Week 9 Paper Review

Niharika Shrivastava
School of Computing
National University of Singapore
Singapore, 119077
niharika@comp.nus.edu.sg

Abstract

This is a brief review of [1], [2], [3].

1 Conservative Q-Learning for Offline Reinforcement Learning

The authors introduce Conservation Q-Learning (CQL) which aims to avoid overestimation of values induced by the distributional shift between the dataset and the learned policy, such that the expected value of a policy under this Q-function lower bounds its true value using a regularizer. This is done by minimizing values for OOD actions with the states always following the marginal state distribution, then further tightening the bound by maximizing the returns from in-distribution actions. This approach, as a whole, provides a lower-bound estimate of the value function by regularizing Q-values during training. It is also known to provide a point-wise lower bound guarantee, but not overall optimality.

The paper also introduces the concept of gap-expanding which states that a Q-function is gap-expanding if it only over-estimates the gap between in-distribution and OOD actions, thereby avoiding OOD actions. This in turn helps to penalize the Q values in an intelligent way, thereby leading to a tighter lower-bound of the true value.

Empirically, it was shown that CQL underestimated the returns on both discrete and continuous control domains whereas all other methods grossly overestimated their returns by a large order of maginitudes, thereby enabling CQL to attain 2-5 times higher final return.

2 Revisiting Design Choices in Offline Model-Based Reinforcement Learning

The paper motivates a deeper understanding of the type of hyperparameter choices prior works have used based on heuristics and assumptions to get state of the art empirical results in the realm of offline Model-Based RL approaches (MBRL). Specifically, they advocate the use of longer horizon rollouts with larger penalties for uncertainty to improve existing methods. Furthermore, penalizing true dynamics error in the form of uncertainty estimation is better suited than penalizing OOD state-action generation for MB methods. Finally, using an ensemble of penalties is known to be more stable with change in other hyperparameters (such as a greater number of models in the ensemble) while incorporating different forms of system errors.

Multiple experiments are conducted in order to validate these design choices. Finally, a formal methodology based on Bayesian Optimization was used to automatically select the choice of hyper-parameters for every setting. Even in this case, a combination of ensemble penalties along with larger values of uncertainty estimation, horizon rollouts, and the number of models used for ensembling provide the best performance, along with providing stability during training over every iteration. This was noted to be a significant factor for offline RL.

Preprint. Under review.

However, their optimizations were only applied to MOPO and may need more evidence in order to validate that tuning in this way would infact work for every prior other MB-RL approach.

3 MOPO: Model-based Offline Policy Optimization

The authors of this paper introduce a model-based RL algorithm for the offline setting (MOPO) which penalizes rewards by estimating uncertainties in the learned environment dynamics. The motivation is to explore out-of-distribution states for a better-generalized policy since there is no corrective learning from new data in an offline RL setting. They theoretically prove that the algorithm maximizes a lower bound of the policy's return under the true MDP and find the optimal trade-off between the return of greater rewards and the risk of exploring unseen states.

According to the authors, a potential benefit of a soft penalty is that the policy is allowed to take a few risky actions and then return to the confident area near the behavioural distribution without being terminated. To achieve this optimal balance, the return from below by the return of a constructed uncertainty-penalized MDP is bounded, followed by the maximization of conservative estimation of the return in unseen states.

Experiments show that MOPO outperforms significantly for most tasks except medium datasets due to the lack of action diversity in them making it difficult to learn nicely generalized models. Therefore, MOPO is a good formulation since it accounts for uncertainty in the environment, appropriately penalizes it, and provides generalization in the form of exploration of unseen states.

References

[1] Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-Learning for Offline Reinforcement Learning. ArXiv. https://doi.org/10.48550/arXiv.2006.04779

[2] Lu, C., Ball, P. J., Osborne, M. A., & Roberts, S. J. (2021). Revisiting Design Choices in Offline Model-Based Reinforcement Learning. ArXiv. https://doi.org/10.48550/arXiv.2110.04135

[3] Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., & Ma, T. (2020). MOPO: Model-based Offline Policy Optimization. ArXiv. https://doi.org/10.48550/arXiv.2005.13239