Learning (BCQ)

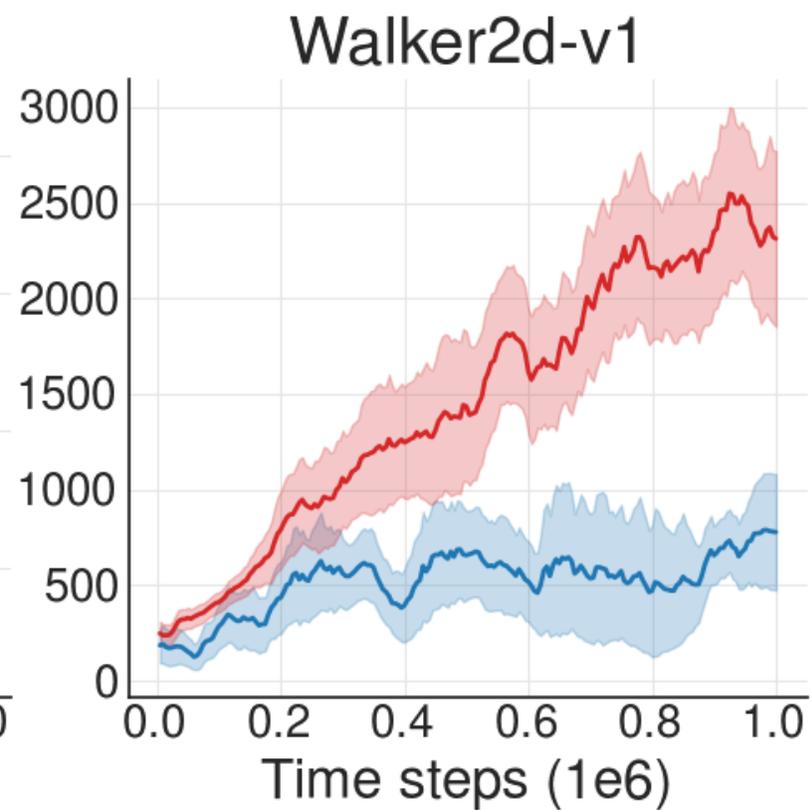
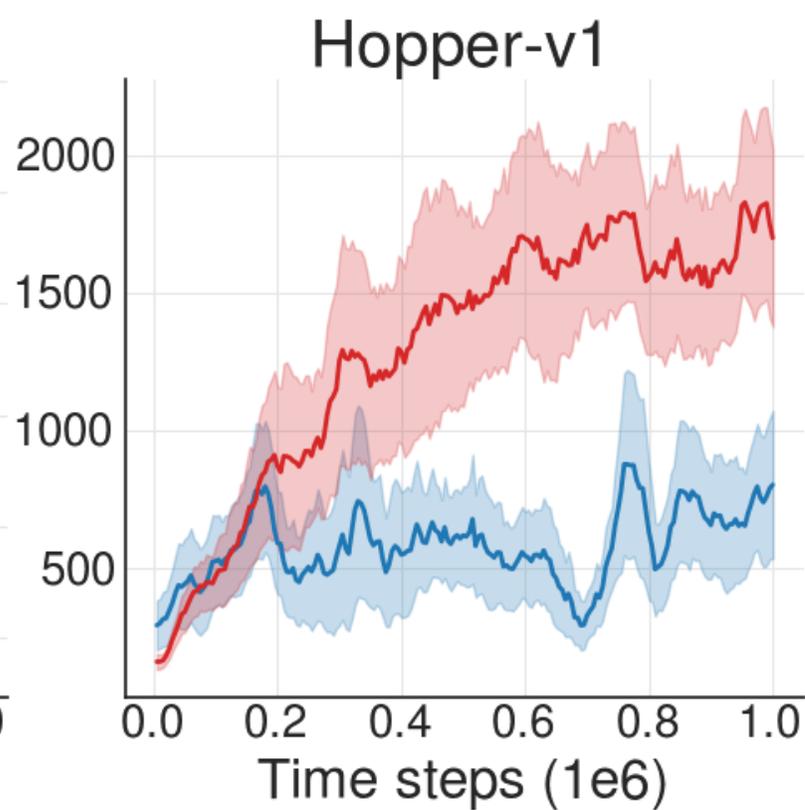
