A. Illustration of Suboptimality via a Special Case: MAB

We consider the MAB, a special case of the MDP, where $\mathcal{S}$ is a singleton, $\mathcal{A}$ is discrete, and $H = 1$. To simplify the subsequent discussion, we assume without loss of generality

$$r(a) = \mu(a) + \epsilon, \quad \text{where } \epsilon \sim \mathrm{N}(0, 1).$$

Here $\mu(a)$ is the expected reward of each action $a \in \mathcal{A}$ and ϵ is independently drawn. For notational simplicity, we omit the dependency on $h \in [H]$ and $x \in \mathcal{S}$, as $H = 1$ and $\mathcal{S}$ is a singleton. Based on the dataset $\mathcal{D} = \{(a^\tau, r^\tau)\}_{\tau=1}^K$, where $r^\tau = r(a^\tau)$, we consider the sample average estimator

$$\widehat{\mu}(a) = \frac{1}{N(a)} \sum_{\tau=1}^{K} r^\tau \cdot \mathbb{1}\{a^\tau = a\}, \quad \text{where } N(a) = \sum_{\tau=1}^{K} \mathbb{1}\{a^\tau = a\}.$$

Note that $\widehat{\mu}$ serves as the estimated Q-function. Under Assumption 2.2, we have

$$\widehat{\mu}(a) \sim \mathrm{N}\big(\mu(a), 1/N(a)\big). \tag{A.1}$$

In particular, $\{\widehat{\mu}(a)\}_{a \in \mathcal{A}}$ are independent across each action $a \in \mathcal{A}$ conditioning on $\{a^\tau\}_{\tau=1}^K$. We consider the policy

$$\widehat{\pi}(\cdot) = \operatorname*{argmax}_{\pi} \langle \widehat{\mu}(\cdot), \pi(\cdot) \rangle_{\mathcal{A}}, \tag{A.2}$$

which is greedy with respect to $\widehat{\mu}$, as it takes the action $\operatorname{argmax}_{a \in \mathcal{A}} \widehat{\mu}(a)$ with probability one.

By Equation (3.1), Lemma 3.1, and Equation (A.2), we have

$$\text{SubOpt}(\widehat{\pi}; x) \le \underbrace{-\mathbb{E}_{\widehat{\pi}}\big[\iota(a)\big]}_{\text{(i)}} + \underbrace{\mathbb{E}_{\pi^*}\big[\iota(a)\big]}_{\text{(ii)}}, \qquad \text{where } \iota(a) = \mu(a) - \widehat{\mu}(a).$$

Note that $\iota(a)$ is mean zero with respect to $\mathbb{P}_{\mathcal{D}}$ for each action $a \in \mathcal{A}$. Therefore, assuming hypothetically $\widehat{\pi}$ and ι are independent, term (i) is mean zero with respect to $\mathbb{P}_{\mathcal{D}}$. Meanwhile, as π^* and ι are independent, term (ii) is also mean zero with respect to $\mathbb{P}_{\mathcal{D}}$. However, as $\widehat{\pi}$ and ι are spuriously correlated due to their dependency on $\mathcal{D}$, term (i) can be rather large in expectation. See Figure 1 for an illustration. Specifically, we have

$$\begin{aligned} -\mathbb{E}_{\widehat{\pi}}\big[\iota(a)\big] &= \langle \widehat{\mu}(\cdot) - \mu(\cdot), \widehat{\pi}(\cdot) \rangle_{\mathcal{A}} \\ &= \big\langle \widehat{\mu}(\cdot) - \mu(\cdot), \operatorname*{argmax}_{\pi} \langle \widehat{\mu}(\cdot), \pi(\cdot) \rangle_{\mathcal{A}} \big\rangle_{\mathcal{A}}. \end{aligned} \tag{A.3}$$

For example, assuming $\mu(a) = 0$ for each action $a \in \mathcal{A}$, term (i) is the maximum of $|\mathcal{A}|$ Gaussians $\{\mathrm{N}(0, 1/N(a))\}_{a \in \mathcal{A}}$, which can be rather large in expectation, especially when $N(a^\sharp)$ is relatively small for a certain action $a^\sharp \in \mathcal{A}$, e.g., $N(a^\sharp) = 1$. More generally, it is quite possible that $\widehat{\pi}$ takes a certain action $a^\natural \in \mathcal{A}$ with probability one only because $N(a^\natural)$ is relatively small, which allows $\widehat{\mu}(a^\natural)$ to be rather large, even when $\mu(a^\natural)$ is relatively small. Due to such a spurious correlation, $\langle \widehat{\mu}(\cdot) - \mu(\cdot), \widehat{\pi}(\cdot) \rangle_{\mathcal{A}} = \widehat{\mu}(a^\natural) - \mu(a^\natural)$ in Equation (A.3) can be rather large in expectation, which incurs a significant suboptimality. More importantly, such an undesired situation can be quite common in practice, as $\mathcal{D}$ does not necessarily have a "uniform coverage" over each action $a \in \mathcal{A}$. In other words, $N(a^\sharp)$ is often relatively small for at least a certain action $a^\sharp \in \mathcal{A}$.

Going beyond the MAB, that is, $H \ge 1$, such a spurious correlation is further exacerbated, as it is more challenging to ensure each state $x \in \mathcal{S}$ and each action $a \in \mathcal{A}$ are visited sufficiently many times in $\mathcal{D}$. To this end, existing literature (Antos et al., 2007; 2008; Munos and Szepesvári, 2008; Farahmand et al., 2010; 2016; Scherrer et al., 2015; Liu et al., 2018; Nachum et al., 2019a;b; Chen and Jiang, 2019; Tang et al., 2019; Kallus and Uehara, 2019; 2020; Fan et al., 2020; Xie and Jiang, 2020a;b; Jiang and Huang, 2020; Uehara et al., 2020; Duan et al., 2020; Yin et al., 2020; Qu and Wierman, 2020; Li et al., 2020; Liao et al., 2020; Nachum and Dai, 2020; Yang et al., 2020a; Zhang et al., 2020a;b) relies on various assumptions on the "uniform coverage" of $\mathcal{D}$, e.g., finite concentrability coefficients and uniformly lower bounded densities of visitation measures, which however often fail to hold in practice.

B. Well-Explored Dataset

Corollary B.1 (Well-Explored Dataset)**.** Suppose $\mathcal{D}$ consists of K trajectories $\{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$ independently and identically induced by a fixed behavior policy $\bar{\pi}$ in the linear MDP. Meanwhile, suppose there exists an absolute constant

$\underline{c} > 0$ such that

$$\lambda_{\min}(\Sigma_h) \geq \underline{c}, \quad \text{where } \Sigma_h = \mathbb{E}_{\bar{\pi}}\big[\phi(s_h, a_h)\phi(s_h, a_h)^\top\big]$$

at each step $h \in [H]$. Here $\mathbb{E}_{\bar{\pi}}$ is taken with respect to the trajectory induced by $\bar{\pi}$ in the underlying MDP. In Algorithm 2, we set

$$\lambda = 1, \quad \beta = c \cdot dH\sqrt{\zeta}, \quad \text{where } \zeta = \log(4dHK/\xi).$$

Here $c > 0$ is an absolute constant and $\xi \in (0, 1)$ is the confidence parameter. Suppose we have $K \geq C \cdot \log(4dH/\xi)$, where $C > 0$ is a sufficiently large absolute constant that depends on $\underline{c}$. For $\texttt{Pess}(\mathcal{D})$ in Algorithm 2, the event

$$\mathcal{E}^* = \Big\{\text{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) \leq c' \cdot dH^2K^{-1/2}\sqrt{\zeta} \text{ for all } x \in \mathcal{S}\Big\} \tag{B.1}$$

satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}^*) \geq 1 - \xi$. Here $c' > 0$ is an absolute constant that only depends on $\underline{c}$ and c.

Proof of Corollary B.1. See Appendix E.4 for a detailed proof. □

C. Proof Sketch

In this section, we sketch the proofs of the main results in Section 4. In Section C.1, we sketch the proof of Theorem 4.2, which handles any general MDP. In Section C.2, we specialize it to the linear MDP, which is handled by Theorem 4.4. In Section C.3, we sketch the proof of Theorem 4.6, which establishes the information-theoretic lower bound.

C.1. Suboptimality of PEVI: General MDP

Recall that we define the model evaluation errors $\{\iota_h\}_{h=1}^H$ in Equation (3.1), which are based on the (action- and state-)value functions $\{(\widehat{Q}_h, \widehat{V}_h)\}_{h=1}^H$ constructed by PEVI. Also, recall that we define the ξ-uncertainty quantifiers $\{\Gamma_h\}_{h=1}^H$ in Definition 4.1. The key to the proof of Theorem 4.2 is to show that for all $h \in [H]$, the constructed Q-function $\widehat{Q}_h$ in Algorithm 1 is a pessimistic estimator of the optimal Q-function Q_h^*. To this end, in the following lemma, 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 $h \in [H]$. Recall that $\mathbb{P}_\mathcal{D}$ is the joint distribution of the data collecting process.

Lemma C.1 (Pessimism for General MDP)**.** Suppose that $\{\Gamma_h\}_{h=1}^H$ in Algorithm 1 are ξ-uncertainty quantifiers. Under $\mathcal{E}$ defined in Equation (4.1), which satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}) \geq 1 - \xi$, we have

$$0 \leq \iota_h(x, a) \leq 2\Gamma_h(x, a), \quad \text{for all } (x, a) \in \mathcal{S} \times \mathcal{A},\ h \in [H]. \tag{C.1}$$

Proof of Lemma C.1. See Appendix E.1 for a detailed proof. □

In Equation (C.1), the nonnegativity of $\{\iota_h\}_{h=1}^H$ implies the pessimism of $\{\widehat{Q}_h\}_{h=1}^H$, that is, $\widehat{Q}_h \leq Q_h^*$ in a pointwise manner for all $h \in [H]$. To see this, note that the definition of $\{\iota_h\}_{h=1}^H$ in Equation (3.1) gives

$$Q_h^*(x, a) - \widehat{Q}_h(x, a) \geq (\mathbb{B}_h V_{h+1}^*)(x, a) - (\mathbb{B}_h \widehat{V}_{h+1})(x, a) = (\mathbb{P}_h V_{h+1}^*)(x, a) - (\mathbb{P}_h \widehat{V}_{h+1})(x, a), \tag{C.2}$$

which together with Equations (2.3) and (2.5) further implies

$$\begin{aligned} Q_h^*(x, a) - \widehat{Q}_h(x, a) &\geq \mathbb{E}\big[\max_{a' \in \mathcal{A}} Q_{h+1}^*(s_{h+1}, a') - \langle \widehat{Q}_{h+1}(s_{h+1}, \cdot), \widehat{\pi}_{h+1}(\cdot \,|\, s_{h+1})\rangle_\mathcal{A} \,\big|\, s_h = x, a_h = a\big] \\ &\geq \mathbb{E}\big[\langle Q_{h+1}^*(s_{h+1}, \cdot) - \widehat{Q}_{h+1}(s_{h+1}, \cdot), \widehat{\pi}_{h+1}(\cdot \,|\, s_{h+1})\rangle_\mathcal{A} \,\big|\, s_h = x, a_h = a\big] \end{aligned} \tag{C.3}$$

for all $(x, a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$. Also, note that $\widehat{V}_{H+1} = V_{H+1}^* = 0$. Therefore, Equation (C.2) implies $Q_H^* \geq \widehat{Q}_H$ in a pointwise manner. Moreover, by recursively applying Equation (C.3), we have $Q_h^* \geq \widehat{Q}_h$ in a pointwise manner for all $h \in [H]$. In other words, Lemma C.1 implies that the pessimism of $\{\widehat{Q}_h\}_{h=1}^H$ holds with probability at least $1 - \xi$ as long as $\{\Gamma_h\}_{h=1}^H$ in Algorithm 1 are ξ-uncertainty quantifiers, which serves as a sufficient condition that can be verified. Meanwhile, the upper bound of $\{\iota_h\}_{h=1}^H$ in Equation (C.1) controls the underestimation bias of $\{\widehat{Q}_h\}_{h=1}^H$, which arises from pessimism.

