Is Pessimism Provably Efficient for Offline RL?

Kai-Jie Lin
UIUC 2024 Fall
CS 542 Final Project
kjlin2@illinois.edu

Abstract

In this project, we present the analysis of pessimistic variant of the value iteration algorithm (PEVI), which is proposed by Jin et al. (2021). We primalrily focus on the analysis provided by Jin et al. (2021) that establish a data-dependent upper bound on the suboptimality of PEVI for general Markov decision processes (MDPs) and linear MDPs. Sec. 1 first introduces the relevant offline Reinforcement Learning settings. Then Sec. 2 introdices the PEVI algorithm. Finally, the Sec. 3 provides the analysis and proofs.

1 Introduction

Offline Reinforcement Learning (Offline RL) aims to learn an optimal policy from a previously collected dataset without further interaction with the environment. Due to such a lack of continuing exploration, which plays a key role in online RL, any algorithm for offline RL possibly suffers from the insufficient coverage of the dataset. We consider the episodic MDP:

1.1 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

$$(\mathcal{T}_h f)(x, a) = \mathbb{E} \left[r_h(s_h, a_h) + \langle P(s, a), \max_{a \in \mathcal{A}} f(\cdot, a) \rangle \Big| s_h = x, a_h = a \right]. \tag{3}$$

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

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

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 $s_1 = x$.

1.2 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

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

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).

2 Algorithm

2.1 Pessimistic Value Iteration: General MDP

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

$$\mathcal{E} = \Big\{ |(\hat{\mathcal{T}}_h \hat{Q}_{h+1})(x,a) - (\mathcal{T}_h \hat{Q}_{h+1})(x,a)| \le \Gamma_h(x,a) \text{ for all } (x,a) \in S \times A, h \in [H] \Big\} \tag{8}$$

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

Here $\Gamma_h(x,a)$ quantifies the uncertainty that arises from approximating $\mathcal{T}_h \hat{Q}_{h+1}$ using $\hat{\mathcal{T}}_h \hat{Q}_{h+1}$.

Algorithm 1 Pessimistic Value Iteration (PEVI): General MDP

Input: Dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$
Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$
for step $h = H, H-1, \ldots, 1$ **do**
  Construct $(\widehat{\mathcal{T}}_h \widehat{Q}_{h+1})(\cdot,\cdot)$ and $\Gamma_h(\cdot,\cdot)$ based on $\mathcal{D}$
  Set $\overline{Q}_h(\cdot,\cdot) \leftarrow (\widehat{\mathcal{T}}_h \widehat{Q}_{h+1})(\cdot,\cdot) - \Gamma_h(\cdot,\cdot)$
  Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\{\overline{Q}_h(\cdot,\cdot), H - h + 1\}^+$
  Set $\widehat{\pi}_h(\cdot|\cdot) \leftarrow \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot|\cdot)\rangle_\mathcal{A}$
  Set $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot|\cdot)\rangle_\mathcal{A}$
end for
Output: $\texttt{Pess}(\mathcal{D}) = \{\widehat{\pi}_h\}_{h=1}^H$

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*

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

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

2.2 Pessimistic Value Iteration: Linear MDP

Definition 4 (Linear MDP)**.** We say an episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a linear MDP with a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist d unknown (signed) measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that

$$\mathbb{P}_h(x'|x, a) = \langle \phi(x, a), \mu_h(x') \rangle, \quad \mathbb{E}[r_h(s_h, a_h)|s_h = x, a_h = a] = \langle \phi(x, a), \theta_h \rangle$$

for all $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ at each step $h \in [H]$. Here we assume $\|\phi(x, a)\| \leq 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$ and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \leq \sqrt{d}$ at each step $h \in [H]$, where with an abuse of notation, we define $\|\mu_h(\mathcal{S})\| = \int_{\mathcal{S}} \|\mu_h(x)\| dx$.