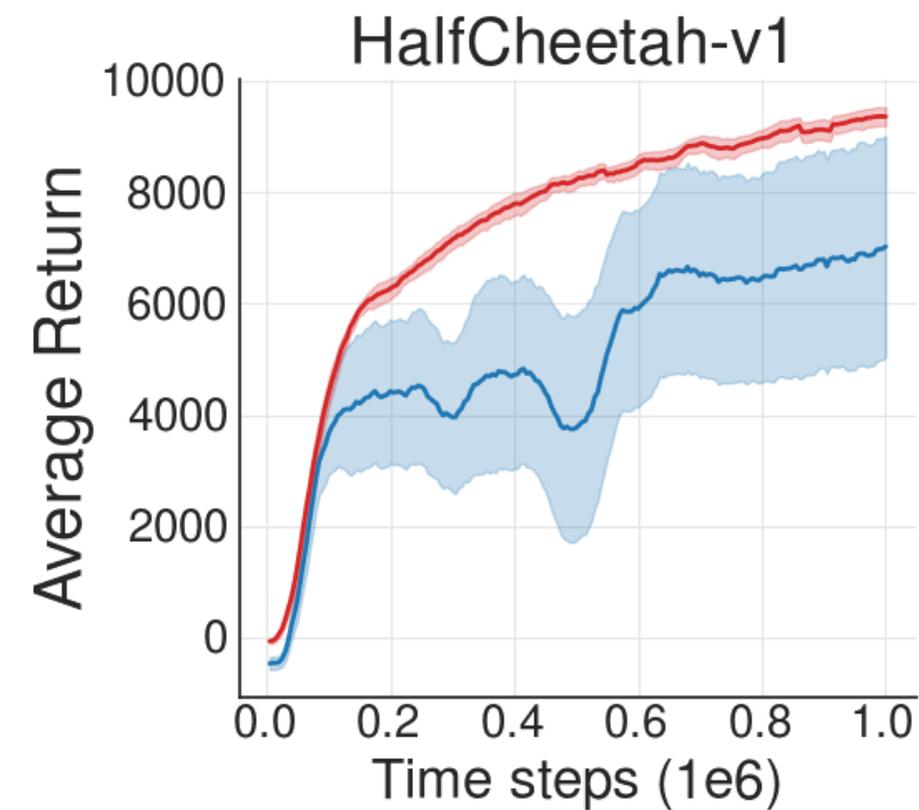
First imitate dataset via generative model:

