CS 542 Stats RL Project Proposal

Name: Kai-Jie Lin
October 25, 2024

1 Proposal:

This project is going to reproduce the theoretical result from [Jin et al., 2022]. We will focus on the analysis of Pessimistic Value Iteration for both General MDP and Linear MDP. For the further extension, we are going to use the idea of [Jin et al., 2022] to analysis empirical offline RL algorithms like [Kumar et al., 2020].

2 Setting:

We consider the same setting as in [Jin et al., 2022].

2.1 Episodic MDP and Performance Metric

We consider an episodic MDP (S, A, H, P, r) with the state space S, action space A, horizon H, transition kernel $P = \{P_h\}_{h=1}^H$, and reward function $r = \{r_h\}_{h=1}^H$. We assume the reward function is bounded, that is, $r_h \in [0, 1]$ for all $h \in [H]$. For any policy $\pi = \{\pi_h\}_{h=1}^H$, we define the (state-)value function $V_h^\pi : S \to \mathbb{R}$ at each step $h \in [H]$ as

$$V_h^\pi(x) = \mathbb{E}_\pi \left[\sum_{i=h}^H r_i(s_i, a_i) \Big| s_h = x \right] \tag{1}$$

and the action-value function (Q-function) $Q_h^\pi : S \times A \to \mathbb{R}$ at each step $h \in [H]$ as

$$Q_h^\pi(x, a) = \mathbb{E}_\pi \left[\sum_{i=h}^H r_i(s_i, a_i) \Big| s_h = x, a_h = a \right]. \tag{2}$$

The Bellman operator at each step $h \in [H]$ as

$$(\mathcal{T}_h f)(x, a) = \mathbb{E} \left[r_h(s_h, a_h) + f(s_{h+1}) \Big| s_h = x, a_h = a \right]. \tag{3}$$

For the episodic MDP (S, A, H, P, r), we use π^*, Q_h^*, and V_h^* to denote the optimal policy, optimal Q-function, and optimal value function, respectively. We have $V_{H+1}^* = 0$ and the Bellman optimality equation

$$V_h^*(x) = \max_{a \in A} Q_h^*(x, a), \quad Q_h^*(x, a) = (B_h V_{h+1}^*)(x, a). \tag{4}$$

Meanwhile, the optimal policy π^* is specified by

$$\pi_h^*(\cdot|x) = \arg\max_{\pi_h} \langle Q_h^*(x, \cdot), \pi_h(\cdot|x) \rangle_A, \quad V_h^*(x) = \langle Q_h^*(x, \cdot), \pi_h^*(\cdot|x) \rangle_A, \tag{5}$$

where the maximum is taken over all functions mapping from S to distributions over A. We aim to learn a policy that maximizes the expected cumulative reward. Correspondingly, we define the performance metric as

$$\text{SubOpt}(\pi; x) = V_1^{\pi^*}(x) - V_1^\pi(x), \tag{6}$$

which is the suboptimality of the policy π given the initial state $s_1 = x$.

2.2 Data Assumptions:

Definition 1 (Compliance). For a dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$, let $\mathbb{P}_\mathcal{D}$ be the joint distribution of the data collecting process. We say $\mathcal{D}$ is compliant with an underlying MDP (S, A, H, P, r) if

$$\mathbb{P}_\mathcal{D}(r_h^\tau = r', x_{h+1}^\tau = x' | \{(x_h^j, a_h^j)\}_{j=1}^\tau, \{(r_h^j, x_{h+1}^j)\}_{j=1}^{\tau-1}) = \mathbb{P}(r_h(s_h, a_h) = r', s_{h+1} = x' | s_h = x_h^\tau, a_h = a_h^\tau) \quad (7)$$

for all $r' \in [0, 1]$ and $x' \in S$ at each step $h \in [H]$ of each trajectory $\tau \in [K]$. Here $\mathbb{P}$ on the right-hand side of Equation (2.7) is taken with respect to the underlying MDP.

Assumption 2 (Data Collecting Process)**.** The dataset $\mathcal{D}$ that the learner has access to is compliant with the underlying MDP (S, A, H, P, r).

3 Theorem to Proof:

Definition 3 (ξ-Uncertainty Quantifier)**.** We say $\{\Gamma_h\}_{h=1}^H$ ($\Gamma_h : S \times A \to \mathbb{R}$) is a ξ-uncertainty quantifier with respect to $\mathbb{P}_\mathcal{D}$ if the event

$$\mathcal{E} = \left\{ |(\hat{\mathcal{B}}_h \hat{V}_{h+1})(x, a) - (\mathcal{B}_h \hat{V}_{h+1})(x, a)| \le \Gamma_h(x, a) \text{ for all } (x, a) \in S \times A, h \in [H] \right\} \quad (8)$$

satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}) \ge 1 - \xi$.

Theorem 1 (Suboptimality for General MDP)**.** *Suppose $\{\Gamma_h\}_{h=1}^H$ in PEVI is a ξ-uncertainty quantifier. Under $\mathcal{E}$ defined in Equation (4.1), which satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}) \ge 1 - \xi$, for any $x \in S$, Pess($\mathcal{D}$) in Algorithm 1 satisfies*

$$\textit{SubOpt}(\textit{Pess}(\mathcal{D}); x) \le 2 \sum_{h=1}^H \mathbb{E}_{\pi^*}[\Gamma_h(s_h, a_h) | s_1 = x]. \quad (9)$$

Here $\mathbb{E}_{\pi^*}$ is with respect to the trajectory induced by π^* in the underlying MDP given the fixed function Γ_h.

Theorem 2 (Suboptimality for Linear MDP)**.** *Suppose Assumption 2.2 holds and the underlying MDP is a linear MDP. In Algorithm 2, we set*

$$\lambda = 1, \quad \beta = c \cdot dH\sqrt{\zeta}, \text{ where } \zeta = \log(2dHK/\xi). \quad (10)$$

Here $c > 0$ is an absolute constant and $\xi \in (0, 1)$ is the confidence parameter. The following statements hold:

(i) $\{\Gamma_h\}_{h=1}^H$ in Algorithm 2, which is specified in Equation (4.7), is a ξ-uncertainty quantifier, and hence

(ii) under $\mathcal{E}$ defined in Equation (4.1), which satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}) \ge 1 - \xi$, for any $x \in S$, Pess($\mathcal{D}$) in Algorithm 2 satisfies

$$\textit{SubOpt}(\textit{Pess}(\mathcal{D}); x) \le 2\beta \sum_{h=1}^H \mathbb{E}_{\pi^*} \left[(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h))^{1/2} \middle| s_1 = x \right]. \quad (11)$$

Here $\mathbb{E}_{\pi^*}$ is with respect to the trajectory induced by π^* in the underlying MDP given the fixed matrix Λ_h.

Extension to the CQL part is not sure yet.

References

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? 2022. URL https://arxiv.org/abs/2012.15085.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. 2020. URL https://arxiv.org/abs/2006.04779.

[Jin et al., 2022] [Kumar et al., 2020]