Based on Lemma C.1, we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. We upper bound the three terms on the right-hand side of Equation (3.2) respectively. Specifically, we apply Lemma 3.1 by setting $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ as the output of Algorithm 1, that is, $\widehat{\pi} = \texttt{Pess}(\mathcal{D})$. As $\widehat{\pi}_h$ is greedy with respect to $\widehat{Q}_h$ for all $h \in [H]$, term (iii) in Equation (3.2) is nonpositive. Therefore, we have

$$\text{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) \leq \underbrace{-\sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}\big[\iota_h(s_h, a_h) \,\big|\, s_1 = x\big]}_{\text{(i)}} + \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\iota_h(s_h, a_h) \,\big|\, s_1 = x\big]}_{\text{(ii)}} \tag{C.4}$$

for all $x \in \mathcal{S}$, where terms (i) and (ii) characterize the spurious correlation and intrinsic uncertainty, respectively. To upper bound such two terms, we invoke Lemma C.1, which implies

$$\text{(i)} \leq 0, \quad \text{(ii)} \leq 2\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\Gamma_h(s_h, a_h) \,\big|\, s_1 = x\big] \tag{C.5}$$

for all $x \in \mathcal{S}$. Combining Equations (C.4) and (C.5), we obtain Equation (4.2) under $\mathcal{E}$ defined in Equation (4.1). Meanwhile, by Definition 4.1, we have $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$. Therefore, we conclude the proof of Theorem 4.2. □

C.2. Suboptimality of PEVI: Linear MDP

Based on Theorem 4.2, we are ready to prove Theorem 4.4, which is specialized to the linear MDP defined in Definition 4.3.

Proof of Theorem 4.4. It suffices to show that $\{\Gamma_h\}_{h=1}^H$ specified in Equation (4.7) are ξ-uncertainty quantifiers, which are defined in Definition 4.1. In the following lemma, we prove that such a statement holds when the regularization parameter $\lambda > 0$ and scaling parameter $\beta > 0$ in Algorithm 2 are properly chosen.

Lemma C.2 (ξ-Uncertainty Quantifier for Linear MDP)**.** Suppose that Assumption 2.2 holds and the underlying MDP is a linear MDP. In Algorithm 2, we set

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

Here $c > 0$ is an absolute constant and $\xi \in (0, 1)$ is the confidence parameter. It holds that $\{\Gamma_h\}_{h=1}^H$ specified in Equation (4.7) are ξ-uncertainty quantifiers, where $\{\widehat{V}_{h+1}\}_{h=1}^H$ used in Equation (4.1) are obtained by Algorithm 2.

Proof of Lemma C.2. See Appendix E.2 for a detailed proof. □

As Lemma C.2 proves that $\{\Gamma_h\}_{h=1}^H$ specified in Equation (4.7) are ξ-uncertainty quantifiers, $\mathcal{E}$ defined in Equation (4.1) satisfies $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$. Recall that $\mathbb{P}_{\mathcal{D}}$ is the joint distribution of the data collecting process. By specializing Theorem 4.2 to the linear MDP, we have

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

for all $x \in \mathcal{S}$ under $\mathcal{E}$ defined in Equation (4.1). Here the last equality follows from Equation (4.7). Therefore, we conclude the proof of Theorem 4.4. □

C.3. Minimax Optimality of PEVI

In this section, we sketch the proof of Theorem 4.6, which establishes the minimax optimality of Theorem 4.4 for the linear MDP. Specifically, in Section C.3.1, we construct a class $\mathfrak{M}$ of linear MDPs and a worst-case dataset $\mathcal{D}$, while in Section C.3.2, we prove Theorem 4.6 via the information-theoretic lower bound.

C.3.1. Construction of a Hard Instance

In the sequel, we construct a class $\mathfrak{M}$ of linear MDPs and a worst-case dataset $\mathcal{D}$, which is compliant with the underlying MDP as defined in Definition 2.1.

Linear MDP: We define the following class of linear MDPs

$$\mathfrak{M} = \big\{M(p_1, p_2, p_3) : p_1, p_2, p_3 \in [1/4, 3/4] \text{ with } p_3 = \min\{p_1, p_2\}\big\}, \tag{C.6}$$

where $M(p_1, p_2, p_3)$ is an episodic MDP with the horizon $H \geq 2$, state space $\mathcal{S} = \{x_0, x_1, x_2\}$, and action space $\mathcal{A} = \{b_j\}_{j=1}^{A}$ with $|\mathcal{A}| = A \geq 3$. In particular, we fix the initial state as $s_1 = x_0$. For the transition kernel, at the first step $h = 1$, we set

$$\begin{aligned} &\mathcal{P}_1(x_1 \,|\, x_0, b_1) = p_1, \quad \mathcal{P}_1(x_2 \,|\, x_0, b_1) = 1 - p_1, \\ &\mathcal{P}_1(x_1 \,|\, x_0, b_2) = p_2, \quad \mathcal{P}_1(x_2 \,|\, x_0, b_2) = 1 - p_2, \\ &\mathcal{P}_1(x_1 \,|\, x_0, b_j) = p_3, \quad \mathcal{P}_1(x_2 \,|\, x_0, b_j) = 1 - p_3, \quad \text{for all } \ j \in \{3, \ldots, A\}. \end{aligned} \tag{C.7}$$

Meanwhile, at any subsequent step $h \in \{2, \ldots, H\}$, we set

$$\mathcal{P}_h(x_1 \,|\, x_1, a) = \mathcal{P}_h(x_2 \,|\, x_2, a) = 1, \quad \text{for all } \ a \in \mathcal{A}.$$

In other words, $x_1, x_2 \in \mathcal{S}$ are the absorbing states. Here $\mathcal{P}_1(x_1 \,|\, x_0, b_1)$ abbreviates $\mathcal{P}_1(s_2 = x_1 \,|\, s_1 = x_0, a_1 = b_1)$. For the reward function, we set

$$\begin{aligned} &r_1(x_0, a) = 0, \quad \text{for all } \ a \in \mathcal{A}, \\ &r_h(x_1, a) = 1, \quad r_h(x_2, a) = 0, \quad \text{for all } \ a \in \mathcal{A}, \ h \in \{2, \ldots, H\}. \end{aligned} \tag{C.8}$$

See Figure 3 for an illustration. Note that $M(p_1, p_2, p_3)$ is a linear MDP, which is defined in Definition 4.3 with the dimension $d = A + 2$. To see this, we set the corresponding feature map $\phi \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ as

$$\begin{aligned} \phi(x_0, b_j) &= (\boldsymbol{e}_j, 0, 0) \in \mathbb{R}^{A+2}, \quad \text{for all } \ j \in [A], \\ \phi(x_1, a) &= (\mathbf{0}_A, 1, 0) \in \mathbb{R}^{A+2}, \quad \text{for all } \ a \in \mathcal{A}, \\ \phi(x_2, a) &= (\mathbf{0}_A, 0, 1) \in \mathbb{R}^{A+2}, \quad \text{for all } \ a \in \mathcal{A}, \end{aligned} \tag{C.9}$$

where $\{\boldsymbol{e}_j\}_{j=1}^{A}$ and $\mathbf{0}_A$ are the canonical basis and zero vector in $\mathbb{R}^A$, respectively.

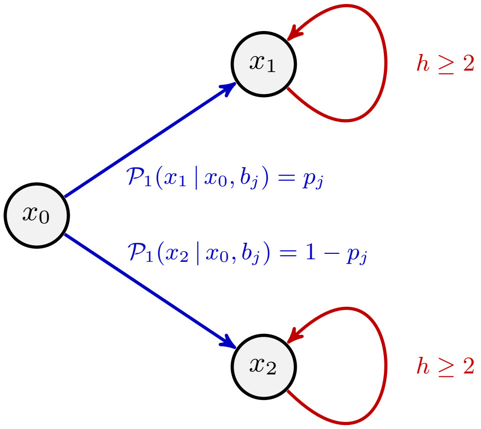
x_1
$h \geq 2$
$\mathcal{P}_1(x_1 \,|\, x_0, b_j) = p_j$
x_0
$\mathcal{P}_1(x_2 \,|\, x_0, b_j) = 1 - p_j$
x_2
$h \geq 2$

Figure 3. An illustration of the episodic MDP $\mathcal{M} = M(p_1, p_2, p_3) \in \mathfrak{M}$ with the state space $\mathcal{S} = \{x_0, x_1, x_2\}$ and action space $\mathcal{A} = \{b_j\}_{j=1}^{A}$. Here we fix the initial state as $s_1 = x_0$, where the agent takes the action $a \in \mathcal{A}$ and transits into the second state $s_2 \in \{x_1, x_2\}$. At the first step $h = 1$, the transition probability satisfies $\mathcal{P}_1(x_1 \,|\, x_0, b_j) = p_j$ and $\mathcal{P}_1(x_2 \,|\, x_0, b_j) = 1 - p_j$ for all $j \in [A]$, where for notational simplicity, we define $p_j = p_3$ for all $j \in \{3, \ldots, A\}$. Also, $x_1, x_2 \in \mathcal{S}$ are the absorbing states. Meanwhile, the reward function satisfies $r_h(x_0, a) = 0$, $r_h(x_1, a) = 1$, and $r_h(x_2, a) = 0$ for all $a \in \mathcal{A}$ and $h \in [H]$.

As $x_1, x_2 \in \mathcal{S}$ are the absorbing states, the optimal policy π_1^* at the first step $h = 1$ is a deterministic policy, which by Equation (C.8) selects the action $a \in \mathcal{A}$ that induces the largest transition probability into the desired state x_1. In other words, at the first step $h = 1$, we have

$$\pi_1^*(b_{j^*} \,|\, x_0) = 1, \quad \text{where } \; j^* = \operatorname*{argmax}_{j \in \{1,2\}} p_j. \tag{C.10}$$

Here we assume without loss of generality $p_1 \neq p_2$ in Equation (C.7). Meanwhile, at any subsequent step $h \in \{2, \ldots, H\}$, an arbitrary policy π_h is optimal, as the action $a \in \mathcal{A}$ selected by π_h does not affect the transition probability. Therefore, for any policy $\pi = \{\pi_h\}_{h=1}^H$, the suboptimality of π for the linear MDP $\mathcal{M} = M(p_1, p_2, p_3)$ takes the form

$$\text{SubOpt}(\mathcal{M}, \pi; x_0) = p_{j^*} \cdot (H - 1) - \sum_{j=1}^A p_j \cdot \pi_1(b_j \,|\, x_0) \cdot (H - 1), \tag{C.11}$$

where for notational simplicity, we define $p_j = p_3$ for all $j \in \{3, \ldots, A\}$. Here with an abuse of notation, we incorporate the explicit dependency on the underlying MDP $\mathcal{M} \in \mathfrak{M}$ into the suboptimality $\text{SubOpt}(\pi; x_0)$.

Dataset: We specify the worst-case data collecting process $\mathbb{P}_{\mathcal{D}}$ as follows. Given a linear MDP $\mathcal{M} \in \mathfrak{M}$, the dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$ consists of K trajectories starting from the same initial state x_0, that is, $x_1^\tau = x_0$ for all $\tau \in [K]$. The initial actions $\{a_1^\tau\}_{\tau \in [K]}$ are predetermined. The subsequent states $\{x_h^\tau\}_{\tau \in [K], h \geq 2}$ are sampled from the underlying MDP $\mathcal{M} = M(p_1, p_2, p_3)$, while the subsequent actions $\{a_h^\tau\}_{\tau \in [K], h \geq 2}$ are arbitrarily chosen, as they do not affect the state transitions. The state transitions in the different trajectories are independent. The immediate reward r_h^τ satisfies $r_h^\tau = r_h(x_h^\tau, a_h^\tau)$. Note that such a dataset $\mathcal{D}$ satisfies Assumption 2.2, that is, $\mathcal{D}$ is compliant with the linear MDP $\mathcal{M} \in \mathfrak{M}$.

