Offline Reinforcement Learning Hands-On

Jakub Kmec • Louis Monier • Alexandre Laterre • Thomas Pierrot • Valentin Courgeau • Olivier Sigaud • Karim Beguir

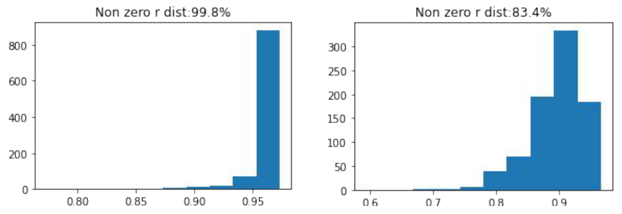
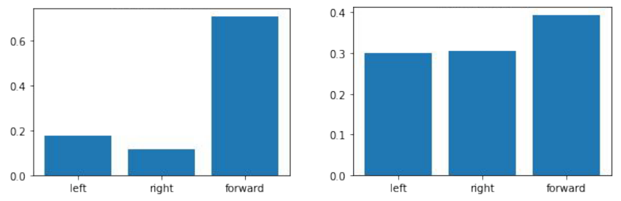
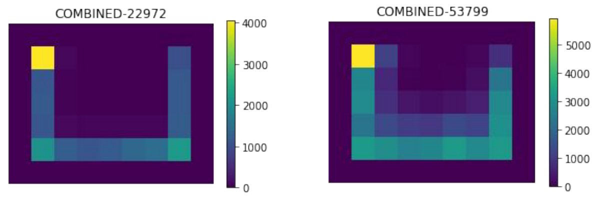
Introduction

Offline Reinforcement Learning (RL) aims to turn large datasets into powerful decision-making engines without any online interactions with the environment. This great promise has motivated a large amount of research that hopes to replicate the success RL has experienced in simulation settings. This work ambitions to reflect upon these efforts from a practitioner viewpoint. We start by discussing the dataset properties that we hypothesise can suggest the potential for applicability of offline RL methods. We then verify these claims through a set of experiments and specifically designed datasets which aim to highlight strengths and weaknesses of current methods.

Key Observation

Online RL uses effectively a **feedback loop**: every action determines the future data and the amount of exploratory behaviour is a parameter.

In offline RL, the amount of exploratory behavior is already fixed. Therefore, the best achievable policy is fully **determined by the dataset**, not the environment.

Property (quality proxy)	EXPERTNESS ⟷ STOCHASTICITY
Reward distribution	
Actions distribution	
State coverage	

How much does Performance vary with Dataset Quality?

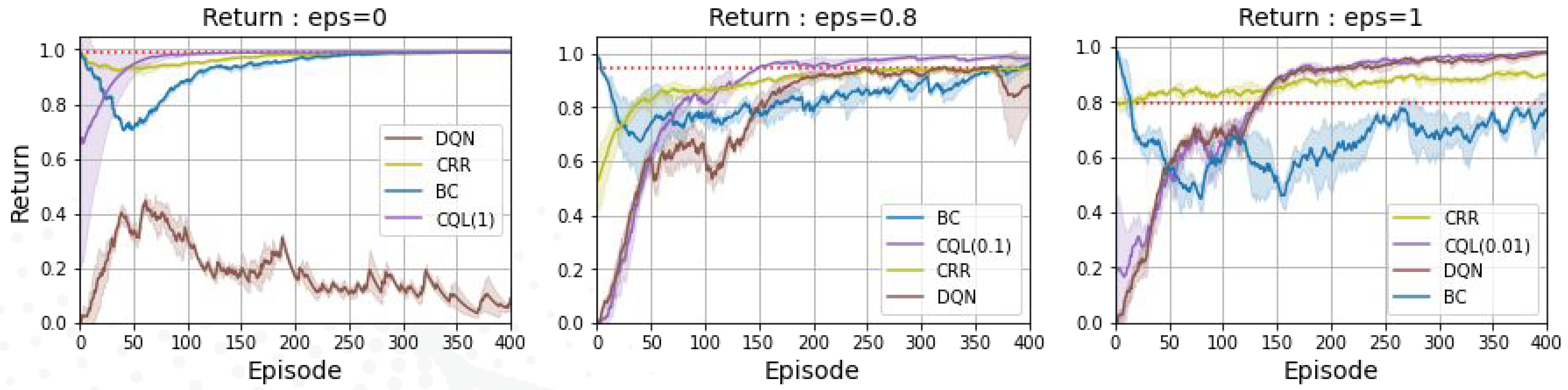
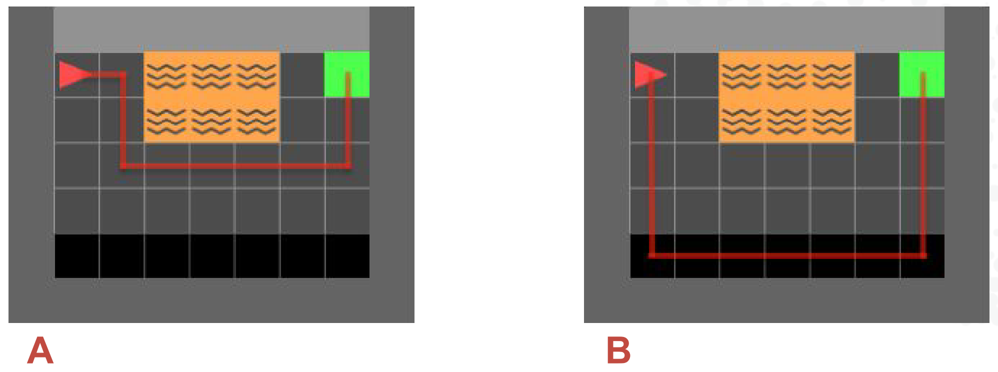


Figure 1 : Comparison of CQL, CRR (with exponential filter), DQN and BC on datasets of different quality. Here, we use an epsilon greedy expert with 3 different values of epsilon to vary the dataset characteristics. The red dotted line shows the average episode return for each dataset (0.991, 0.947, 0.796 respectively).

