Provably Pessimism/ Exploration via Regularization

Motivation. Offline RL algorithms promise to learn effective policies from previously-collected, static datasets without further interaction. However, in practice, offline RL presents a major challenge, and standard off-policy RL methods can fail due to overestimation of values induced by the distributional shift between the dataset and the learned policy, especially when training on complex and multi-modal data distributions. In theory, there are two different kinds of algorithms that achieve the near-optimal gap bound, i.e., LCB (Lower Confidence Bound) type algorithms and TS (Thompson Sampling) type algorithms [1–3]. However, the LCB type algorithms [4] enjoy well-established theoretical guarantees but suffer from difficult implementation (because of the hardness to maintain a counter, especially impossible in the infinite states setting). The TS type algorithms [5], inject noise into the data several times to establish high probability LCB, but significantly increase the computational cost (typically a $\log(S)$ multiplicative factor). Therefore, we propose to study alternate ways to achieve pessimism or optimism (exploration), and focus on the prevalent method (in practice) that based on adding regularization term. We show that not only it's easy to implement in practice, but also nearly optimal in theory.

Algorithm (Contextual Bandit). As a starter, let's consider the Contextual Bandit (CB) setting. The batch data set is $\mathcal{D} = \{(s_i, a_i, r_i)\}_{i=1}^N$. The offline learning objective is to find a policy $\widehat{\pi}$ based on $\mathcal{D}$ that minimizes the expected sub-optimality $\mathbb{E}_{\mathcal{D},\rho}[r(s, \pi^*(s)) - r(s, \widehat{\pi}(s))]$, where $\rho(\cdot)$ is the initial distribution of the state space. We propose to solve the problem in a Q learning manner, our final goal is to get $\widetilde{Q}$ (a pessimism estimation of the ground truth), and directly take the greedy action.

We first learn the empirical value estimation $\widehat{Q}$ by solving the least square TD error problem, which is given by the equation below:

$$\widehat{Q} = \operatorname*{argmin}_{Q} \sum_{i=1}^{N} (Q(s_i, a_i) - r_i)^2$$

After we construct the empirical value $\widehat{Q}$, we can use $\widehat{Q}$ to design a simple regularization term,

$$\mathcal{R}(\widehat{Q}, Q) = \sum_{(s,a)\in\mathcal{A}\times\mathcal{S}} \log \frac{1}{\widehat{Q}(s,a) - Q(s,a)}.$$

We solve the following least square TD error problem with regularization to get the pessimistic $\widetilde{Q}$,

$$\widetilde{Q} = \operatorname*{argmin}_{Q} \sum_{i=1}^{N} (Q(s_i, a_i) - r_i)^2 + \alpha \mathcal{R}(\widehat{Q}, Q).$$

This algorithmic framework contains the SOTA algorithms as special cases. For example, in [2], they take $\mathcal{R}(\widehat{Q}, Q) = \mathbb{E}_{s\sim\rho, a\sim\exp(\widehat{Q}(s,\cdot)),} Q(s,a)^2$. In [3], $\mathcal{R}(\widehat{Q}, Q) = \mathbb{E}_{\rho\sim s}\mathrm{softmax}(Q(s,\cdot))$.

Proof. The proof is simple, directly take derivative over $Q(s,a)$ we get,

$$2n(s,a)(\widetilde{Q}(s,a) - \widehat{r}_{s,a}) + \alpha \frac{1}{\widehat{Q}(s,a) - \widetilde{Q}(s,a)} = 0$$

Combining with the fact that $\widehat{Q} = \widehat{r}_{s,a}$, we get,

$$\widetilde{Q}(s,a) = \widehat{r}_{s,a} - \sqrt{\frac{\alpha}{2n(s,a)}}.$$

This is the LCB type pessimism, [4] showed that by choosing an appropriate α we can guarantee that the sub-optimality is minimax optimal. □

To do.

- Extend to Tabular MDP (done)
- Linear function approximation (?)
- Other regularization terms

References

[1] Tianhe Yu, Aviral Kumar, Yevgen Chebotar, Karol Hausman, Chelsea Finn, and Sergey Levine. How to leverage unlabeled data in offline reinforcement learning. *arXiv preprint arXiv:2202.01741*, 2022.

[2] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

[3] Yevgen Chebotar, Karol Hausman, Yao Lu, Ted Xiao, Dmitry Kalashnikov, Jake Varley, Alex Irpan, Benjamin Eysenbach, Ryan Julian, Chelsea Finn, et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. *arXiv preprint arXiv:2104.07749*, 2021.

[4] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] Zhihan Xiong, Ruoqi Shen, and Simon S Du. Randomized exploration is near-optimal for tabular mdp. *arXiv preprint arXiv:2102.09703*, 2021.