To provide an algorithm for linear MDP, we construct $\widehat{\mathcal{T}}\widehat{Q}_{h+1}$ based on $\mathcal{D}$. We can define the empirical mean square Bellman error (MSBE) as

$$M_h(w) = \sum_{\tau=1}^{K} \left(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau) - \phi(x_h^\tau, a_h^\tau)^\top w \right)^2$$

at each step $h \in [H]$. Correspondingly, we set

$$(\widehat{\mathcal{T}}_h \widehat{Q}_{h+1})(x, a) = \phi(x, a)^\top \widehat{w}_h, \text{where } \widehat{w}_h = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} M_h(w) + \lambda \cdot \|w\|_2^2$$

at each step $h \in [H]$. Here $\lambda > 0$ is the regularization parameter. Note that $\widehat{w}_h$ has the closed form

$$\widehat{w}_h = \Lambda_h^{-1} \left(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot (r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)) \right), \text{where } \Lambda_h = \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I$$

We construct $\Gamma_h(x, a) = \beta \cdot (\phi(x, a)^\top \Lambda_h^{-1} \phi(x, a))^{1/2}$ at each step $h \in [H]$. Here $\beta > 0$ is scaling parameter.

Algorithm 2 Pessimistic Value Iteration (PEVI): Linear MDP

Input: Dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$
Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$
for step $h = H, H-1, \ldots, 1$ **do**
  Set $\Lambda_h \leftarrow \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I$
  Set $\widehat{w}_h \leftarrow \Lambda_h^{-1} (\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot (r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)))$
  Set $\Gamma_h(\cdot, \cdot) \leftarrow \beta \cdot (\phi(\cdot, \cdot)^\top \Lambda_h^{-1} \phi(\cdot, \cdot))^{1/2}$
  Set $\overline{Q}_h(\cdot, \cdot) \leftarrow \phi(\cdot, \cdot)^\top \widehat{w}_h - \Gamma_h(\cdot, \cdot)$
  Set $\widehat{Q}_h(\cdot, \cdot) \leftarrow \min\{\overline{Q}_h(\cdot, \cdot), H - h + 1\}^+$
  Set $\widehat{\pi}_h(\cdot|\cdot) \leftarrow \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot, \cdot), \pi_h(\cdot|\cdot) \rangle_{\mathcal{A}}$
  Set $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot, \cdot), \widehat{\pi}_h(\cdot|\cdot) \rangle_{\mathcal{A}}$
end for
Output: $\texttt{Pess}(\mathcal{D}) = \{\widehat{\pi}_h\}_{h=1}^{H}$

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

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

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 $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$, for any $x \in \mathcal{S}$, $\texttt{Pess}(\mathcal{D})$ in Algorithm 2 satisfies

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

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

3 Analysis

3.1 Suboptimality Decomposition

We want to give the proof of Theorem(1) and Theorem(2). But first we need to find out what cause suboptimality. The Lemma 3 here decompose the suboptimality into three component that we will use it for further analysis. We define the model evaluation error for every step $h \in [H]$:

$$\iota_h(x,a) = (\mathcal{T}_h \widehat{Q}_{h+1})(x,a) - \widehat{Q}_h(x,a). \tag{10}$$

Lemma 3 (Decomposition of Suboptimality)**.** *Let $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ be the policy such that $\widehat{V}_h(x) = \langle \widehat{Q}_h(x,\cdot), \widehat{\pi}_h(\cdot|x)\rangle_{\mathcal{A}}$. For any $\widehat{\pi}$ and $x \in \mathcal{S}$, we have*

$$\begin{aligned}\textit{SubOpt}(\widehat{\pi}; x) = &\underbrace{-\sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}[\iota_h(s_h,a_h)|s_1 = x]}_{\textit{(i): Spurious Correlation}} + \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}[\iota_h(s_h,a_h)|s_1 = x]}_{\textit{(ii): Intrinsic Uncertainty}} \\ &+ \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}[\langle \widehat{Q}_h(s_h,\cdot), \pi_h^*(\cdot|s_h) - \widehat{\pi}_h(\cdot|s_h)\rangle_{\mathcal{A}}|s_1 = x]}_{\textit{(iii): Optimization Error}}.\end{aligned}$$

Here $\mathbb{E}_{\widehat{\pi}}$ and $\mathbb{E}_{\pi^*}$ are taken with respect to the trajectories induced by $\widehat{\pi}$ and π^* in the underlying MDP given the fixed functions $\widehat{V}_{h+1}$ and $\widehat{Q}_h$, which determine ι_h.