We define

$$n_j = \sum_{\tau=1}^K \mathbb{1}\{a_1^\tau = b_j\}, \quad \{\kappa_j^i\}_{i=1}^{n_j} = \bigl\{r_2^\tau \colon a_1^\tau = b_j \text{ with } \tau \in [K]\bigr\}, \quad \text{for all } \; j \in [A]. \tag{C.12}$$

In other words, assuming that $1 \leq \tau_1 < \tau_2 < \cdots < \tau_{n_j} \leq K$ are the episode indices such that $a_1^{\tau_i} = b_j$ for all $i \in [n_j]$, we define $\kappa_j^i = r_2^{\tau_i}$ for all $j \in [A]$. By such a construction, $\{\kappa_j^i\}_{i,j=1}^{n_j,A}$ are the realizations of K independent Bernoulli random variables, which satisfy

$$\mathbb{E}_{\mathcal{D}}[\kappa_j^i] = p_j, \quad \text{for all } \; i \in [n_j], \; j \in [A]. \tag{C.13}$$

Note that knowing the value of the immediate reward r_2^τ is sufficient for determining the value of the second state x_2^τ. Meanwhile, recall that $x_1, x_2 \in \mathcal{S}$ are the absorbing states. Therefore, for learning the optimal policy π^*, the original dataset $\mathcal{D}$ contains the same information as the reduced dataset $\mathcal{D}_1 = \{(x_1^\tau, a_1^\tau, x_2^\tau, r_2^\tau)\}_{\tau=1}^K$, where the randomness only comes from the state transition at the first step $h = 1$ of each trajectory $\tau \in [K]$. Correspondingly, the probability of observing the dataset $\mathcal{D}_1$ takes the form

$$\begin{aligned} \mathbb{P}_{\mathcal{D} \sim \mathcal{M}}(\mathcal{D}_1) &= \prod_{\tau=1}^K \mathbb{P}_{\mathcal{M}}\bigl(r_2(s_2, a_2) = r_2^\tau \,\big|\, s_1 = x_1^\tau, a_1 = a_1^\tau\bigr) \\ &= \prod_{j=1}^A \Bigl(\prod_{i=1}^{n_j} p_j^{\kappa_j^i} \cdot (1 - p_j)^{1 - \kappa_j^i}\Bigr) = \prod_{j=1}^A \Bigl(p_j^{\sum_{i=1}^{n_j} \kappa_j^i} \cdot (1 - p_j)^{n_j - \sum_{i=1}^{n_j} \kappa_j^i}\Bigr). \end{aligned} \tag{C.14}$$

Here $\mathbb{P}_{\mathcal{D} \sim \mathcal{M}}$ denotes the randomness of the dataset $\mathcal{D}$, which is compliant with the underlying MDP $\mathcal{M} = M(p_1, p_2, p_3)$, while $\mathbb{P}_{\mathcal{M}}$ denotes the randomness of the immediate rewards and state transitions. In the second equality, we apply the definition of $\{\kappa_j^i\}_{i=1}^{n_j}$ in Equation (C.12). By such a definition, $\mathbb{P}_{\mathcal{D} \sim \mathcal{M}}(\mathcal{D}_1)$ in Equation (C.14) is the likelihood of the linear MDP $\mathcal{M} \in \mathfrak{M}$ given the reduced dataset $\mathcal{D}_1$ (or equivalently, the original dataset $\mathcal{D}$, assuming that the subsequent actions $\{a_h^\tau\}_{\tau \in [K], h \geq 2}$ are predetermined).

C.3.2. Information-Theoretic Lower Bound

The proof of Theorem 4.6 is based on the Le Cam method (Le Cam, 2012; Yu, 1997). Specifically, we construct two linear MDPs $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$, where the class $\mathfrak{M}$ of linear MDPs is defined in Equation (C.6). Such a construction ensures that (i) the distribution of the dataset $\mathcal{D}$, which is compliant with the underlying MDP, is similar across $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$, and (ii) the suboptimality of any policy π, which is constructed based on the dataset, is different across $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$. In other words, it is hard to distinguish $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$ based on $\mathcal{D}$, while π obtained from $\mathcal{D}$ can not achieve a desired suboptimality for $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$ simultaneously. Such a construction captures the fundamental hardness of offline RL for the linear MDP.

For any $p, p^* \in [1/4, 3/4]$, where $p < p^*$, we set

$$\mathcal{M}_1 = M(p^*, p, p), \quad \mathcal{M}_2 = M(p, p^*, p). \tag{C.15}$$

Based on $\mathcal{D}$, whose likelihood is specified in Equation (C.14), we aim to test whether the underlying MDP is $\mathcal{M}_1$ or $\mathcal{M}_2$. The following lemma establishes a reduction from learning the optimal policy π^* to testing the underlying MDP $\mathcal{M} \in \mathfrak{M}$. Recall that for any $\ell \in \{1, 2\}$, n_ℓ is defined in Equation (C.12).

Lemma C.3. For the dataset $\mathcal{D}$ specified in Section C.3.1, the output $\texttt{Algo}(\mathcal{D})$ of any algorithm satisfies

$$\begin{aligned}
&\max_{\ell \in \{1,2\}} \sqrt{n_\ell} \cdot \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_\ell}\Big[\text{SubOpt}\big(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\big)\Big] \\
&\quad \geq \frac{\sqrt{n_1 n_2}}{\sqrt{n_1} + \sqrt{n_2}} \cdot (p^* - p) \cdot (H - 1) \cdot \Big(\mathbb{E}_{\mathcal{D} \sim \mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_2}\big[\pi_1(b_1 \,|\, x_0)\big]\Big),
\end{aligned}$$

where $\pi = \{\pi_h\}_{h=1}^H = \texttt{Algo}(\mathcal{D})$. For any $\ell \in \{1, 2\}$, $\mathbb{E}_{\mathcal{D} \sim \mathcal{M}_\ell}$ is taken with respect to the randomness of $\mathcal{D}$, which is compliant with the underlying MDP $\mathcal{M}_\ell$.

Proof of Lemma C.3. See Appendix F.1 for a detailed proof. □

As specified in Equation (C.10), for the underlying MDP $\mathcal{M}_1$, the optimal policy π_1^* takes the initial action b_1 with probability one at the initial state x_0, while for $\mathcal{M}_2$, π_1^* takes b_2 with probability one at x_0. We consider the following hypothesis testing problem

$$H_0 \colon \mathcal{M} = \mathcal{M}_1 \quad \text{versus} \quad H_1 \colon \mathcal{M} = \mathcal{M}_2 \tag{C.16}$$

based on the dataset $\mathcal{D}$. For such a problem, any test function ψ is a binary map such that $\psi(\mathcal{D}) = 0$ means the null hypothesis H_0 is accepted, while $\psi(\mathcal{D}) = 1$ means H_0 is rejected. For the output $\pi = \{\pi_h\}_{h=1}^H = \texttt{Algo}(\mathcal{D})$ of any algorithm, we define

$$\psi_{\texttt{Algo}}(\mathcal{D}) = \mathbb{1}\{a \neq b_1\}, \quad \text{where } a \sim \pi_1(\cdot \,|\, x_0). \tag{C.17}$$

Correspondingly, the risk of the (randomized) test function $\psi_{\texttt{Algo}}$ takes the form

$$\begin{aligned}
\text{Risk}(\psi_{\texttt{Algo}}) &= \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_1}\big[\mathbb{1}\{\psi_{\texttt{Algo}}(\mathcal{D}) = 1\}\big] + \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_2}\big[\mathbb{1}\{\psi_{\texttt{Algo}}(\mathcal{D}) = 0\}\big] \\
&= \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_2}\big[\pi_1(b_1 \,|\, x_0)\big].
\end{aligned} \tag{C.18}$$

Therefore, Lemma C.3 lower bounds the suboptimality of any policy $\pi = \{\pi_h\}_{h=1}^H = \texttt{Algo}(\mathcal{D})$ by the risk of a (randomized) test function, which is induced by π, for the corresponding hypothesis testing problem defined in Equation (C.16). Such an approach mirrors the Le Cam method (Le Cam, 2012; Yu, 1997) for establishing the minimax optimality in statistical estimation. In particular, a careful choice of $p, p^* \in [1/4, 3/4]$ leads to the information-theoretic lower bound established in Theorem 4.6. See Appendix F.3 for a detailed proof.

D. Proofs of Suboptimality Decomposition

Proof of Lemma 3.1. By the definition in Equation (2.6), the suboptimality of the policy $\widehat{\pi}$ given any initial state $x \in \mathcal{S}$ can be decomposed as

$$\text{SubOpt}(\widehat{\pi}; x) = \underbrace{\big(V_1^{\pi^*}(x) - \widehat{V}_1(x)\big)}_{\text{(i)}} + \underbrace{\big(\widehat{V}_1(x) - V_1^{\widehat{\pi}}(x)\big)}_{\text{(ii)}}, \tag{D.1}$$

where $\{\widehat{V}_h\}_{h=1}^H$ are the estimated value functions constructed by the meta-algorithm. Term (i) in Equation (D.1) is the difference between the estimated value function $\widehat{V}_1$ and the optimal value function $V_1^{\pi^*}$, while term (ii) is the difference between $\widehat{V}_1$ and the value function $V_1^{\widehat{\pi}}$ of $\widehat{\pi}$. To further decompose terms (i) and (ii), we utilize the following lemma, which is obtained from (Cai et al., 2020), to characterize the difference between an estimated value function and the value function of a policy.

Lemma D.1 (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 $\{\widehat{Q}_h\}_{h=1}^H$ be any estimated Q-functions. For all $h \in [H]$, we define the estimated value function $\widehat{V}_h \colon \mathcal{S} \to \mathbb{R}$ by setting

$\widehat{V}_h(x) = \langle \widehat{Q}_h(x,\cdot), \pi_h(\cdot \,|\, x)\rangle_{\mathcal{A}}$ for all $x \in \mathcal{S}$. For all $x \in \mathcal{S}$, we have

$$\begin{aligned}\widehat{V}_1(x) - V_1^{\pi'}(x) &= \sum_{h=1}^{H} \mathbb{E}_{\pi'}\bigl[\langle \widehat{Q}_h(s_h,\cdot), \pi_h(\cdot \,|\, s_h) - \pi'_h(\cdot \,|\, s_h)\rangle_{\mathcal{A}} \,\big|\, s_1 = x\bigr] \\ &\quad + \sum_{h=1}^{H} \mathbb{E}_{\pi'}\bigl[\widehat{Q}_h(s_h,a_h) - (\mathbb{B}_h \widehat{V}_{h+1})(s_h,a_h) \,\big|\, s_1 = x\bigr],\end{aligned}$$

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

Proof. See Section B.1 in (Cai et al., 2020) for a detailed proof. □

Applying Lemma D.1 with $\pi = \widehat{\pi}$, $\pi' = \pi^*$, and $\{\widehat{Q}_h\}_{h=1}^H$ being the estimated Q-functions constructed by the meta-algorithm, we have

$$\begin{aligned}\widehat{V}_1(x) - V_1^{\pi^*}(x) &= \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\bigl[\langle \widehat{Q}_h(s_h,\cdot), \widehat{\pi}_h(\cdot \,|\, s_h) - \pi^*_h(\cdot \,|\, s_h)\rangle_{\mathcal{A}} \,\big|\, s_1 = x\bigr] \\ &\quad + \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\bigl[\widehat{Q}_h(s_h,a_h) - (\mathbb{B}_h \widehat{V}_{h+1})(s_h,a_h) \,\big|\, s_1 = x\bigr],\end{aligned}$$

where $\mathbb{E}_{\pi^*}$ is taken with respect to the trajectory generated by π^*. By the definition of the model evaluation error ι_h in Equation (3.1), we have

$$V_1^{\pi^*}(x) - \widehat{V}_1(x) = \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\bigl[\langle \widehat{Q}_h(s_h,\cdot), \pi^*_h(\cdot \,|\, s_h) - \widehat{\pi}_h(\cdot \,|\, s_h)\rangle_{\mathcal{A}} \,\big|\, s_1 = x\bigr] + \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\bigl[\iota_h(s_h,a_h) \,\big|\, s_1 = x\bigr].$$

Similarly, applying Lemma D.1 with $\pi = \pi' = \widehat{\pi}$ and $\{\widehat{Q}_h\}_{h=1}^H$ being the estimated Q-functions constructed by the meta-algorithm, we have

$$\widehat{V}_1(x) - V_1^{\widehat{\pi}}(x) = \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\bigl[\widehat{Q}_h(s_h,a_h) - (\mathbb{B}_h \widehat{V}_{h+1})(s_h,a_h) \,\big|\, s_1 = x\bigr] = -\sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\bigl[\iota_h(s_h,a_h) \,\big|\, s_1 = x\bigr],$$

where $\mathbb{E}_{\widehat{\pi}}$ is taken with respect to the trajectory generated by $\widehat{\pi}$. By Equation (D.1), we conclude the proof of Lemma 3.1. □

E. Proofs of Pessimistic Value Iteration

E.1. Proof of Lemma C.1

Proof of Lemma C.1. We first show that on the event $\mathcal{E}$ defined in Equation (4.1), the model evaluation errors $\{\iota_h\}_{h=1}^H$ are nonnegative. In the sequel, we assume that $\mathcal{E}$ holds. Recall the construction of $\overline{Q}_h$ in Line 1 of Algorithm 1 for all $h \in [H]$. For all $h \in [H]$ and all $(x,a) \in \mathcal{S} \times \mathcal{A}$, if $\overline{Q}_h(x,a) < 0$, we have

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

By the definition of ι_h in Equation (3.1), we have

$$\iota_h(x,a) = (\mathbb{B}_h \widehat{V}_{h+1})(x,a) - \widehat{Q}_h(x,a) = (\mathbb{B}_h \widehat{V}_{h+1})(x,a) \geq 0,$$

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

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

As $\{\Gamma_h\}_{h=1}^H$ are ξ-uncertainty quantifiers, which are defined in Definition 4.1, we have

$$\iota_h(x,a) \geq (\mathbb{B}_h \widehat{V}_{h+1})(x,a) - \overline{Q}_h(x,a) = (\mathbb{B}_h \widehat{V}_{h+1})(x,a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x,a) + \Gamma_h(x,a) \geq 0.$$

Here the last inequality follows from the definition of $\mathcal{E}$ in Equation (4.1). Therefore, we conclude the proof of $\iota_h(x,a) \geq 0$ for all $h \in [H]$ and all $(x,a) \in \mathcal{S} \times \mathcal{A}$ on $\mathcal{E}$.

