Semi-Supervised Offline Reinforcement Learning with Action-Free Trajectories

Qinqing Zheng*, Mikael Henaff*, Brandon Amos*, Aditya Grover+

* Meta AI Research
+ UCLA

ICML 2023

UCLA Meta

Learning From Heterogenous Data

Natural Agents learn from multiple data sources that differ in size, quality, and types of measurements

Standard Offline RL learn from homogeneous and completely measured data

Learning From Heterogenous Data

Natural Agents learn from multiple data sources that differ in size, quality, and types of measurements

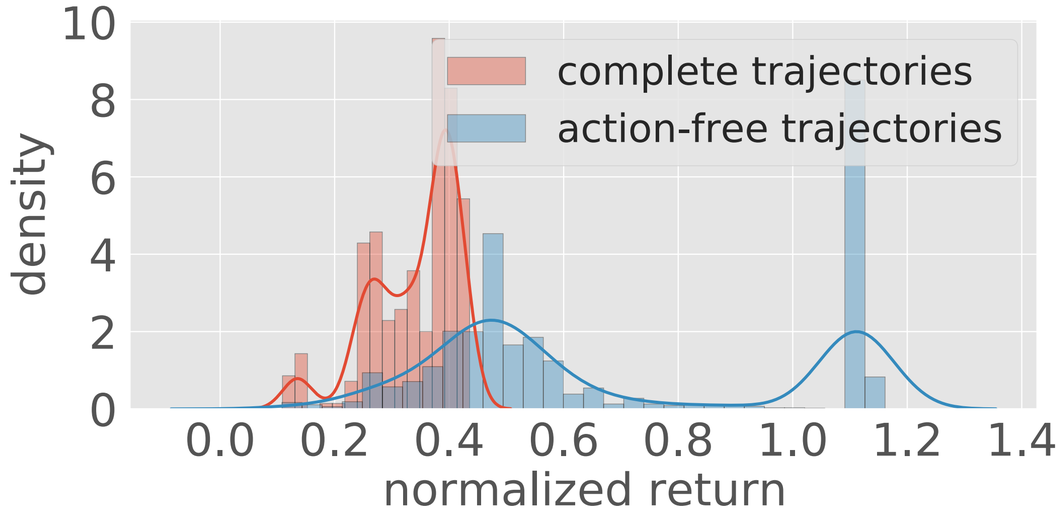
Standard Offline RL learn from homogeneous and completely measured data

Our work studies this heterogeneity in the context of offline RL

- propose a new, practically motivated semi-supervised setting
- propose a simple but highly successful pipeline SS-ORL
- perform large-scale controlled experiment to investigate the interplay of data-centric properties and algorithmic design choices

A Semi-Supervised Offline RL Setting

Offline Dataset = Action-Free Trajectories (unlabelled) ∪ Action-Complete Trajectories (labelled)

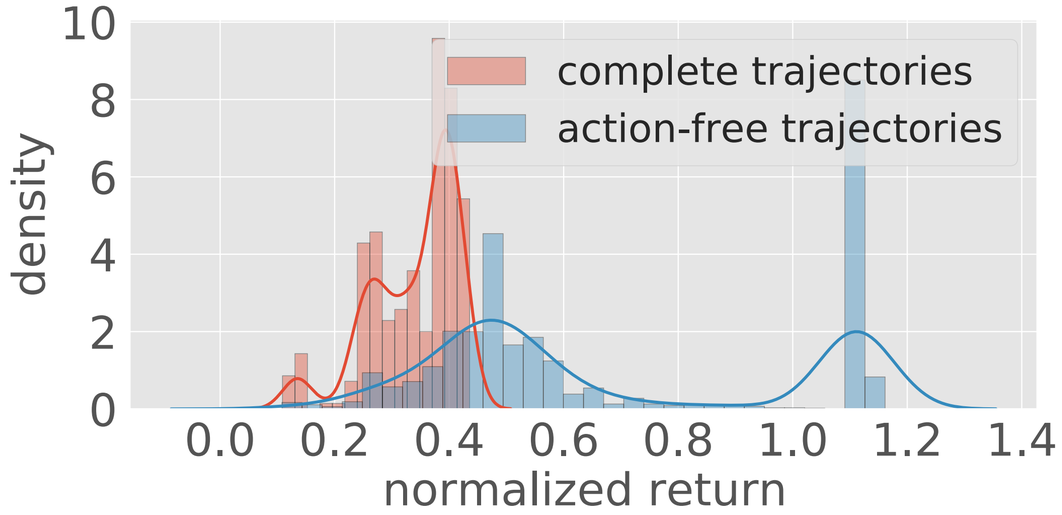


Interesting case:

1. Majority of the data are unlabelled
2. Labelled & unlabelled data have different qualities, especially when labelled data are of low quality, and high quality data are all unlabelled

A Semi-Supervised Offline RL Setting

Offline Dataset = Action-Free Trajectories (unlabelled) ∪ **Action-Complete Trajectories (labelled)**



Interesting case:

1. Majority of the data are unlabelled
2. Labelled & unlabelled data have different qualities, especially when labelled data are of low quality, and high quality data are all unlabelled

Motivating Examples: Learning from Videos and 3rd Person Imitation Learning

There are tremendous amounts of internet videos can be potentially used to train RL agents, yet they are without action labels and are of varying quality.