Lemma 4 (Extended Value Difference (Cai et al. 2020))**.** *Let $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ be any two policies and let $\{\hat{Q}_h\}_{h=1}^H$ be any estimated Q-functions. For all $h \in [H]$, we define the estimated value function $\hat{V}_h : \mathcal{S} \to \mathbb{R}$ by setting $\hat{V}_h(x) = \langle \hat{Q}_h(x,\cdot), \pi_h(\cdot|x)\rangle_{\mathcal{A}}$ for all $x \in \mathcal{S}$. For all $x \in \mathcal{S}$, we have*

$$\begin{aligned}\hat{V}_1(x) - V_1^{\pi'}(x) = &\sum_{h=1}^H \mathbb{E}_{\pi'}\big[\langle \hat{Q}_h(s_h,\cdot), \pi_h(\cdot|s_h) - \pi'_h(\cdot|s_h)\rangle_{\mathcal{A}}\big|s_1 = x\big] \\ &+ \sum_{h=1}^H \mathbb{E}_{\pi'}\big[\hat{Q}_h(s_h,a_h) - (\mathcal{T}_h \hat{Q}_{h+1})(s_h,a_h)\big|s_1 = x\big]\end{aligned}$$

where $\mathbb{E}_{\pi'}$ is taken with respect to the trajectory generated by π', while B_h is the Bellman operator defined in Equation (2.4).

Proof of Lemma 3.

$$\begin{aligned}&\text{SubOpt}(\widehat{\pi}; x) = V_1^{\pi^*} - V_1^{\widehat{\pi}} = (V_1^{\pi^*}(x) - \widehat{V}_1(x)) + (\widehat{V}_1(x) - V_1^{\pi_b}(x)) \\ &= \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\langle \widehat{Q}_h(s_h,\cdot), \widehat{\pi}_h(\cdot|s_h) - \pi_h^*(\cdot|s_h)\rangle_{\mathcal{A}}\big|s_1 = x\big] \\ &+ \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\widehat{Q}_h(s_h,a_h) - (\mathcal{T}_h \widehat{Q}_{h+1})(s_h,a_h)\big|s_1 = x\big] \\ &+ \sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}\big[\widehat{Q}_h(s_h,a_h) - (\mathcal{T}_h \widehat{Q}_{h+1})(s_h,a_h)\big|s_1 = x\big] \qquad \text{(Lemma 4)} \\ &= -\sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}\big[\iota_h(s_h,a_h)\big|s_1 = x\big] + \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\iota_h(s_h,a_h)\big|s_1 = x\big] \\ &+ \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\langle \widehat{Q}_h(s_h,\cdot), \widehat{\pi}_h(\cdot|s_h) - \pi_h^*(\cdot|s_h)\rangle_{\mathcal{A}}\big|s_1 = x\big]\end{aligned}$$

□

3.2 Suboptimality for General MDP

Proof of Theorem 1. We first show that for all $h \in [H]$, the constructed Q-function $\widehat{Q}_h$ in Algorithm 1 (2.1) is a pessimistic estimator of the optimal Q-function Q_h^*. In the following, we prove that under the event $\mathcal{E}$ defined in Equation (4.1), ι_h lies within $[0, 2\Gamma_h]$ in a pointwise manner for all $h \in [H]$.
(i) $\iota_h \geq 0$:
For all $h \in [H]$ and all $(x, a) \in \mathcal{S} \times \mathcal{A}$, if $\bar{Q}_h(x, a) < 0$, we have

$$\hat{Q}_h(x, a) = \min\{\bar{Q}_h(x, a), H - h + 1\}^+ = 0.$$

By the definition of ι_h, we have

$$\iota_h(x, a) = (\mathcal{T}_h \hat{Q}_{h+1})(x, a) - \hat{Q}_h(x, a) = (\mathcal{T}_h \hat{Q}_{h+1})(x, a) \geq 0,$$

as r_h and $\hat{Q}_{h+1}$ are nonnegative. Otherwise, if $\bar{Q}_h(x, a) \geq 0$, we have

$$\hat{Q}_h(x, a) = \min\{\bar{Q}_h(x, a), H - h + 1\}^+ \leq \bar{Q}_h(x, a).$$

As $\{\Gamma_h\}_{h=1}^H$ are ξ-uncertainty quantifiers, which are defined in Definition 3 (3), we have

$$\begin{aligned}\iota_h(x, a) &\geq (\mathcal{T}_h \hat{Q}_{h+1})(x, a) - \bar{Q}_h(x, a) \\ &= (\mathcal{T}_h \hat{Q}_{h+1})(x, a) - (\hat{\mathcal{T}}_h \hat{V}_{h+1})(x, a) + \Gamma_h(x, a) \geq 0.\end{aligned}$$