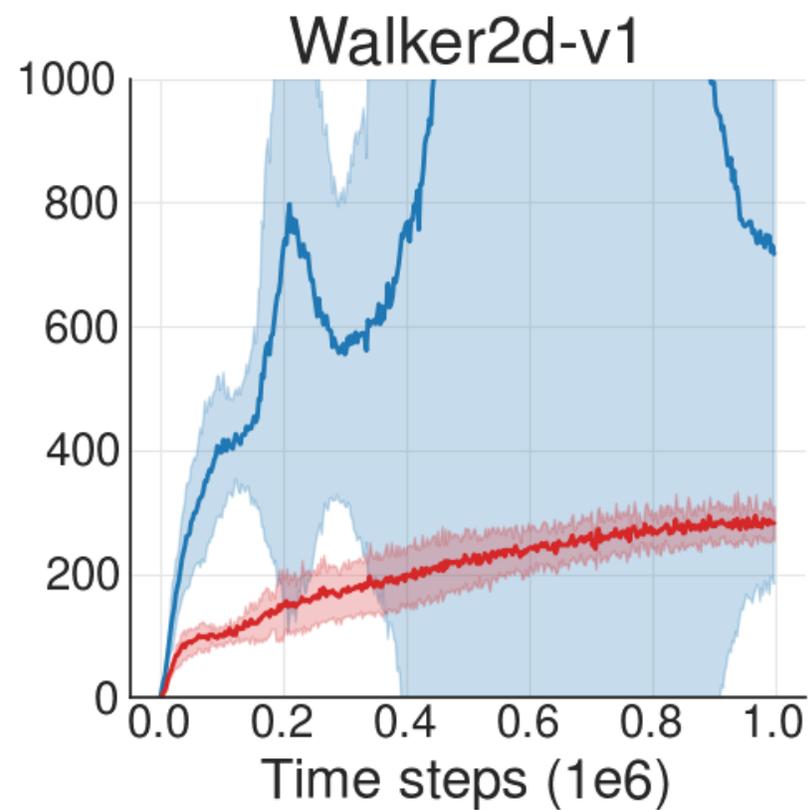
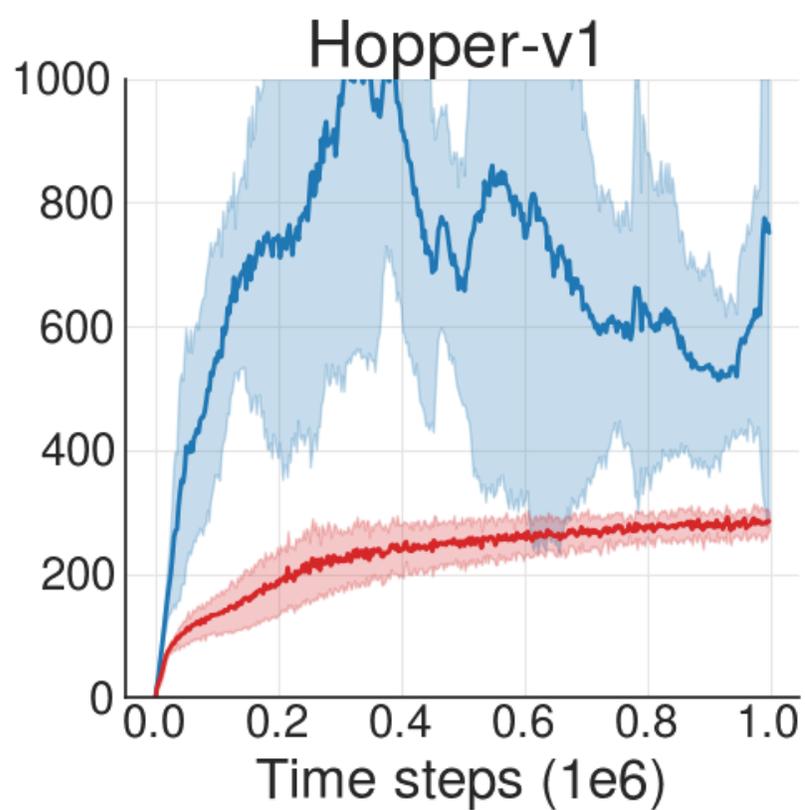
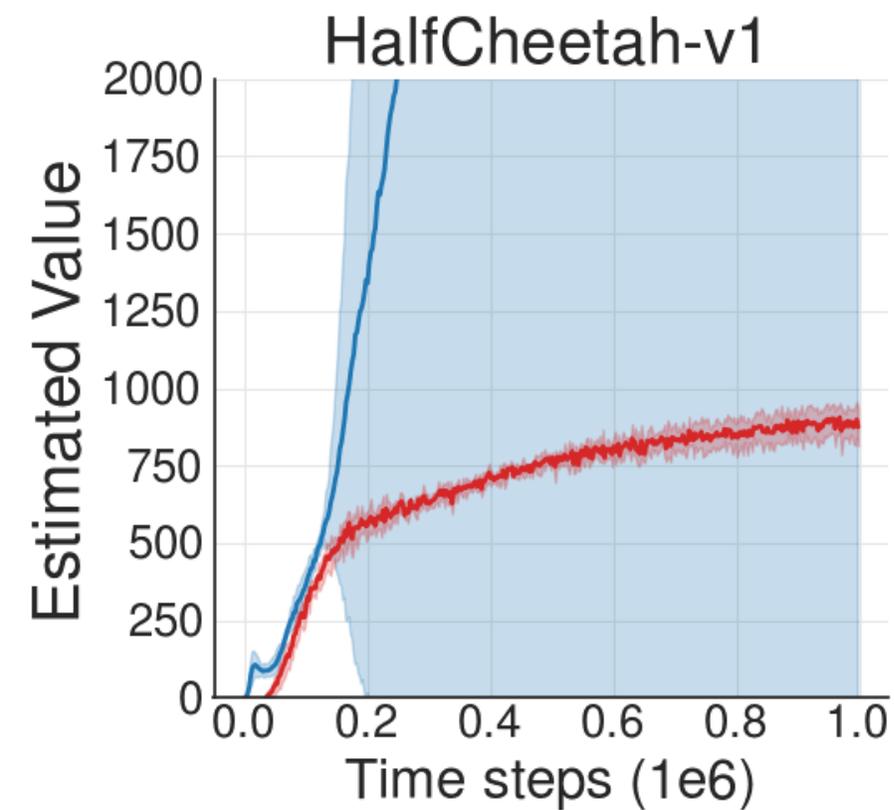
$$G(a|s) \approx P_{Dataset}(a|s).$$

