Missing Figures/Experimental Results of Appendix

A. Data Generation

We generate the whole dataset for 60 Atari 2600 games with 5 seeds, then split them into low, medium, and high dataset.

Please refer to Fig. A-1 for the mean episode return of on line data generation process.

B. Additional Experimental Results

B.1. Results of exploitation-tentative algorithms

For BCQ, a τ of higher value denotes a more strict standard for data selection, i.e., less data would be selected.A τ of higher value denotes a more strict standard for data selection, i.e., less (S, A) pairs would be selected. Even though equation (4) shows that the extrapolation error would decrease with a larger τ, the variance would increase along with decrease of selected data, which could be attributed to the worse performance in Fig. A-2, A-3, and A-4.

On the contrary, for BAIL, a τ of lower value denotes a more strict standard for data selection. Note that the balance between extrapolation error and off line performance is also fitted in BAIL. Out of this experiment, we keep $\tau = 0.7$ in BAIL all the time.

Overall, τ affects BCQ more on the aspect of variance. Carefully choosing an appropriate τ for BCQ may lead to a stable policy, which is robust from the off line learning iterations. For BAIL, τ affects more on the aspect of both off line performance and extrapolation error, and it is a trivial work to balance between them to acquire a policy with better on line performance.

It is noted that other than this experiment, we run BAIL with a fix $\tau = 0.7$ due to its better performance. Besides, the reason why we put the results in the appendix is two-fold, (1) space limitation, (2) It is not the focus of our paper.

B.2. All 60 Atari 2600 games on poor dataset

B.3. All 60 Atari 2600 games on medium dataset

B.4. All 60 Atari 2600 games on high dataset