It remains to establish the upper bound in Equation (C.1). For all $h \in [H]$ and all $(x,a) \in \mathcal{S} \times \mathcal{A}$, combining the definition of event $\mathcal{E}$ in Equation (4.1) as well as the construction of $\overline{Q}_h$ in Line 1 of Algorithm 1 gives

$$\overline{Q}_h(x,a) = (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x,a) - \Gamma_h(x,a) \leq (\mathbb{B}_h \widehat{V}_{h+1})(x,a) \leq H - h + 1,$$

where the first inequality follows from the triangle inequality, while the second inequality follows from the fact that $r_h \in [0,1]$ and $\widehat{V}_{h+1} \in [0, H-h]$. Hence, we have

$$\widehat{Q}_h(x,a) = \min\{\overline{Q}_h(x,a), H-h+1\}^+ = \max\{\overline{Q}_h(x,a), 0\} \geq \overline{Q}_h(x,a),$$

which by the definition of ι_h in Equation (3.1) implies

$$\begin{aligned}\iota_h(x,a) &= (\mathbb{B}_h \widehat{V}_{h+1})(x,a) - \widehat{Q}_h(x,a) \leq (\mathbb{B}_h \widehat{V}_{h+1})(x,a) - \overline{Q}_h(x,a) \\ &= (\mathbb{B}_h \widehat{V}_{h+1})(x,a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x,a) + \Gamma_h(x,a) \leq 2\Gamma_h(x,a).\end{aligned}$$

Here the last inequality follows from the definition of $\mathcal{E}$ in Equation (4.1). Therefore, we complete the proof of $\iota_h(x,a) \leq 2\Gamma_h(x,a)$ for all $h \in [H]$ and all $(x,a) \in \mathcal{S} \times \mathcal{A}$ on $\mathcal{E}$.

In summary, we conclude that on $\mathcal{E}$,

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

Therefore, we conclude the proof of Lemma C.1. □

E.2. Proof of Lemma C.2

Proof of Lemma C.2. It suffices to show that under Assumption 2.2, the event $\mathcal{E}$ defined in Equation (4.1) satisfies $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$ with the ξ-uncertainty quantifiers $\{\Gamma_h\}_{h=1}^H$ defined in Equation (4.7). To this end, we upper bound the difference between $(\mathbb{B}_h \widehat{V}_{h+1})(x,a)$ and $(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x,a)$ for all $h \in [H]$ and all $(x,a) \in \mathcal{S} \times \mathcal{A}$, where the Bellman operator $\mathbb{B}_h$ is defined in Equation (2.4), the estimated Bellman operator $\widehat{\mathbb{B}}_h$ is defined in Equation (4.5), and the estimated value function $\widehat{V}_{h+1}$ is constructed in Line 2 of Algorithm 2.

For any function $V \colon \mathcal{S} \to [0, H]$, Definition 4.3 ensures that $\mathbb{P}_h V$ and $\mathbb{B}_h V$ are linear in the feature map ϕ for all $h \in [H]$. To see this, note that Equation (4.4) implies

$$(\mathbb{P}_h V)(x,a) = \Big\langle \phi(x,a), \int_{\mathcal{S}} V(x')\mu_h(x')\mathrm{d}x' \Big\rangle, \qquad \forall (x,a) \in \mathcal{S} \times \mathcal{A},\ \forall h \in [H].$$

Also, Equation (4.4) ensures that the expected reward is linear in ϕ for all $h \in [H]$, which implies

$$(\mathbb{B}_h V)(x,a) = \langle \phi(x,a), \theta_h \rangle + \Big\langle \phi(x,a), \int_{\mathcal{S}} V(x')\mu_h(x')\mathrm{d}x' \Big\rangle, \qquad \forall (x,a) \in \mathcal{S} \times \mathcal{A},\ \forall h \in [H]. \tag{E.1}$$

Hence, there exists an unknown vector $w_h \in \mathbb{R}^d$ such that

$$(\mathbb{B}_h \widehat{V}_{h+1})(x,a) = \phi(x,a)^\top w_h, \qquad \forall (x,a) \in \mathcal{S} \times \mathcal{A},\ \forall h \in [H]. \tag{E.2}$$

Recall the definition of $\widehat{w}_h$ in Equation (4.6) and the construction of $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$ in Equation (4.5). The following lemma upper bounds the norms of w_h and $\widehat{w}_h$, respectively.

Lemma E.1 (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

$$\|w_h\| \leq (1 + V_{\max})\sqrt{d}, \qquad \|\widehat{w}_h\| \leq H\sqrt{Kd/\lambda},$$

where w_h and $\widehat{w}_h$ are defined in Equations (E.2) and (4.6), respectively.

Proof of Lemma E.1. For all $h \in [H]$, Equations (E.1) and (E.2) imply

$$w_h = \theta_h + \int_{\mathcal{S}} V(x')\mu_h(x')\mathrm{d}x'.$$

By the triangle inequality and the fact that $\|\mu_h(\mathcal{S})\| \leq \sqrt{d}$ in Definition 4.3 with the notation $\|\mu_h(\mathcal{S})\| = \int_{\mathcal{S}} \|\mu_h(x')\|\mathrm{d}x'$, we have

$$\begin{aligned}\|w_h\| &\le \|\theta_h\| + \Big\| \int_{\mathcal{S}} V(x')\mu_h(x')\mathrm{d}x' \Big\| \le \|\theta_h\| + \int_{\mathcal{S}} \|V(x')\mu_h(x')\|\mathrm{d}x' \\ &\le \sqrt{d} + V_{\max} \cdot \|\mu_h(\mathcal{S})\| \le (1+V_{\max})\sqrt{d}, \end{aligned} \tag{E.3}$$

where the third inequality follows from the fact that $V \in [0, V_{\max}]$.

Meanwhile, by the definition of $\widehat{w}_h$ in Equation (4.6) and the triangle inequality, we have

$$\begin{aligned}\|\widehat{w}_h\| &= \Big\| \Lambda_h^{-1}\Big(\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)\big)\Big)\Big\| \\ &\le \sum_{\tau=1}^K \big\| \Lambda_h^{-1}\phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)\big)\big\|.\end{aligned}$$

Note that $|r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)| \le H$, which follows from the fact that $r_h^\tau \in [0,1]$ and $\widehat{V}_{h+1} \in [0, H-1]$ by Line 2 of Algorithm 2. Also, note that $\Lambda_h \succeq \lambda \cdot I$, which follows from the definition of Λ_h in Equation (4.6). Hence, we have

$$\begin{aligned}\|\widehat{w}_h\| \le H \cdot \sum_{\tau=1}^K \|\Lambda_h^{-1}\phi(x_h^\tau, a_h^\tau)\| &= H \cdot \sum_{\tau=1}^K \sqrt{\phi(x_h^\tau, a_h^\tau)^\top \Lambda_h^{-1/2}\Lambda_h^{-1}\Lambda_h^{-1/2}\phi(x_h^\tau, a_h^\tau)} \\ &\le \frac{H}{\sqrt{\lambda}} \cdot \sum_{\tau=1}^K \sqrt{\phi(x_h^\tau, a_h^\tau)^\top \Lambda_h^{-1}\phi(x_h^\tau, a_h^\tau)},\end{aligned}$$

where the last inequality follows from the fact that $\|\Lambda_h^{-1}\|_{\mathrm{op}} \le \lambda^{-1}$. Here $\|\cdot\|_{\mathrm{op}}$ denotes the matrix operator norm. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|\widehat{w}_h\| &\le H\sqrt{K/\lambda} \cdot \sqrt{\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)^\top \Lambda_h^{-1}\phi(x_h^\tau, a_h^\tau)} = H\sqrt{K/\lambda} \cdot \sqrt{\mathrm{Tr}\Big(\Lambda_h^{-1}\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top\Big)} \\ &= H\sqrt{K/\lambda} \cdot \sqrt{\mathrm{Tr}\big(\Lambda_h^{-1}(\Lambda_h - \lambda \cdot I)\big)} \le H\sqrt{K/\lambda} \cdot \sqrt{\mathrm{Tr}(\Lambda_h^{-1}\Lambda_h)} = H\sqrt{Kd/\lambda},\end{aligned} \tag{E.4}$$

where the second equality follows from the definition of Λ_h in Equation (4.6).

Therefore, combining Equations (E.3) and (E.4), we conclude the proof of Lemma E.1. □

We upper bound the difference between $\mathbb{B}_h\widehat{V}_{h+1}$ and $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}$. For all $h \in [H]$ and all $(x,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}&(\mathbb{B}_h\widehat{V}_{h+1})(x,a) - (\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x,a) = \phi(x,a)^\top(w_h - \widehat{w}_h) \\ &\quad= \phi(x,a)^\top w_h - \phi(x,a)^\top\Lambda_h^{-1}\Big(\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)\big)\Big) \\ &\quad= \underbrace{\phi(x,a)^\top w_h - \phi(x,a)^\top\Lambda_h^{-1}\Big(\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau) \cdot (\mathbb{B}_h\widehat{V}_{h+1})(x_h^\tau, a_h^\tau)\Big)}_{\text{(i)}} \\ &\qquad \underbrace{- \phi(x,a)^\top\Lambda_h^{-1}\Big(\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau) - (\mathbb{B}_h\widehat{V}_{h+1})(x_h^\tau, a_h^\tau)\big)\Big)}_{\text{(ii)}}.\end{aligned} \tag{E.5}$$

Here the first equality follows from the definition of the Bellman operator $\mathbb{B}_h$ in Equation (2.4), the decomposition of $\mathbb{B}_h$ in Equation (E.1), and the definition of the estimated Bellman operator $\widehat{\mathbb{B}}_h$ in Equation (4.5), while the second equality follows from the definition of $\widehat{w}_h$ in Equation (4.6). By the triangle inequality, we have

$$\big|(\mathbb{B}_h\widehat{V}_{h+1})(x,a) - (\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x,a)\big| \le |\text{(i)}| + |\text{(ii)}|.$$

In the sequel, we upper bound terms (i) and (ii) respectively. By the construction of the estimated value function $\widehat{V}_{h+1}$ in Line 2 of Algorithm 2, we have $\widehat{V}_{h+1} \in [0, H-1]$. By Lemma E.1, we have $\|w_h\| \le H\sqrt{d}$. Hence, term (i) defined in Equation (E.5) is upper bounded by

$$\begin{aligned} |\text{(i)}| &= \Big|\phi(x,a)^\top w_h - \phi(x,a)^\top \Lambda_h^{-1}\Big(\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top w_h\Big)\Big| \\ &= \big|\phi(x,a)^\top w_h - \phi(x,a)^\top \Lambda_h^{-1}(\Lambda_h - \lambda\cdot I)w_h\big| = \lambda \cdot \big|\phi(x,a)^\top \Lambda_h^{-1} w_h\big| \\ &\le \lambda \cdot \|w_h\|_{\Lambda_h^{-1}} \cdot \|\phi(x,a)\|_{\Lambda_h^{-1}} \le H\sqrt{d\lambda}\cdot\sqrt{\phi(x,a)^\top \Lambda_h^{-1}\phi(x,a)}. \end{aligned} \tag{E.6}$$

Here the second equality follows from the definition of Λ_h in Equation (4.6). Also, the first inequality follows from the Cauchy-Schwarz inequality, while the last inequality follows from the fact that

$$\|w_h\|_{\Lambda_h^{-1}} = \sqrt{w_h^\top \Lambda_h^{-1} w_h} \le \|\Lambda_h^{-1}\|_{\text{op}}^{1/2} \cdot \|w_h\| \le H\sqrt{d/\lambda}.$$

Here $\|\cdot\|_{\text{op}}$ denotes the matrix operator norm and we use the fact that $\|\Lambda_h^{-1}\|_{\text{op}} \le \lambda^{-1}$.