Experiment 1

- CQL achieves the highest returns, but comes at the cost of high sensitivity to the value of alpha.
- CRR(exp) achieves lower returns but is more robust to the choice of hyper-parameters.
- BC robustly reaches its upper-bound, given by the average return of the trajectories in each dataset.
- DQN fails in the expert settings due to distributional shift, but performs well on a fully random dataset.

Extracting optimal behaviour from multi-modal datasets



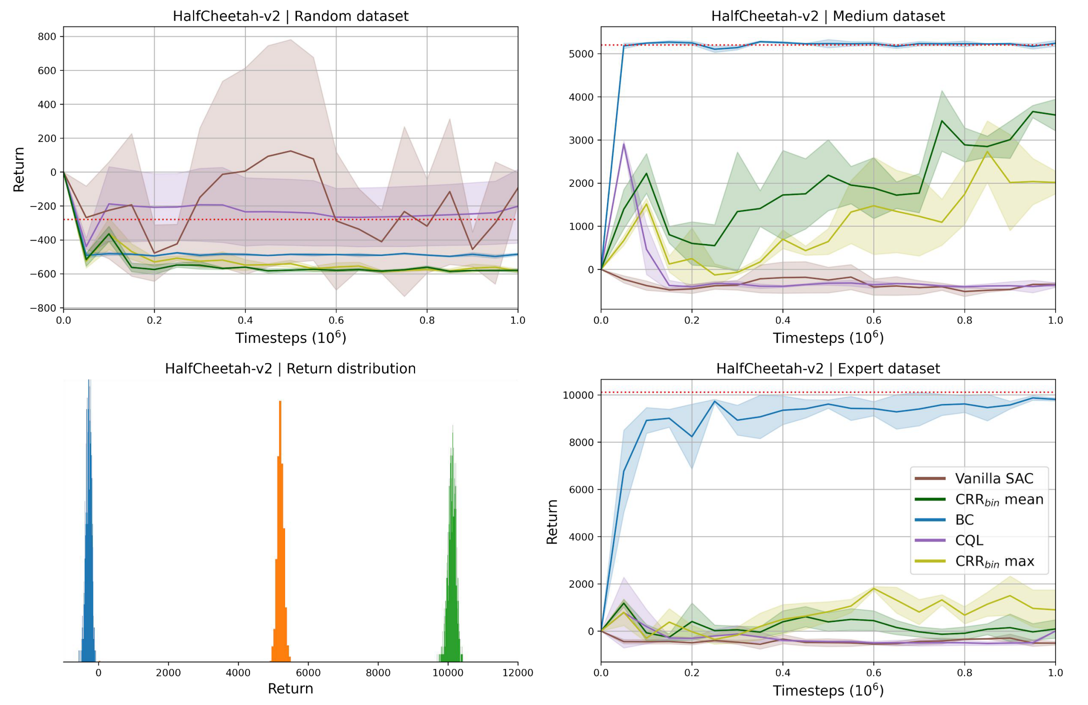
Method	Hyperparameters	Ep. length
CQL	$\alpha = 0.001$	500
	$\alpha = 0.01$	13
	$\alpha = 0.1, \alpha = 1$	17
$\mathrm{CRR}_{\mathrm{exp}}$	$\beta = 1$	17
	$\beta = 0.01$	500
CCRR	$\alpha = 0.01, \beta = 1$	17
	$\alpha = 0.01, \beta = 0.01$	13
BC		17

Figure 2 : This experiment aims to test the agents ability to recover the optimal behaviour (20% of A) from a dataset containing a majority of suboptimal trajectories (80% of B). Agents constrained to lie too close to the behaviour are expected to fail.

Experiment 2

- There is a value of alpha for which CQL converges to the correct policy, but the algorithm is very sensitive.
- BC and CRR with an exponential filter fail to learn the correct policy and converge to the long path.
- A new variant of CRR that incorporates the CQL penalty (CCRR) works but still requires extensive tuning.

Comparison in more Complex Environments



Experiment 3

- CQL remains unstable and is unable to learn robustly across all the dataset-task pairs
- CRR only partly retrieves the performance contained in the medium quality dataset
- BC remains a strong and robust baseline for any continuous settings with qualitative data

Conclusion

In this offline RL hands-on, we returned to practical considerations: data and algorithms. Our contribution provides foundations to help the reader better grasp the issues and suitability of different offline RL methods. That is, for now, no method performs uniformly better independently of the use case. CQL and CRR show good recovery properties but dealing with custom datasets uncovers many practical pitfalls.

References

1- Aviral Kumar, Aurick Zhou, G. Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning.ArXiv, abs/2006.04779, 2020.

2- Ziyu Wang, A. Novikov, Konrad Zolna, Jost Tobias Springenberg, Scott Reed, B. Shahriari,N. Siegel, Josh Merel, Caglar Gulcehre, Nicolas Heess, and N. D. Freitas. Critic regularized regression.ArXiv, abs/2006.15134, 2020.

3- Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems.arXiv preprint arXiv:2005.01643, 2020.