$$\pi(s) = \operatorname{argmax}_{a_i} Q(s, a_i), \text{ where } a_i \sim G$$

(i.e. select the best action that is likely under the dataset)

(+ some additional deep RL **magic**)





Come say Hi
@ Pacific
Ballroom #38
(6:30 Tonight)

<https://github.com/sfujim/BCQ>

Off-Policy Deep Reinforcement Learning without Exploration

Scott Fujimoto, David Meger, Doina Precup

McGill

Abstract

Off-policy deep reinforcement learning algorithms fail in **off-policy settings** where they are unable to interact with the environment (Batch RL).

This failure can be explained by a phenomenon we call *extrapolation error*, which is induced by a mismatch between the dataset and true state-action visitation of the current policy.

We introduce batch-constrained reinforcement learning which restricts the action space in order to force the agent towards behaving close with respect to a subset of the given data.

We present the first continuous control deep reinforcement learning algorithm which can learn effectively from arbitrary, fixed behavior data.

Batch-Constrained Reinforcement Learning

1. Minimize the distance of selected actions to the data $\max_{\pi} \mathbb{E}_{s \sim P_{\pi}}(d(s))$
2. Lead to states where familiar data can be observed. $\max_{\pi} \mathbb{E}_{s, a \sim P_{\pi}}(P_{\pi}(s'))$

Extrapolation Error

Extrapolation error is an error in off-policy value estimation caused by a mismatch between the dataset and true state-action visitation. This is problematic in the Bellman update where the target action a' , without consideration of the dataset.

Absent Data: Attempting to evaluate an out-of-distribution action a' ($s', \pi(s')$). $Q(s', \pi(s'))$ may be arbitrarily bad without sufficient data.

Model Bias: When using data sampled from a batch \mathcal{B} , we have a biased estimate of Bellman update, by taking an expectation over the transitions in the batch \mathcal{B} , rather than the true MDP: $T^{\pi} Q(s, a) \approx \mathbb{E}_{r, s', a'} [r + \gamma Q(s', \pi(s'))]$

Training Mismatch: Transitions are sampled uniformly from the dataset, giving a loss weighted with respect to the likelihood of data in the batch, rather than weighted in regions of interest to the policy.

Experiments

Final buffer: We train a DDPG agent for 1 million time steps, adding $N(0, 0.5)$ Gaussian noise to actions for high exploration, with the aim of sufficient coverage, and use the final replay buffer.

Concurrent: We concurrently train the off-policy and behavioral DDPG agents, for 1 million time steps.

Imitation: A trained DDPG agent acts as an expert, and is used to collect a dataset of 1 million transitions.

Imperfect Demonstrations: An expert is used to collect a dataset of 100k transitions, while selecting actions randomly with probability 0.3 and adding Gaussian noise $N(0, 0.5)$ to the remaining actions.

Off-Policy Experiments

Value Estimates

Speech bubble: I ♥ DQN

(Artist's rendition of poster session)