(ii) $\iota \leq 2\Gamma_h$
For all $h \in [H]$ and all (x, a),

$$\begin{aligned}&\bar{Q}_h(x, a) = (\hat{B}_h \hat{V}_{h+1})(x, a) - \Gamma_h(x, a) \leq (B_h \hat{V}_{h+1})(x, a) \leq H - h + 1 \\ \implies &\hat{Q}_h(x, a) = \min\{\bar{Q}_h(x, a), H - h + 1\}^+ = \max\{\bar{Q}_h(x, a), 0\} \leq \bar{Q}_h(x, a) \\ \implies &\iota_h(x, a) = (B_h \hat{V}_{h+1})(x, a) - \hat{Q}_h(x, a) \leq (B_h \hat{V}_{h+1})(x, a) - \bar{Q}_h(x, a) \\ &= (B_h \hat{V}_{h+1})(x, a) - (\hat{B}_h \hat{V}_{h+1})(x, a) + \Gamma_h(x, a) \leq 2\Gamma_h(x, a).\end{aligned}$$

We can conclude that on $\mathcal{E}$,

$$0 \leq \iota_h(x, a) \leq 2\Gamma_h(x, a), \quad \forall (x, a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H].$$

Next, based on Lemma 3 (3),

$$\begin{aligned}\text{SubOpt}(\text{Pess}(\mathcal{D}); x) &\leq -\sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}[\iota_h(s_h, a_h)|s_1 = x] + \sum_{h=1}^H \mathbb{E}_{\pi^*}[\iota_h(s_h, a_h)|s_1 = x] \\ &\leq 2\sum_{h=1}^H \mathbb{E}_{\pi^*}[\Gamma_h(s_h, a_h)|s_1 = x]\end{aligned}$$

□

3.3 Suboptimality for Linear MDP

Proof of Theorem 2 (2). We first show that $\{\Gamma_h\}_{h=1}^H$ specified in algorithm 2 (2.2) are ξ-uncertainty quantifiers. To this end, we upper bound the difference between $(\mathcal{T}_h \hat{Q}_{h+1})(x, a)$ and $(\hat{\mathcal{T}}_h \hat{Q}_{h+1})(x, a)$ for all $h \in [H]$ and all $(x, a) \in \mathcal{S} \times \mathcal{A}$.
Based on Assumption 2 and Linear MDP: $\forall (x, a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H]$.

$$\begin{aligned}&(\mathbb{P}_h V)(x, a) = \left\langle \phi(x, a), \int_{\mathcal{S}} V(x')\mu_h(x')dx' \right\rangle \\ \implies &(\mathcal{T}_h Q)(x, a) = \left\langle \phi(x, a), \theta_h \right\rangle + \left\langle \phi(x, a), \int_{\mathcal{S}} V(x')\mu_h(x')dx' \right\rangle \\ \implies &\exists w_h \in \mathbb{R}^d, (\mathcal{T}_h \hat{Q}_{h+1})(x, a) = \phi(x, a)^\top w_h\end{aligned}$$

Recall the definition of $\hat{w}_h$ in Algorithm 2.2 and the construction of $\hat{\mathcal{T}}_h \hat{Q}_{h+1}$ in Algorithm 2.2. The following lemma upper bounds the norms of w_h and $\hat{w}_h$, respectively.

Lemma 5 (Bounded Weights of Value Functions)**.** *Let $V_{\max} > 0$ be an absolute constant. For any function $V : \mathcal{S} \to [0, V_{\max}]$ and any $h \in [H]$, we have*

$$\|w_h\| \le (1 + V_{\max})\sqrt{d}, \quad \|\hat{w}_h\| \le H\sqrt{Kd}/\lambda,$$

Proof.