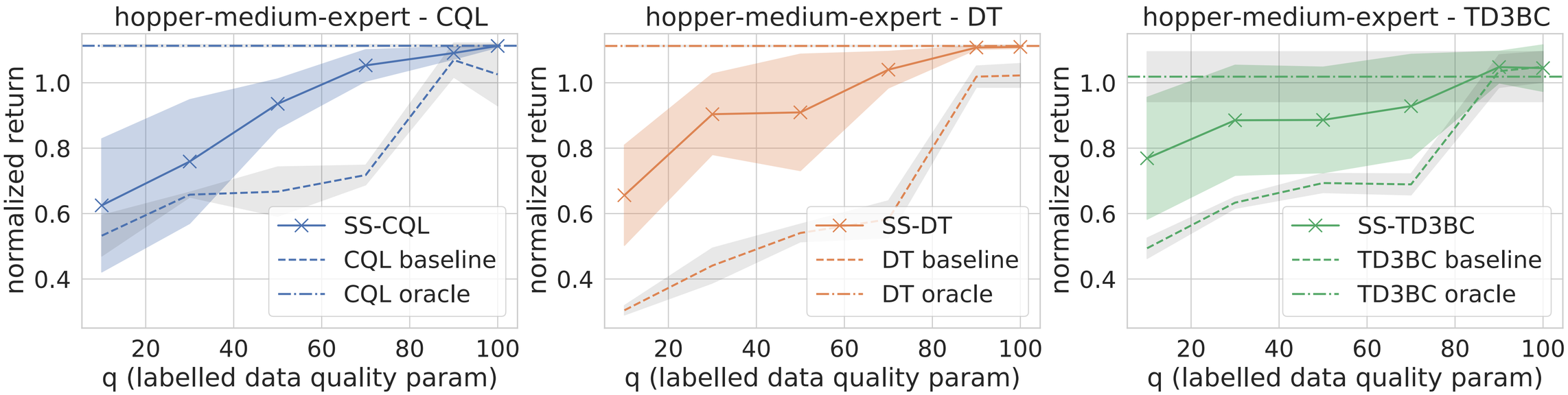
SS-ORL: A Simple & Generic Pipeline

1. Train an inverse dynamics model (IDM) on the labelled dataset
   to predict actions from state transitions

2. Fill in pseudo-actions for those unlabelled (action-free) trajectories using the trained IDM

3. Train an offline RL agent using both the labelled and proxy-labelled data

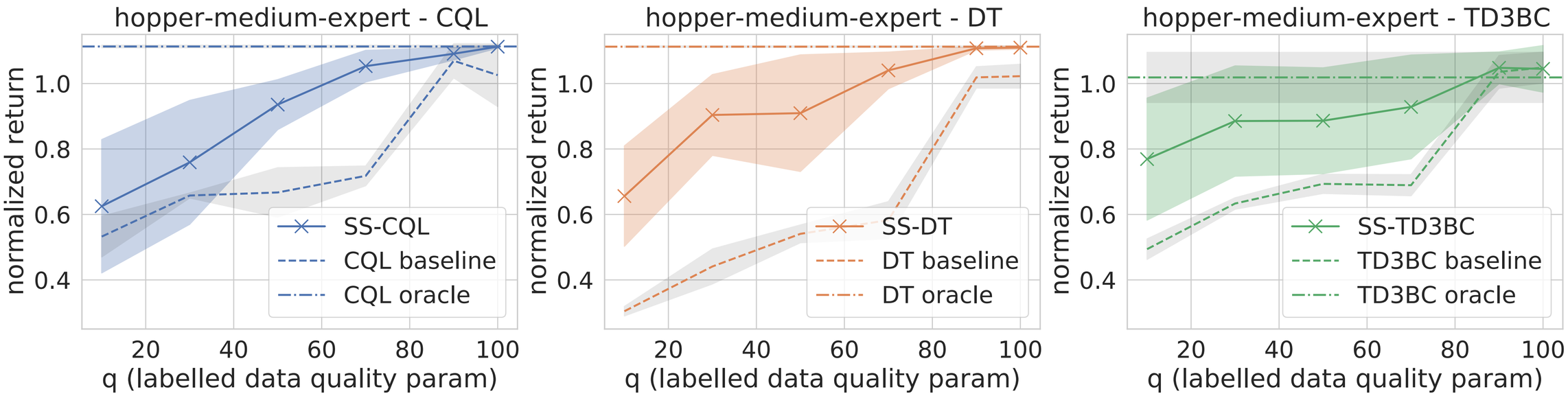
SS-ORL: Highly Successful on D4RL Datasets



Data Setup

- Labelled: 10% of the total number of trajectories, randomly sampled from the bottom q% trajectories
- Unlabelled: the other trajectories with actions removed

SS-ORL: Highly Successful on D4RL Datasets



SS-ORL can effectively utilize the unlabelled trajectories to improve performance, and even match the oracle performance when the quality of the labelled trajectories improves.

Baseline – Trained on the labelled data only (~ performance lower bound)

Oracle – Trained on the original action-complete dataset (~ performance upper bound)

July 25, Tue, Exhibit Hall 1 #105 Thank you!

More Results of Large Scale Controlled Experiments

- Other D4RL datasets
- Different design choices for training the IDM including classic semi-supervised learning techniques
- Influence of the data-centric properties such as size and quality of the labelled and unlabelled datasets
- How different offline RL methods respond differently to various setups of the dataset size and quality