It remains to upper bound term (ii). For notational simplicity, for any $h \in [H]$, any $\tau \in [K]$, and any function $V\colon \mathcal{S} \to [0, H]$, we define the random variable

$$\epsilon_h^\tau(V) = r_h^\tau + V(x_{h+1}^\tau) - (\mathbb{B}_h V)(x_h^\tau, a_h^\tau). \tag{E.7}$$

By the Cauchy-Schwarz inequality, term (ii) defined in Equation (E.5) is upper bounded by

$$\begin{aligned} |\text{(ii)}| &= \Big|\phi(x,a)^\top \Lambda_h^{-1}\Big(\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot \epsilon_h^\tau(\widehat{V}_{h+1})\Big)\Big| \\ &\le \Big\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot \epsilon_h^\tau(\widehat{V}_{h+1})\Big\|_{\Lambda_h^{-1}} \cdot \|\phi(x,a)\|_{\Lambda_h^{-1}} \\ &= \underbrace{\Big\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot \epsilon_h^\tau(\widehat{V}_{h+1})\Big\|_{\Lambda_h^{-1}}}_{\text{(iii)}} \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1}\phi(x,a)}. \end{aligned} \tag{E.8}$$

In the sequel, we upper bound term (iii) via concentration inequalities. An obstacle is that $\widehat{V}_{h+1}$ depends on $\{(x_h^\tau, a_h^\tau)\}_{\tau\in[K]}$ via $\{(x_{h'}^\tau, a_{h'}^\tau)\}_{\tau\in[K], h'>h}$, as it is constructed based on the dataset $\mathcal{D}$. To this end, we resort to uniform concentration inequalities to upper bound

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

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

$$\begin{aligned} &\mathcal{V}_h(R,B,\lambda) = \big\{V_h(x;\theta,\beta,\Sigma)\colon \mathcal{S} \to [0,H] \text{ with } \|\theta\| \le R, \beta \in [0,B], \Sigma \succeq \lambda\cdot I\big\}, \\ &\text{where } \ V_h(x;\theta,\beta,\Sigma) = \max_{a\in\mathcal{A}}\Big\{\min\big\{\phi(x,a)^\top\theta - \beta\cdot\sqrt{\phi(x,a)^\top\Sigma^{-1}\phi(x,a)}, H-h+1\big\}^+\Big\}. \end{aligned} \tag{E.9}$$

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

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

Meanwhile, among all ε-covers of $\mathcal{V}_h(R, B, \lambda)$ defined by such a property, we choose $\mathcal{N}_h(\varepsilon; R, B, \lambda)$ as the one with the minimal cardinality.

By Lemma E.1, we have $\|\widehat{w}_h\| \le H\sqrt{Kd/\lambda}$. Hence, for all $h \in [H]$, we have

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

Here $\lambda > 0$ is the regularization parameter and $\beta > 0$ is the scaling parameter, which are specified in Algorithm 2. For notational simplicity, we use $\mathcal{V}_{h+1}$ and $\mathcal{N}_{h+1}(\varepsilon)$ to denote $\mathcal{V}_{h+1}(R_0, B_0, \lambda)$ and $\mathcal{N}_{h+1}(\varepsilon; R_0, B_0, \lambda)$, respectively. As it holds that $\widehat{V}_{h+1} \in \mathcal{V}_{h+1}$ and $\mathcal{N}_{h+1}(\varepsilon)$ is an ε-cover of $\mathcal{V}_{h+1}$, there exists a function $V_{h+1}^\dagger \in \mathcal{N}_{h+1}(\varepsilon)$ such that

$$\sup_{x\in\mathcal{S}} \bigl|\widehat{V}_{h+1}(x) - V_{h+1}^\dagger(x)\bigr| \le \varepsilon. \tag{E.10}$$

Hence, given $V_{h+1}^\dagger$ and $\widehat{V}_{h+1}$, the monotonicity of conditional expectations implies

$$\begin{aligned}
&\bigl|(\mathbb{P}_h V_{h+1}^\dagger)(x,a) - (\mathbb{P}_h \widehat{V}_{h+1})(x,a)\bigr| \\
&\quad = \Bigl|\mathbb{E}\bigl[V_{h+1}^\dagger(s_{h+1}) \,\big|\, s_h = x, a_h = a\bigr] - \mathbb{E}\bigl[\widehat{V}_{h+1}(s_{h+1}) \,\big|\, s_h = x, a_h = a\bigr]\Bigr| \\
&\quad \le \mathbb{E}\Bigl[\bigl|V_{h+1}^\dagger(s_{h+1}) - \widehat{V}_{h+1}(s_{h+1})\bigr| \,\Big|\, s_h = x, a_h = a\Bigr] \le \varepsilon, \qquad \forall (x,a) \in \mathcal{S}\times\mathcal{A},\ \forall h \in [H].
\end{aligned} \tag{E.11}$$

Here the conditional expectation is induced by the transition kernel $\mathcal{P}_h(\cdot\,|\,x,a)$. Combining Equation (E.11) and the definition of the Bellman operator $\mathbb{B}_h$ in Equation (2.4), we have

$$\bigl|(\mathbb{B}_h V_{h+1}^\dagger)(x,a) - (\mathbb{B}_h \widehat{V}_{h+1})(x,a)\bigr| \le \varepsilon, \qquad \forall (x,a) \in \mathcal{S}\times\mathcal{A},\ \forall h \in [H]. \tag{E.12}$$

By the triangle inequality, Equations (E.10) and (E.12) imply

$$\Bigl|\bigl(r_h(x,a) + \widehat{V}_{h+1}(x') - (\mathbb{B}_h \widehat{V}_{h+1})(x,a)\bigr) - \bigl(r_h(x,a) + V_{h+1}^\dagger(x') - (\mathbb{B}_h V_{h+1}^\dagger)(x,a)\bigr)\Bigr| \le 2\varepsilon \tag{E.13}$$

for all $h \in [H]$ and all $(x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$. Setting $(x,a,x') = (x_h^\tau, a_h^\tau, x_{h+1}^\tau)$ in Equation (E.13), we have

$$\bigl|\epsilon_h^\tau(\widehat{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger)\bigr| \le 2\varepsilon, \qquad \forall \tau \in [K],\ \forall h \in [H]. \tag{E.14}$$

Also, recall the definition of term (iii) in Equation (E.8). By the Cauchy-Schwarz inequality, for any two vectors $a, b \in \mathbb{R}^d$ and any positive definite matrix $\Lambda \in \mathbb{R}_+^{d\times d}$, it holds that $\|a+b\|_\Lambda^2 \le 2\cdot\|a\|_\Lambda^2 + 2\cdot\|b\|_\Lambda^2$. Hence, for all $h \in [H]$, we have

$$\text{(iii)} \le 2\cdot\Bigl\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot\epsilon_h^\tau(V_{h+1}^\dagger)\Bigr\|_{\Lambda_h^{-1}}^2 + 2\cdot\Bigl\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot\bigl(\epsilon_h^\tau(\widehat{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger)\bigr)\Bigr\|_{\Lambda_h^{-1}}^2. \tag{E.15}$$

The second term on the right-hand side of Equation (E.15) is upper bounded by

$$\begin{aligned}
&2\cdot\Bigl\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot\bigl(\epsilon_h^\tau(\widehat{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger)\bigr)\Bigr\|_{\Lambda_h^{-1}}^2 \\
&\quad = 2\cdot\sum_{\tau,\tau'=1}^K \phi(x_h^\tau, a_h^\tau)^\top \Lambda_h^{-1}\phi(x_h^{\tau'}, a_h^{\tau'})\cdot\bigl(\epsilon_h^\tau(\widehat{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger)\bigr)\cdot\bigl(\epsilon_h^{\tau'}(\widehat{V}_{h+1}) - \epsilon_h^{\tau'}(V_{h+1}^\dagger)\bigr) \\
&\quad \le 8\varepsilon^2\cdot\sum_{\tau,\tau'=1}^K \bigl|\phi(x_h^\tau, a_h^\tau)^\top \Lambda_h^{-1}\phi(x_h^{\tau'}, a_h^{\tau'})\bigr| \le 8\varepsilon^2\cdot\sum_{\tau,\tau'=1}^K \|\phi(x_h^\tau, a_h^\tau)\|\cdot\|\phi(x_h^{\tau'}, a_h^{\tau'})\|\cdot\|\Lambda_h^{-1}\|_{\mathrm{op}},
\end{aligned}$$

where the first inequality follows from Equation (E.14). As it holds that $\Lambda_h \succeq \lambda\cdot I$ by the definition of Λ_h in Equation (4.6) and $\|\phi(x,a)\| \le 1$ for all $(x,a) \in \mathcal{S}\times\mathcal{A}$ by Definition 4.3, for all $h \in [H]$, we have

$$2\cdot\Bigl\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot\bigl(\epsilon_h^\tau(\widehat{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger)\bigr)\Bigr\|_{\Lambda_h^{-1}}^2 \le 8\varepsilon^2 K^2/\lambda. \tag{E.16}$$

Combining Equations (E.15) and (E.16), for all $h \in [H]$, we have

$$\text{(iii)} \le 2\cdot\sup_{V\in\mathcal{N}_{h+1}(\varepsilon)}\Bigl\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot\epsilon_h^\tau(V)\Bigr\|_{\Lambda_h^{-1}}^2 + 8\varepsilon^2 K^2/\lambda. \tag{E.17}$$

Note that the right-hand side of Equation (E.17) does not involve the estimated value functions $\widehat{Q}_h$ and $\widehat{V}_{h+1}$, which are constructed based on the dataset $\mathcal{D}$. Hence, it allows us to upper bound the first term via uniform concentration inequalities. We utilize the following lemma to characterize the first term for any fixed function $V \in \mathcal{N}_{h+1}(\varepsilon)$. Recall the definition of $\epsilon_h^\tau(V)$ in Equation (E.7). Also recall that $\mathbb{P}_\mathcal{D}$ is the joint distribution of the data collecting process.

Lemma E.2 (Concentration of Self-Normalized Processes)**.** Let $V : \mathcal{S} \to [0, H-1]$ be any fixed function. Under

Assumption 2.2, for any fixed $h \in [H]$ and any $\delta \in (0,1)$, we have

$$\mathbb{P}_{\mathcal{D}}\biggl(\Bigl\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(V)\Bigr\|_{\Lambda_h^{-1}}^2 > H^2\cdot\bigl(2\cdot\log(1/\delta)+d\cdot\log(1+K/\lambda)\bigr)\biggr)\le\delta.$$

Proof of Lemma E.2. For the fixed $h \in [H]$ and all $\tau \in \{0,\ldots,K\}$, we define the σ-algebra

$$\mathcal{F}_{h,\tau}=\sigma\bigl(\{(x_h^j,a_h^j)\}_{j=1}^{(\tau+1)\wedge K}\cup\{(r_h^j,x_{h+1}^j)\}_{j=1}^{\tau}\bigr),$$

where $\sigma(\cdot)$ denotes the σ-algebra generated by a set of random variables and $(\tau+1)\wedge K$ denotes $\min\{\tau+1,K\}$. For all $\tau\in[K]$, we have $\phi(x_h^\tau,a_h^\tau)\in\mathcal{F}_{h,\tau-1}$, as (x_h^τ,a_h^τ) is $\mathcal{F}_{h,\tau-1}$-measurable. Also, for the fixed function $V\colon\mathcal{S}\to[0,H-1]$ and all $\tau\in[K]$, we have

$$\epsilon_h^\tau(V)=r_h^\tau+V(x_{h+1}^\tau)-(\mathbb{B}_hV)(x_h^\tau,a_h^\tau)\in\mathcal{F}_{h,\tau},$$

as (r_h^τ,x_{h+1}^τ) is $\mathcal{F}_{h,\tau}$-measurable. Hence, $\{\epsilon_h^\tau(V)\}_{\tau=1}^K$ is a stochastic process adapted to the filtration $\{\mathcal{F}_{h,\tau}\}_{\tau=0}^K$. By Assumption 2.2, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\bigl[\epsilon_h^\tau(V)\,\big|\,\mathcal{F}_{h,\tau-1}\bigr]&=\mathbb{E}_{\mathcal{D}}\bigl[r_h^\tau+V(x_{h+1}^\tau)\,\big|\,\{(x_h^j,a_h^j)\}_{j=1}^{\tau},\{(r_h^j,x_{h+1}^j)\}_{j=1}^{\tau-1}\bigr]-(\mathbb{B}_hV)(x_h^\tau,a_h^\tau)\\&=\mathbb{E}\bigl[r_h(s_h,a_h)+V(s_{h+1})\,\big|\,s_h=x_h^\tau,a_h=a_h^\tau\bigr]-(\mathbb{B}_hV)(x_h^\tau,a_h^\tau)=0,\end{aligned}$$

where the second equality follows from Equation (2.7) and the last equality follows from the definition of the Bellman operator $\mathbb{B}_h$ in Equation (2.4). Here $\mathbb{E}_{\mathcal{D}}$ is taken with respect to $\mathbb{P}_{\mathcal{D}}$, while $\mathbb{E}$ is taken with respect to the immediate reward and next state in the underlying MDP.

Moreover, as it holds that $r_h^\tau\in[0,1]$ and $V\in[0,H-1]$, we have $r_h^\tau+V(x_{h+1}^\tau)\in[0,H]$. Meanwhile, we have $(\mathbb{B}_hV)(x_h^\tau,a_h^\tau)\in[0,H]$, which implies $|\epsilon_h^\tau(V)|\le H$. Hence, for the fixed $h\in[H]$ and all $\tau\in[K]$, the random variable $\epsilon_h^\tau(V)$ defined in Equation (E.7) is mean-zero and H-sub-Gaussian conditioning on $\mathcal{F}_{h,\tau-1}$.