$$\begin{aligned}
&\forall h \in [H], w_h = \theta_h + \int_{\mathcal{S}} V(x')\mu_h(x')dx', \\
&\|w_h\| \le \|\theta_h\| + \Big\| \int_{\mathcal{S}} V(x')\mu_h(x')dx' \Big\| \le \|\theta_h\| + \int_{\mathcal{S}} \|V(x')\mu_h(x')\|dx' \\
&\quad \le \sqrt{d} + V_{\max} \cdot \|\mu_h(\mathcal{S})\| \le (1 + V_{\max})\sqrt{d} \\
&\|\hat{w}_h\| = \|\Lambda_h^{-1}\Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot (r_h^\tau + \hat{V}_{h+1}(x_{h+1}^\tau))\Big)\| \\
&\quad \le \sum_{\tau=1}^{K} \|\Lambda_h^{-1}\phi(x_h^\tau, a_h^\tau) \cdot (r_h^\tau + \hat{V}_{h+1}(x_{h+1}^\tau)) \quad \text{(triangle inequality)} \\
&\quad \le H \sum_{\tau=1}^{K} \|\Lambda_h^{-1}\phi(x_h^\tau, a_h^\tau)\| \quad (|r_h^\tau + \hat{V}_{h+1}(x_{h+1}^\tau)| \le H) \\
&\quad = H \sum_{\tau=1}^{K} \sqrt{\phi(x_h^\tau, a_h^\tau)^\top \Lambda_h^{-1/2}\Lambda_h^{-1}\Lambda_h^{-1/2}\phi(x_h^\tau, a_h^\tau)} \\
&\quad \le \frac{H}{\sqrt{\lambda}} \sum_{\tau=1}^{K} \sqrt{\phi(x_h^\tau, a_h^\tau)^\top \Lambda_h^{-1}\phi(x_h^\tau, a_h^\tau)} \\
&\quad \le H\sqrt{\frac{K}{\lambda}} \sqrt{\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)^\top \Lambda_h^{-1}\phi(x_h^\tau, a_h^\tau)} \quad \text{(Cauchy-Schwarz inequality)} \\
&\quad = H\sqrt{\frac{K}{\lambda}} \cdot \sqrt{\mathrm{Tr}(\Lambda_h^{-1} \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top)} = H\sqrt{Kd/\lambda}
\end{aligned}$$

□

We upperbound the difference between $\mathcal{T}_h\hat{V}_{Q+1}$ and $\hat{\mathcal{T}}_h\hat{Q}_{h+1}$,

$$\begin{aligned}
&(\mathcal{T}_h\hat{Q}_{h+1})(x,a) - (\hat{\mathcal{T}}_h\hat{Q}_{h+1})(x,a) = \phi(x,a)^\top(w_h - \hat{w}_h) \\
&= \phi(x,a)^\top w_h - \phi(x,a)^\top\Lambda_h^{-1}\Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot (r_h^\tau + \hat{V}_{h+1}(x_{h+1}^\tau))\Big) \\
&= \underbrace{\phi(x,a)^\top w_h - \phi(x,a)^\top\Lambda_h^{-1}\Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot (\mathcal{T}_h\hat{Q}_{h+1})(x_h^\tau, a_h^\tau)\Big)}_{(i)} \\
&\underbrace{- \phi(x,a)^\top\Lambda_h^{-1}\Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \hat{V}_{h+1}(x_{h+1}^\tau) - (\mathcal{T}_h\hat{Q}_{h+1})(x_h^\tau, a_h^\tau)\big)\Big)}_{(ii)}
\end{aligned}$$

We upper bound (i) and (ii) one by one:

$$\begin{aligned}
&|(i)| = |\phi(x,a)^\top w_h - \phi(x,a)^\top\Lambda_h^{-1}(\Lambda_h - \lambda \cdot I)w_h| = \lambda|\phi(x,a)^\top\Lambda_h^{-1}w_h| \\
&\quad \le \lambda\|w_h\|_{\Lambda_h^{-1}}\|\phi(x,a)\|_{\Lambda_h^{-1}} \quad \text{(Cauchy-Schwarz inequality)} \\
&\quad \le H\sqrt{d}\sqrt{\phi(x,a)^\top\Lambda_h^{-1}\phi(x,a)} \quad (\|w_h\|_{\Lambda_h^{-1}} = \sqrt{w_h^\top\Lambda_h^{-1}w_h} \le \|\Lambda_h^{-1}\|_{op}^{1/2} \cdot \|w_h\| \le H\sqrt{d/\lambda})
\end{aligned}$$