We invoke Lemma G.2 with $M_0=\lambda\cdot I$ and $M_k=\lambda\cdot I+\sum_{\tau=1}^k\phi(x_h^\tau,a_h^\tau)\,\phi(x_h^\tau,a_h^\tau)^\top$ for all $k\in[K]$. For the fixed function $V\colon\mathcal{S}\to[0,H-1]$ and fixed $h\in[H]$, we have

$$\mathbb{P}_{\mathcal{D}}\biggl(\Bigl\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(V)\Bigr\|_{\Lambda_h^{-1}}^2>2H^2\cdot\log\Bigl(\frac{\det(\Lambda_h)^{1/2}}{\delta\cdot\det(\lambda\cdot I)^{1/2}}\Bigr)\biggr)\le\delta$$

for all $\delta\in(0,1)$. Here we use the fact that $M_K=\Lambda_h$. Note that $\|\phi(x,a)\|\le1$ for all $(x,a)\in\mathcal{S}\times\mathcal{A}$ by Definition 4.3. We have

$$\|\Lambda_h\|_{\mathrm{op}}=\Bigl\|\lambda\cdot I+\sum_{\tau=1}^K\phi(x_h^\tau,a_h^\tau)\phi(x_h^\tau,a_h^\tau)^\top\Bigr\|_{\mathrm{op}}\le\lambda+\sum_{\tau=1}^K\|\phi(x_h^\tau,a_h^\tau)\phi(x_h^\tau,a_h^\tau)^\top\|_{\mathrm{op}}\le\lambda+K,$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the matrix operator norm. Hence, it holds that $\det(\Lambda_h)\le(\lambda+K)^d$ and $\det(\lambda\cdot I)=\lambda^d$, which implies

$$\begin{aligned}&\mathbb{P}_{\mathcal{D}}\biggl(\Bigl\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(V)\Bigr\|_{\Lambda_h^{-1}}^2>H^2\cdot\bigl(2\cdot\log(1/\delta)+d\cdot\log(1+K/\lambda)\bigr)\biggr)\\&\quad\le\mathbb{P}_{\mathcal{D}}\biggl(\Bigl\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(V)\Bigr\|_{\Lambda_h^{-1}}^2>2H^2\cdot\log\Bigl(\frac{\det(\Lambda_h)^{1/2}}{\delta\cdot\det(\lambda\cdot I)^{1/2}}\Bigr)\biggr)\le\delta.\end{aligned}$$

Therefore, we conclude the proof of Lemma E.2. □

Applying Lemma E.2 and the union bound, for any fixed $h\in[H]$, we have

$$\mathbb{P}_{\mathcal{D}}\biggl(\sup_{V\in\mathcal{N}_{h+1}(\varepsilon)}\Bigl\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(V)\Bigr\|_{\Lambda_h^{-1}}^2>H^2\cdot\bigl(2\cdot\log(1/\delta)+d\cdot\log(1+K/\lambda)\bigr)\biggr)\le\delta\cdot|\mathcal{N}_{h+1}(\varepsilon)|.$$

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

$$\sup_{V \in \mathcal{N}_{h+1}(\varepsilon)} \Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \epsilon_h^\tau(V) \Big\|^2_{\Lambda_h^{-1}} \\ \leq H^2 \cdot \Big(2 \cdot \log\big(H \cdot |\mathcal{N}_{h+1}(\varepsilon)|/\xi\big) + d \cdot \log(1 + K/\lambda)\Big) \tag{E.18}$$

with probability at least $1 - \xi/H$, which is taken with respect to $\mathbb{P}_{\mathcal{D}}$. Combining Equations (E.17) and (E.18), we have

$$\mathbb{P}_{\mathcal{D}}\bigg(\bigcap_{h \in [H]} \bigg\{ \Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \epsilon_h^\tau(\widehat{V}_{h+1}) \Big\|^2_{\Lambda_h^{-1}} \tag{E.19} \\ \leq 2H^2 \cdot \Big(2 \cdot \log\big(H \cdot |\mathcal{N}_{h+1}(\varepsilon)|/\xi\big) + d \cdot \log(1 + K/\lambda)\Big) + 8\varepsilon^2 K^2/\lambda \bigg\} \bigg) \geq 1 - \xi,$$

which follows from the union bound.

It remains to choose a proper $\varepsilon > 0$ and upper bound the ε-covering number $|\mathcal{N}_{h+1}(\varepsilon)|$. In the sequel, we set $\varepsilon = dH/K$ and $\lambda = 1$. By Equation (E.19), for all $h \in [H]$, it holds that

$$\Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \epsilon_h^\tau(\widehat{V}_{h+1}) \Big\|^2_{\Lambda_h^{-1}} \leq 2H^2 \cdot \Big(2 \cdot \log\big(H \cdot |\mathcal{N}_{h+1}(\varepsilon)|/\xi\big) + d \cdot \log(1 + K) + 4d^2\Big) \tag{E.20}$$

with probability at least $1 - \xi$, which is taken with respect to $\mathbb{P}_{\mathcal{D}}$. To upper bound $|\mathcal{N}_{h+1}(\varepsilon)|$, we utilize the following lemma, which is obtained from (Jin et al., 2020). Recall the definition of the function class $\mathcal{V}_h(R, B, \lambda)$ in Equation (E.9). Also, recall that $\mathcal{N}_h(\varepsilon; R, B, \lambda)$ is the minimal ε-cover of $\mathcal{V}_h(R, B, \lambda)$ with respect to the supremum norm.

Lemma E.3 (ε-Covering Number (Jin et al., 2020))**.** For all $h \in [H]$ and all $\varepsilon > 0$, we have

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

Proof of Lemma E.3. See Lemma D.6 in (Jin et al., 2020) for a detailed proof. □

Recall that

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

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. Recall that $\mathcal{N}_{h+1}(\varepsilon) = \mathcal{N}_{h+1}(\varepsilon; R_0, B_0, \lambda)$ is the minimal ε-cover of $\mathcal{V}_{h+1} = \mathcal{V}_{h+1}(R_0, B_0, \lambda)$ with respect to the supremum norm. Applying Lemma E.3 with $\varepsilon = dH/K$, we have

$$\log |\mathcal{N}_{h+1}(\varepsilon)| \leq d \cdot \log(1 + 4d^{-1/2}K^{3/2}) + d^2 \cdot \log(1 + 32c^2 \cdot d^{1/2}K^2\zeta) \\ \leq d \cdot \log(1 + 4d^{1/2}K^2) + d^2 \cdot \log(1 + 32c^2 \cdot d^{1/2}K^2\zeta). \tag{E.21}$$

As it holds that $\zeta > 1$, we set $c \geq 1$ to ensure that the second term on the right-hand side of Equation (E.21) is the dominating term, where $32c^2 \cdot d^{1/2}K^2\zeta \geq 1$. Hence, we have

$$\log |\mathcal{N}_{h+1}(\varepsilon)| \leq 2d^2 \cdot \log(1 + 32c^2 \cdot d^{1/2}K^2\zeta) \leq 2d^2 \cdot \log(64c^2 \cdot d^{1/2}K^2\zeta). \tag{E.22}$$

By Equations (E.20) and (E.22), for all $h \in [H]$, it holds that

$$\Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \epsilon_h^\tau(\widehat{V}_{h+1}) \Big\|^2_{\Lambda_h^{-1}} \\ \leq 2H^2 \cdot \big(2 \cdot \log(H/\xi) + 4d^2 \cdot \log(64c^2 \cdot d^{1/2}K^2\zeta) + d \cdot \log(1 + K) + 4d^2\big) \tag{E.23}$$

with probability at least $1 - \xi$, which is taken with respect to $\mathbb{P}_{\mathcal{D}}$. Note that $\log(1 + K) \leq \log(2K) \leq \zeta$ and $\log \zeta \leq \zeta$. Hence, we have

$$2 \cdot \log(H/\xi) + 4d^2 \cdot \log(d^{1/2}K^2\zeta) + d \cdot \log(1 + K) + 4d^2 \\ \leq 2d^2 \cdot \log(dHK^4/\xi) + d\zeta + 8d^2\zeta \leq 18d^2\zeta.$$

As it holds that $\zeta > 1$ and $\log \zeta \leq \zeta$, Equation (E.23) implies

$$\Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \epsilon_h^\tau(\widehat{V}_{h+1}) \Big\|_{\Lambda_h^{-1}}^2 \leq d^2 H^2 \zeta \cdot \big(36 + 8 \cdot \log(64c^2)\big). \tag{E.24}$$

We set $c \geq 1$ to be sufficiently large, which ensures that $36 + 8 \cdot \log(64c^2) \leq c^2/4$ on the right-hand side of Equation (E.24). By Equations (E.8) and (E.24), for all $h \in [H]$, it holds that

$$|\text{(ii)}| \leq c/2 \cdot dH\sqrt{\zeta} \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)} = \beta/2 \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)} \tag{E.25}$$

with probability at least $1 - \xi$, which is taken with respect to $\mathbb{P}_{\mathcal{D}}$.

By Equations (4.7), (E.5), (E.6), and (E.25), for all $h \in [H]$ and all $(x,a) \in \mathcal{S} \times \mathcal{A}$, it holds that

$$\big|(\mathbb{B}_h \widehat{V}_{h+1})(x,a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x,a)\big| \leq (H\sqrt{d} + \beta/2) \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)} \leq \Gamma_h(x,a)$$

with probability at least $1 - \xi$, which is taken with respect to $\mathbb{P}_{\mathcal{D}}$. In other words, $\{\Gamma_h\}_{h=1}^H$ defined in Equation (4.7) are ξ-uncertainty quantifiers. Therefore, we conclude the proof of Lemma C.2. □

E.3. Proof of Corollary 4.5

Proof of Corollary 4.5. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\pi^*}\Big[\big(\phi(s_h,a_h)^\top \Lambda_h^{-1} \phi(s_h,a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big] \\
&\quad = \mathbb{E}_{\pi^*}\Big[\sqrt{\text{Tr}\left(\phi(s_h,a_h)^\top \Lambda_h^{-1} \phi(s_h,a_h)\right)} \,\Big|\, s_1 = x\Big] \\
&\quad = \mathbb{E}_{\pi^*}\Big[\sqrt{\text{Tr}\left(\phi(s_h,a_h)\phi(s_h,a_h)^\top \Lambda_h^{-1}\right)} \,\Big|\, s_1 = x\Big] \\
&\quad \leq \sqrt{\text{Tr}\Big(\mathbb{E}_{\pi^*}\big[\phi(s_h,a_h)\phi(s_h,a_h)^\top \,\big|\, s_1 = x\big] \Lambda_h^{-1}\Big)}
\end{aligned} \tag{E.26}$$

for all $x \in \mathcal{S}$ and all $h \in [H]$. We define the event

$$\mathcal{E}^\ddagger = \bigg\{ \text{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) \leq 2\beta \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h,a_h)^\top \Lambda_h^{-1} \phi(s_h,a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big] \text{ for all } x \in \mathcal{S} \bigg\}. \tag{E.27}$$

For notational simplicity, we define

$$\Sigma_h(x) = \mathbb{E}_{\pi^*}\big[\phi(s_h,a_h)\phi(s_h,a_h)^\top \,\big|\, s_1 = x\big]$$

for all $x \in \mathcal{S}$ and all $h \in [H]$. On the event $\mathcal{E}^\dagger \cap \mathcal{E}^\ddagger$, where $\mathcal{E}^\dagger$ and $\mathcal{E}^\ddagger$ are defined in Equations (4.9) and (E.27), respectively, we have

$$\begin{aligned}
\text{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) &\leq 2\beta \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h,a_h)^\top \Lambda_h^{-1} \phi(s_h,a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big] \\
&\leq 2\beta \sum_{h=1}^{H} \sqrt{\text{Tr}\Big(\Sigma_h(x) \cdot \big(I + c^\dagger \cdot K \cdot \Sigma_h(x)\big)^{-1}\Big)} \\
&= 2\beta \sum_{h=1}^{H} \sqrt{\sum_{j=1}^{d} \frac{\lambda_{h,j}(x)}{1 + c^\dagger \cdot K \cdot \lambda_{h,j}(x)}}.
\end{aligned}$$