For (ii), we define the variable $\epsilon_h^\tau(V) = r_h^\tau + \hat{V}_{h+1}(x_{h+1}^\tau) - (\mathcal{T}_h\hat{Q}_{h+1})(x_h^\tau, a_h^\tau)$,

$$\begin{aligned}
|(ii)| &= \Big|\phi(x,a)^\top \Lambda_h^{-1}\Big(\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(\hat{V}_{h+1})\Big)\Big| \\
&\le \|\phi(x,a)\|_{\Lambda_h^{-1}} \cdot \Big\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(\hat{V}_{h+1})\Big\|_{\Lambda_h^{-1}} \\
&= \sqrt{\phi(x,a)^\top \Lambda_h^{-1}\phi(x,a)} \cdot \underbrace{\Big\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(\hat{V}_{h+1})\Big\|_{\Lambda_h^{-1}}}_{(iii)}
\end{aligned}$$

We want to upper bound (iii). But $\hat{V}_{h+1}$ depends on dataset $\mathcal{D}$, we need uniform concentration inequalities to bound:

$$\sup_{V\in\mathcal{V}_{h+1}(R,B,\lambda)} \|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot\epsilon_h^\tau(V)\|$$

for each $h \in [H]$, where it holds that $\hat{V}_{h+1} \in \mathcal{V}_{h+1}(R, B, \lambda)$. Define the function class,

$$\begin{aligned}
&\mathcal{V}_h(R,B,\lambda) = \Big\{V_h(x;\theta,\beta,\Sigma) : \mathcal{S}\to[0,H] \text{ with } \|\theta\|\le R, \beta\in[0,B], \Sigma \succeq \lambda\cdot I\Big\}, \\
&\text{where } V_h(x;\theta,\beta,\Sigma) = \max_{a\in\mathcal{A}}\Big\{\min\big\{\phi(x,a)^\top\theta - \beta\cdot\sqrt{\phi(x,a)^\top\Sigma^{-1}\phi(x,a)}, H-h+1\big\}^+\Big\}.
\end{aligned}$$

For all $\varepsilon > 0$ and all $h \in [H]$, let $\mathcal{N}(\varepsilon; R, B, \lambda)$ be a minimal ε-cover of $\mathcal{V}_h(R, B, \lambda)$. with respect to supremum norm.
$\Longrightarrow$ For any function $V \in \mathcal{V}_h(R, B, \lambda)$, $\exists\, V^\dagger \in \mathcal{N}_h(R, B, \lambda)$ such that

$$\sup_{x\in\mathcal{S}} |V(x) - V^\dagger(x)| \le \varepsilon$$

By lemma 5, we have $\|\hat{w}_h\| \le H\sqrt{\frac{Kd}{\lambda}}$. Hence, $\forall h \in [H]$, we have

$$\hat{V}_{h+1} \in \mathcal{V}(R_0, B_0, \lambda), \text{where } R_0 = H\sqrt{\frac{Kd}{\lambda}}, B_0 = 2\beta.$$

Therefore, $\exists V_{h+1}^\dagger \in \mathcal{N}_{h+1}(\varepsilon)$ such that $\sup_{x\in\mathcal{S}} |\hat{V}_{h+1}(x) - V_{h+1}^\dagger| \le \varepsilon$

$$\begin{aligned}
&\Longrightarrow |(\mathbb{P}_h V_{h+1}^\dagger)(x,a) - (\mathbb{P}_h\hat{V}_{h+1})(x,a)| \\
&\quad = |\mathbb{E}[V_{h+1}^\dagger(s_{h+1})\big|s_h = x, a_h = a] - \mathbb{E}[\hat{V}_{h+1}(s_{h+1})\big|s_h = x, a_h = a]| \\
&\quad \le \mathbb{E}[|V_{h+1}^\dagger(s_{h+1}) - \hat{V}_{h+1}(s_{h+1})|\big|s_h = x, a_h = a] \le \varepsilon \\
&\Longrightarrow |(\mathcal{T}_h V_{h+1}^\dagger)(x,a) - (\mathcal{T}_h\hat{V}_{h+1})(x,a)| \le \varepsilon \\
&\Longrightarrow |(r_h(x,a) + \hat{V}_{h+1}(x') - (\mathcal{T}_h\hat{V}_{h+1})(x,a)) - (r_h(x,a) + V_{h+1}^\dagger(x') - (\mathcal{T}_h V_{h+1}^\dagger)(x,a))| \le 2\varepsilon \\
&\Longrightarrow |\epsilon_h^\tau(\hat{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger)| \le 2\varepsilon, \quad \forall\tau\in[K], \forall h\in[H] \qquad (6)
\end{aligned}$$