Here $\{\lambda_{h,j}(x)\}_{j=1}^d$ are the eigenvalues of $\Sigma_h(x)$ for all $x \in \mathcal{S}$ and all $h \in [H]$, the first inequality follows from the definition of $\mathcal{E}^\ddagger$ in Equation (E.27), and the second inequality follows from Equation (E.26) and the definition of $\mathcal{E}^\dagger$ in Equation (4.9). Meanwhile, by Definition 4.3, we have $\|\phi(s,a)\| \leq 1$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$. By Jensen's inequality, we have

$$\|\Sigma_h(x)\|_{\text{op}} \leq \mathbb{E}_{\pi^*}\big[\|\phi(s_h,a_h)\phi(s_h,a_h)^\top\|_{\text{op}} \,\big|\, s_1 = x\big] \leq 1$$

for all $x \in \mathcal{S}$ and all $h \in [H]$. As $\Sigma_h(x)$ is positive semidefinite, we have $\lambda_{h,j}(x) \in [0,1]$ for all $x \in \mathcal{S}$, all $h \in [H]$, and all $j \in [d]$. Hence, on $\mathcal{E}^\dagger \cap \mathcal{E}^\ddagger$, we have

$$\begin{aligned}\text{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) &\le 2\beta \sum_{h=1}^H \sqrt{\sum_{j=1}^d \frac{\lambda_{h,j}(x)}{1 + c^\dagger \cdot K \cdot \lambda_{h,j}(x)}} \\ &\le 2\beta \sum_{h=1}^H \sqrt{\sum_{j=1}^d \frac{1}{1 + c^\dagger \cdot K}} \le c' \cdot d^{3/2} H^2 K^{-1/2} \sqrt{\zeta}\end{aligned}$$

for all $x \in \mathcal{S}$, where the second inequality follows from the fact that $\lambda_{h,j}(x) \in [0,1]$ for all $x \in \mathcal{S}$, all $h \in [H]$, and all $j \in [d]$, while the third inequality follows from the choice of the scaling parameter $\beta > 0$ in Corollary 4.5. Here we define the absolute constant $c' = 2c/\sqrt{c^\dagger} > 0$, where $c^\dagger > 0$ is the absolute constant used in Equation (4.9) and $c > 0$ is the absolute constant used in Theorem 4.4. By the condition in Corollary 4.5, we have $\mathbb{P}_\mathcal{D}(\mathcal{E}^\dagger) \ge 1 - \xi/2$. Also, by Theorem 4.4, we have $\mathbb{P}_\mathcal{D}(\mathcal{E}^\ddagger) \ge 1 - \xi/2$. Hence, by the union bound, we have $\mathbb{P}_\mathcal{D}(\mathcal{E}^\dagger \cap \mathcal{E}^\ddagger) \ge 1 - \xi$, which yields Equation (4.10).

In particular, if $\mathrm{rank}(\Sigma_h(x)) \le r$ for all $x \in \mathcal{S}$ and all $h \in [H]$, on $\mathcal{E}^\dagger \cap \mathcal{E}^\ddagger$, which satisfies $\mathbb{P}_\mathcal{D}(\mathcal{E}^\dagger \cap \mathcal{E}^\ddagger) \ge 1 - \xi$, we have

$$\begin{aligned}\text{SubOpt}\big(\texttt{Pess}(\mathcal{D}); x\big) &\le 2\beta \sum_{h=1}^H \sqrt{\sum_{j=1}^d \frac{\lambda_{h,j}(x)}{1 + c^\dagger \cdot K \cdot \lambda_{h,j}(x)}} \\ &\le 2\beta \sum_{h=1}^H \sqrt{\sum_{j=1}^r \frac{1}{1 + c^\dagger \cdot K}} \le c'' \cdot d H^2 K^{-1/2} \sqrt{\zeta},\end{aligned}$$

where $c'' = 2c\sqrt{r/c^\dagger} > 0$ is an absolute constant. Here the second inequality follows from the fact that $\lambda_{h,j}(x) \in [0,1]$ and $\mathrm{rank}(\Sigma_h(x)) \le r$ for all $x \in \mathcal{S}$, all $h \in [H]$, and all $j \in [d]$, while the third inequality follows from the choice of β in Corollary 4.5. Hence, we obtain Equation (4.11). Therefore, we conclude the proof of Corollary 4.5. □

E.4. Proof of Corollary B.1

Proof of Corollary B.1. For all $h \in [H]$ and all $\tau \in [K]$, we define the random matrices

$$Z_h = \sum_{\tau=1}^K A_h^\tau, \qquad A_h^\tau = \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top - \Sigma_h,$$
$$\text{where } \Sigma_h = \mathbb{E}_{\bar\pi}\big[\phi(s_h, a_h)\phi(s_h, a_h)^\top\big]. \tag{E.28}$$

For all $h \in [H]$ and all $\tau \in [K]$, Equation (E.28) implies $\mathbb{E}_{\bar\pi}[A_h^\tau] = 0$. Here $\mathbb{E}_{\bar\pi}$ is taken with respect to the trajectory induced by the fixed behavior policy $\bar\pi$ in the underlying MDP. As the K trajectories in the dataset $\mathcal{D}$ are i.i.d., for all $h \in [H]$, $\{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau=1}^K$ are also i.i.d.. Hence, for all $h \in [H]$, $\{A_h^\tau\}_{\tau=1}^K$ are i.i.d. and centered.

By Definition 4.3, we have $\|\phi(x,a)\| \le 1$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$. By Jensen's inequality, we have

$$\|\Sigma_h\|_{\text{op}} \le \mathbb{E}_{\bar\pi}\big[\|\phi(s_h, a_h)\phi(s_h, a_h)^\top\|_{\text{op}}\big] \le 1.$$

For any vector $v \in \mathbb{R}^d$ with $\|v\| = 1$, the triangle inequality implies

$$\|A_h^\tau v\| \le \|\phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top v\| + \|\Sigma_h v\| \le \|v\| + \|\Sigma_h\|_{\text{op}} \cdot \|v\| \le 2.$$

Hence, for all $h \in [H]$ and all $\tau \in [K]$, we have

$$\|A_h^\tau\|_{\text{op}} \le 2, \qquad \|A_h^\tau (A_h^\tau)^\top\|_{\text{op}} \le \|A_h^\tau\|_{\text{op}} \cdot \|(A_h^\tau)^\top\|_{\text{op}} \le 4.$$

As $\{A_h^\tau\}_{\tau=1}^K$ are i.i.d. and centered, for all $h \in [H]$, we have

$$\begin{aligned}\|\mathbb{E}_{\bar\pi}[Z_h Z_h^\top]\|_{\text{op}} &= \Big\| \sum_{\tau=1}^K \mathbb{E}_{\bar\pi}[A_h^\tau (A_h^\tau)^\top] \Big\|_{\text{op}} \\ &= K \cdot \|\mathbb{E}_{\bar\pi}[A_h^1 (A_h^1)^\top]\|_{\text{op}} \le K \cdot \mathbb{E}_{\bar\pi}\big[\|A_h^1 (A_h^1)^\top\|_{\text{op}}\big] \le 4K,\end{aligned}$$

where the first inequality follows from Jensen's inequality. Similarly, for all $h \in [H]$ and all $\tau \in [K]$, as it holds that

$$\|(A_h^\tau)^\top A_h^\tau\|_{\text{op}} \le \|(A_h^\tau)^\top\|_{\text{op}} \cdot \|A_h^\tau\|_{\text{op}} \le 4,$$

we have

$$\|\mathbb{E}_{\bar{\pi}}[Z_h^\top Z_h]\|_{\text{op}} \le 4K.$$

Applying Lemma G.1 to Z_h defined in Equation (E.28), for any fixed $h \in [H]$ and any $t \ge 0$, we have

$$\mathbb{P}_{\mathcal{D}}\big(\|Z_h\|_{\text{op}} > t\big) = \mathbb{P}_{\mathcal{D}}\bigg(\Big\|\sum_{\tau=1}^K A_h^\tau\Big\|_{\text{op}} > t\bigg) \le 2d \cdot \exp\Big(-\frac{t^2/2}{4K + 2t/3}\Big). \tag{E.29}$$

For all $\xi \in (0,1)$, we set $t = \sqrt{10K \cdot \log(4dH/\xi)}$. By Equation (E.29), when K is sufficiently large so that $K \ge 5 \cdot \log(4dH/\xi)$, we have $2t/3 \le K$. Hence, for the fixed $h \in [H]$, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{D}}\big(\|Z_h\|_{\text{op}} \le t\big) &\ge 1 - 2d \cdot \exp\big(-t^2/(8K + 4t/3)\big)\\ &\ge 1 - 2d \cdot \exp\big(-t^2/(10K)\big) = 1 - \xi/(2H).\end{aligned} \tag{E.30}$$

By Equation (E.30) and the union bound, for all $h \in [H]$, it holds that

$$\|Z_h/K\|_{\text{op}} = \Big\|\frac{1}{K}\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top - \Sigma_h\Big\|_{\text{op}} \le \sqrt{10/K \cdot \log(4dH/\xi)} \tag{E.31}$$

with probability at least $1 - \xi/2$, which is taken with respect to $\mathbb{P}_{\mathcal{D}}$.

By the definition of Z_h in Equation (E.28), we have

$$Z_h = \sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top - K \cdot \Sigma_h = (\Lambda_h - \lambda \cdot I) - K \cdot \Sigma_h. \tag{E.32}$$

Recall that there exists an absolute constant $\underline{c} > 0$ such that $\lambda_{\min}(\Sigma_h) \ge \underline{c}$, which implies $\|\Sigma_h^{-1}\|_{\text{op}} \le 1/\underline{c}$. By Equations (E.31) and (E.32), when K is sufficiently large so that $K \ge 40/\underline{c} \cdot \log(4dH/\xi)$, for all $h \in [H]$, it holds that

$$\begin{aligned}\lambda_{\min}(\Lambda_h/K) &= \lambda_{\min}(\Sigma_h + \lambda/K \cdot I + Z_h/K)\\ &\ge \lambda_{\min}(\Sigma_h) - \|Z_h/K\|_{\text{op}} \ge \underline{c} - \sqrt{10/K \cdot \log(4dH/\xi)} \ge \underline{c}/2.\end{aligned}$$

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

$$\|\Lambda_h^{-1}\|_{\text{op}} \le \big(K \cdot \lambda_{\min}(\Lambda_h/K)\big)^{-1} \le 2/(K \cdot \underline{c})$$

with probability at least $1 - \xi/2$ with respect to $\mathbb{P}_{\mathcal{D}}$, which implies

$$\sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)} \le \|\phi(x,a)\| \cdot \|\Lambda_h^{-1}\|_{\text{op}}^{1/2} \le c''/\sqrt{K}, \qquad \forall (x,a) \in \mathcal{S} \times \mathcal{A},\ \forall h \in [H]. \tag{E.33}$$

Here we define the absolute constant $c'' = \sqrt{2/\underline{c}}$ and use the fact that $\|\phi(x,a)\| \le 1$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$ in Definition 4.3.

We define the event

$$\mathcal{E}_1^* = \Big\{\sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)} \le c''/\sqrt{K} \text{ for all } (x,a) \in \mathcal{S} \times \mathcal{A} \text{ and all } h \in [H]\Big\}.$$

By Equation (E.33), we have $\mathbb{P}_{\mathcal{D}}(\mathcal{E}_1^*) \ge 1 - \xi/2$ for $K \ge 40/\underline{c} \cdot \log(4dH/\xi)$. Also, we define the event

$$\mathcal{E}_2^* = \bigg\{\text{SubOpt}(\widehat{\pi}; x) \le 2\beta \cdot \sum_{h=1}^H \mathbb{E}_{\pi^*}\Big[\sqrt{\phi(s_h,a_h)^\top \Lambda_h^{-1} \phi(s_h,a_h)}\,\Big|\, s_1 = x\Big] \text{ for all } x \in \mathcal{S}\bigg\}.$$

Here we set $\beta = c \cdot dH\sqrt{\log(4dHK/\xi)}$, where $c > 0$ is the same absolute constant as in Theorem 4.4. By Theorem 4.4, we have $\mathbb{P}_{\mathcal{D}}(\mathcal{E}_2^*) \ge 1 - \xi/2$. Hence, when K is sufficiently large so that $K \ge 40/\underline{c} \cdot \log(4dH/\xi)$, on the event $\mathcal{E}^* = \mathcal{E}_1^* \cap \mathcal{E}_2^*$, we have

$$\text{SubOpt}(\widehat{\pi}; x) \le 2\beta \cdot H \cdot c''/\sqrt{K} = c' \cdot dH^2\sqrt{\log(4dHK/\xi)/K}, \qquad \forall x \in \mathcal{S}.$$

By the union bound, we have $\mathbb{P}_{\mathcal{D}}(\mathcal{E}^*) \ge 1 - \xi$ with $c' = 2c \cdot c''$, where $c'' = \sqrt{2/\underline{c}}$ and $c > 0$ is the same absolute constant as in Theorem 4.4. Therefore, we conclude the proof of Corollary B.1. □