Recall the definition of (iii) and by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|(\text{iii})|^2 &= \Big\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(\hat{V}_{h+1})\Big\|_{\Lambda_h^{-1}}^2 \\
&\le 2\Big\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(V_{h+1}^\dagger)\Big\|_{\Lambda_h^{-1}}^2 + 2\Big\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)(\epsilon_h^\tau(\hat{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger))\Big\|_{\Lambda_h^{-1}}^2
\end{aligned}$$

We can bound the second term by the result in (6) above:

$$
\begin{aligned}
&2\Big\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)(\epsilon_h^\tau(\hat{V}_{h+1})-\epsilon_h^\tau(V_{h+1}^\dagger))\Big\|_{\Lambda_h^{-1}}^2\\
=&2\sum_{\tau,\tau'=1}^{K}\phi(x_h^\tau,a_h^\tau)^\top\Lambda_h^{-1}\phi(x_h^{\tau'},a_h^{\tau'})\cdot(\epsilon_h^\tau(\hat{V}_{h+1})-\epsilon_h^\tau(V_{h+1}^\dagger))\cdot(\epsilon_h^{\tau'}(\hat{V}_{h+1})-\epsilon_h^{\tau'}(V_{h+1}^\dagger))\\
\le&8\varepsilon^2\cdot\sum_{\tau,\tau'=1}^{K}|\phi(x_h^\tau,a_h^\tau)^\top\Lambda_h^{-1}\phi(x_h^{\tau'},a_h^{\tau'})|\le 8\varepsilon^2\cdot\sum_{\tau,\tau'=1}^{K}\|\phi(x_h^\tau,a_h^\tau)\|\cdot\|\phi(x_h^{\tau'},a_h^{\tau'})\|\cdot\|\Lambda_h^{-1}\|\le 8\varepsilon^2K^2/\lambda
\end{aligned}
$$

We left to bound the term $\left\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\epsilon_h^\tau(V_{h+1}^\dagger)\right\|_{\Lambda_h^{-1}}^2$ via uniform concentration inequalities.

Lemma 6 (Concentration of Self-Normalized Processes)**.** *Let $V:S\to[0,H-1]$ be any fixed function. Under Assumption 2, for any fixed $h\in[H]$ and any $\delta\in(0,1)$, we have*

$$
\mathbb{P}_D\left(\left\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(V)\right\|_{\Lambda_h^{-1}}^2>H^2\cdot(2\cdot\log(1/\delta)+d\cdot\log(1+K/\lambda))\right)\le\delta.
$$

Applying Lemma 6 and the union bound, for any fixed $h\in[H]$, we have

$$
\mathbb{P}_D\left(\sup_{V\in\mathcal{N}_{h+1}(\varepsilon)}\left\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(V)\right\|_{\Lambda_h^{-1}}^2>H^2\cdot(2\cdot\log(1/\delta)+d\cdot\log(1+K/\lambda))\right)\le\delta\cdot|\mathcal{N}_{h+1}(\varepsilon)|
$$

For all $\xi\in(0,1)$ and for all $\varepsilon>0$, we set $\delta=\xi/(H\cdot|\mathcal{N}_{h+1}(\varepsilon)|)$. Therefore, for any fixed $h\in[H]$,

$$
\sup_{V\in\mathcal{N}_{h+1}(\varepsilon)}\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(V)\|_{\Lambda_h^{-1}}^2\le H^2\cdot(2\cdot\log(H\cdot|\mathcal{N}_{h+1}(\varepsilon)|/\xi)+d\cdot\log(1+K/\lambda))
$$

with probability greater than $1-\xi$.
It remains to choose a proper $\varepsilon>0$ and upper bound the ε-covering number $|\mathcal{N}_{h+1}(\varepsilon)|$. In the sequal, we set $\varepsilon=dH/K$ and $\lambda=1$. We have

$$
\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\epsilon_h^\tau(\hat{V}_{h+1})\|_{\Lambda_h^{-1}}^2\le 2H^2((2\cdot\log(H\cdot|\mathcal{N}_{h+1}(\varepsilon)|/\xi)+d\cdot\log(1+K/\lambda)+4d^2)\quad\text{w.p.}\ge1-\xi
$$

Lemma 7 (ε-Covering Number (Jin et al., 2020))**.** *For all $h\in[H]$ and all $\varepsilon>0$, we have*

$$
\log|\mathcal{N}_h(\varepsilon;R,B,\lambda)|\le d\cdot\log(1+4R/\varepsilon)+d^2\cdot\log(1+8d^{1/2}B^2/(\varepsilon^2\lambda)).
$$

Recall that

$$
\hat{V}_{h+1}\in\mathcal{V}_{h+1}(R_0,B_0,\lambda),\ \text{where}\ R_0=H\sqrt{Kd/\lambda},B_0=2\beta,\lambda=1,\beta=c\cdot dH\sqrt{\zeta}
$$

Here $c>0$ is an absolute constant, $\xi\in(0,1)$ is the confidence parameter, and $\zeta=log(2dHK/\xi)$ is specified in Algorithm 2 2.2. Applying Lemma 7 with $\varepsilon=dH/K$, we have

$$
\begin{aligned}
\log|\mathcal{N}_{h+1}(\varepsilon)|&\le d\cdot\log(1+4d^{-1/2}K^{3/2})+d^2\cdot\log(1+32c^2\cdot d^{1/2}K^2\zeta)\\
&\le d\cdot\log(1+4d^{1/2}K^2)+d^2\cdot\log(1+32c^2\cdot d^{1/2}K^{2\zeta})\\
&\le 2d^2\cdot\log(1+32c^2\cdot d^{1/2}K^2\zeta)\le 2d^2\cdot\log(64c^2\cdot d^{1/2}K^2\zeta)
\end{aligned}
$$

For all $h\in[H]$, it holds that

$$
\begin{aligned}
\left\|\sum_{k=1}^{K}\phi(x_h^k,a_h^k)\cdot\epsilon_h^k(\hat{V}_{h+1})\right\|_{\Lambda_h^{-1}}^2&\le 2H^2\cdot\Big(2\cdot\log(H/\xi)+4d^2\cdot\log(64c^2\cdot d^{1/2}K^{2\zeta})+d\cdot\log(1+K)+4d^2\Big)\\
&\le d^2H^2\zeta\cdot(36+8\cdot\log(64c^2))
\end{aligned}
$$

We set $c \geq 1$ to be sufficiently large, which ensure that $36 + 8 \cdot \log(64c^2) \leq c^2/4$. Therefore,

$$|(\text{ii})| \leq c/2 \cdot dH\sqrt{\zeta} \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)} = \beta/2 \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)} \quad \text{w.p} \geq 1 - \xi$$

$$\implies |(\mathcal{T}_h \hat{Q}_{h+1})(x,a) - (\hat{\mathcal{T}}_h \hat{Q}_{h+1})(x,a)| \leq (H\sqrt{d} + \beta/2) \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)} \leq \Gamma_h(x,a)$$

with probability at least $1 - \xi$.
We can conclude that $\{\Gamma_h\}_{h=1}^H$ are ξ-uncertainty quantifiers.
By specializing Theorem 1 to the linear MDP, we have

$$\begin{aligned}
\text{SubOpt}(\texttt{Pess}(\mathcal{D}); x) &\leq 2 \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\Gamma_h(s_h, a_h) | s_1 = x] \\
&= 2\beta \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h))^{-1/2} \middle| s_1 = x \right]
\end{aligned}$$

for all $x \in \mathcal{S}$ under $\mathcal{E}$ in definition 3. Therefore, we conclude the proof of Theorem 2. □

References

[1] Ying Jin, Zhuoran Yang and Zhaoran Wang (2021) Is Pessimism Provably Efficient for Offline RL? In International Conference on Machine Learning.

[2] Cai, Q., Yang, Z., Jin, C. and Wang, Z. (2020). Provably efficient exploration in policy optimization. In International Conference on Machine Learning.

[3] Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In Conference on Learning Theory.