F. Proofs of Minimax Optimality

F.1. Proof of Lemma C.3

Proof of Lemma C.3. We consider two linear MDPs $\mathcal{M}_1 = M(p^*, p, p)$ and $\mathcal{M}_2 = M(p, p^*, p)$ in the class $\mathfrak{M}$ defined in Equation (C.6). As we have $p^* > p$, by Equations (C.7) and (C.8), the optimal policy for $\mathcal{M}_1$ satisfies $\pi_1^{*,1}(a_1 \,|\, x_0) = \mathbb{1}\{a_1 = b_1\}$, which always chooses the action b_1 for step $h = 1$, while the optimal policy for $\mathcal{M}_2$ satisfies $\pi_1^{*,2}(a_1 \,|\, x_0) = \mathbb{1}\{a_1 = b_2\}$, which always chooses the action b_2 for step $h = 1$. Given the dataset $\mathcal{D}$, we denote by $\pi = \{\pi_h\}_{h=1}^H = \texttt{Algo}(\mathcal{D})$ the output of any RL algorithm. Recall that $\sum_{j=1}^A \pi_1(b_j \,|\, x_0) = 1$. By Equation (C.11), the suboptimality of π for $\mathcal{M}_1$ is

$$
\begin{aligned}
\text{SubOpt}(\mathcal{M}_1, \pi; x_0) &= \Big(p^* - p^* \cdot \pi_1(b_1 \,|\, x_0) - \sum_{j=2}^A p \cdot \pi_1(b_j \,|\, x_0)\Big) \cdot (H-1) \\
&= (p^* - p) \cdot \big(1 - \pi_1(b_1 \,|\, x_0)\big) \cdot (H-1). \qquad \text{(F.1)}
\end{aligned}
$$

Similarly, the suboptimality of π for $\mathcal{M}_2$ is

$$
\text{SubOpt}(\mathcal{M}_2, \pi; x_0) = (p^* - p) \cdot \big(1 - \pi_1(b_2 \,|\, x_0)\big) \cdot (H-1). \qquad \text{(F.2)}
$$

Recall that we define $n_j = \sum_{\tau=1}^K \mathbb{1}\{a_1^\tau = b_j\}$ for all $j \in [A]$. Combining Equations (F.1) and (F.2), we have

$$
\begin{aligned}
&\max_{\ell \in \{1,2\}} \sqrt{n_\ell} \cdot \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Big[\text{SubOpt}\big(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\big)\Big] \\
&\quad \geq \frac{\sqrt{n_1 n_2}}{\sqrt{n_1} + \sqrt{n_2}} \cdot \bigg(\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\Big[\text{SubOpt}\big(\mathcal{M}_1, \texttt{Algo}(\mathcal{D}); x_0\big)\Big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\Big[\text{SubOpt}\big(\mathcal{M}_2, \texttt{Algo}(\mathcal{D}); x_0\big)\Big]\bigg) \\
&\quad = \frac{\sqrt{n_1 n_2}}{\sqrt{n_1} + \sqrt{n_2}} \cdot (p^* - p) \cdot (H-1) \cdot \Big(\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\big[1 - \pi_1(b_2 \,|\, x_0)\big]\Big) \\
&\quad \geq \frac{\sqrt{n_1 n_2}}{\sqrt{n_1} + \sqrt{n_2}} \cdot (p^* - p) \cdot (H-1) \cdot \Big(\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\big[\pi_1(b_1 \,|\, x_0)\big]\Big),
\end{aligned}
$$

where $\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}$ is the expectation taken with respect to the randomness of $\mathcal{D}$, which is compliant with the underlying MDP $\mathcal{M}_\ell$ for all $\ell \in \{1,2\}$. Here the first inequality follows from the fact that $\max\{x, y\} \geq a \cdot x + (1-a) \cdot y$, for all $a \in [0,1]$ and all $x, y \geq 0$. Therefore, we conclude the proof of Lemma C.3. □

F.2. Suboptimality of PEVI on $\mathfrak{M}$

In this section, we establish the suboptimality of PEVI for the linear MDPs in $\mathfrak{M}$. We consider any linear MDP $\mathcal{M} = M(p_1, p_2, p_3) \in \mathfrak{M}$ and the dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$ compliant with $\mathcal{M}$, which is constructed in Section C.3.1. Recall that $n_j = \sum_{\tau=1}^K \mathbb{1}\{a_1^\tau = b_j\}$ for all $j \in [A]$ and $j^* = \operatorname{argmax}_{j\in[A]} p_j$. We define $m_j = \sum_{\tau=1}^K \mathbb{1}\{x_2^\tau = x_j\}$ for all $j \in \{1,2\}$.

Lemma F.1 (Suboptimality of PEVI)**.** Suppose Assumption 2.2 holds and the underlying MDP is $\mathcal{M} \in \mathfrak{M}$. In Algorithm 2, we set $\lambda = 1$ and $\beta = c \cdot dH\sqrt{\log(4dHK/\xi)}$. Here $c > 0$ is an absolute constant and $\xi \in (0,1)$ is the confidence parameter, which are specified in Theorem 4.4. We have

$$
\begin{aligned}
&\sum_{h=1}^H \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x_0\Big] \\
&\quad = \frac{1}{\sqrt{1 + n_{j^*}}} + (H-1) \cdot \Big(\frac{p_{j^*}}{\sqrt{1+m_1}} + \frac{1 - p_{j^*}}{\sqrt{1+m_2}}\Big), \qquad \text{(F.3)}
\end{aligned}
$$

where $\mathbb{E}_{\pi^*}$ is taken with repsect to the trajectory induced by the optimal policy π^* in $\mathcal{M}$. When K is sufficiently large so that $K \geq 32 \cdot \log(8/\xi)$, $\texttt{Pess}(\mathcal{D})$ in Algorithm 2 satisfies

$$
\text{SubOpt}\big(\mathcal{M}, \texttt{Pess}(\mathcal{D}); x_0\big) \leq 9\beta H / \sqrt{n_{j^*}} \qquad \text{(F.4)}
$$

with probability at least $1 - \xi$, which is with respect to $\mathbb{P}_{\mathcal{D}}$.

Proof of Lemma F.1. Recall that $x_1^\tau = x_0$ for all $\tau \in [K]$. By the definition of Λ_h in Equation (4.6), we have

$$\Lambda_1 = \lambda \cdot I + \sum_{\tau=1}^{K} \phi(x_0, a_1^\tau)\phi(x_0, a_1^\tau)^\top = \mathrm{diag}(\lambda + n_1, \ldots, \lambda + n_A, \lambda, \lambda) \in \mathbb{R}^{(A+2)\times(A+2)}, \tag{F.5}$$

where the second equality follows from the definition of ϕ in Equation (C.9). Since $x_1, x_2 \in \mathcal{S}$ are the absorbing states, for all $a \in \mathcal{A}$ and all $h \in \{2, \ldots, H\}$, we have

$$\Lambda_h = \lambda \cdot I + \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top = \mathrm{diag}(\lambda, \ldots, \lambda, \lambda + m_1, \lambda + m_2) \in \mathbb{R}^{(A+2)\times(A+2)}, \tag{F.6}$$

where the second equality follows from the definition of ϕ in Equation (C.9). Also, we have

$$\mathbb{P}_{\pi^*}(s_2 = x_1) = p_{j^*}, \quad \text{and} \quad \mathbb{P}_{\pi^*}(s_2 = x_2) = 1 - p_{j^*},$$

where $\mathbb{P}_{\pi^*}$ is taken with respect to the trajectory induced by π^* in $\mathcal{M}$. Combining Equations (F.5) and (F.6), we have

$$\begin{aligned} &\mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x_0\Big] \\ &\quad = \begin{cases} (1 + n_{j^*})^{-1/2}, & h = 1, \\ p_{j^*} \cdot (1 + m_1)^{-1/2} + (1 - p_{j^*}) \cdot (1 + m_2)^{-1/2}, & h \in \{2, \ldots, H\}, \end{cases} \end{aligned} \tag{F.7}$$

which yields Equation (F.3). Here we use the definition of ϕ in Equation (C.9) and the parameter $\lambda = 1$ in Algorithm 2.

In the sequel, we lower bound m_1 and m_2 via concentration inequalities. By the construction of $\mathcal{D}$ in Section C.3.1, for all $\tau \in [K]$ and all $j \in [A]$, given the action $a_1^\tau = b_j$, $\mathbb{1}\{x_2^\tau = x_1\}$ is a Bernoulli random variable with the success probability p_j. As $p_1, p_2, p_3 \in [1/4, 3/4]$, we have

$$\mathbb{E}_{\mathcal{D}}[m_1] = \sum_{\tau=1}^{K} \mathbb{E}_{\mathcal{D}}\big[\mathbb{1}\{x_2^\tau = x_1\}\big] = \sum_{j=1}^{A} p_j \cdot n_j \geq 1/4 \cdot \sum_{j=1}^{A} n_j = K/4. \tag{F.8}$$

Given the actions $\{a_1^\tau\}_{\tau=1}^K$, m_1 is a sum of K independent Bernoulli random variables. By Hoeffding's inequality, for all $\xi > 0$, it holds that

$$\big|m_1 - \mathbb{E}_{\mathcal{D}}[m_1]\big| \leq \sqrt{K/2 \cdot \log(8/\xi)} \tag{F.9}$$

with probability at least $1 - \xi/4$, which is with respect to $\mathbb{P}_{\mathcal{D}}$. By Equations (F.8) and (F.9), it holds that

$$m_1 \geq K/4 - \sqrt{K/2 \cdot \log(8/\xi)} \tag{F.10}$$

with probability at least $1 - \xi/4$, which is with respect to $\mathbb{P}_{\mathcal{D}}$. Similarly, we have

$$\mathbb{E}_{\mathcal{D}}[m_2] = \sum_{\tau=1}^{K} \mathbb{E}_{\mathcal{D}}\big[\mathbb{1}\{x_2^\tau = x_2\}\big] = \sum_{j=1}^{A} (1 - p_j) \cdot n_j \geq 1/4 \cdot \sum_{j=1}^{A} n_j \geq K/4.$$

By Hoeffding's inequality, it holds that

$$m_2 \geq K/4 - \sqrt{K/2 \cdot \log(8/\xi)} \tag{F.11}$$

with probability at least $1 - \xi/4$, which is taken with respect to $\mathbb{P}_{\mathcal{D}}$. We define the event

$$\overline{\mathcal{E}} = \{m_1 \geq K/8,\ m_2 \geq K/8\}. \tag{F.12}$$

Combining Equations (F.10) and (F.11), by the union bound, when K is sufficiently large so that $K \geq 32 \cdot \log(8/\xi)$, we have $\mathbb{P}_{\mathcal{D}}(\overline{\mathcal{E}}) \geq 1 - \xi/2$.

Meanwhile, by Theorem 4.4 with the parameter $\lambda = 1$ and the confidence parameter $\xi/2$, it holds that

$$\text{SubOpt}\big(\mathcal{M}, \texttt{Pess}(\mathcal{D}); x_0\big) \leq \frac{2\beta}{\sqrt{1 + n_{j^*}}} + 2\beta(H - 1) \cdot \Big(\frac{p_{j^*}}{\sqrt{1 + m_1}} + \frac{1 - p_{j^*}}{\sqrt{1 + m_2}}\Big) \tag{F.13}$$

with probability at least $1 - \xi/2$, which is taken with respect to $\mathbb{P}_{\mathcal{D}}$. By the union bound on the two events defined in

Equations (F.12) and (F.13), respectively, it holds that

$$\begin{aligned}\mathrm{SubOpt}\big(\mathcal{M}, \mathtt{Pess}(\mathcal{D}); x_0\big) &\le \frac{2\beta}{\sqrt{n_{j^*}}} + 2\beta(H-1)\cdot\Big(\frac{3/4}{\sqrt{m_1}} + \frac{3/4}{\sqrt{m_2}}\Big) \\ &\le 2\beta/\sqrt{n_{j^*}} + 9\beta(H-1)/\sqrt{K} \le 9\beta H/\sqrt{n_{j^*}}\end{aligned}$$

with probability at least $1-\xi$, which is with respect to $\mathbb{P}_{\mathcal{D}}$. Here the first inequality follows from the fact that $p_{j^*} \in [1/4, 3/4]$, the second inequality follows from the fact that $m_1, m_2 \ge K/8$ on $\overline{\mathcal{E}}$ defined in (F.12), while the last inequality follows from the fact that $n_{j^*} \le K$. Therefore, we conclude the proof of Lemma F.1. □

F.3. Proof of Theorem 4.6

Proof of Theorem 4.6. We consider two linear MDPs $\mathcal{M}_1 = M(p^*, p, p)$ and $\mathcal{M}_2 = (p, p^*, p)$ in the class $\mathfrak{M}$ and the dataset $\mathcal{D}$ compliant with $\mathcal{M}_1$ or $\mathcal{M}_2$, which is constructed in Section C.3.1. We additionally assume that $n_1, n_2 \ge 4$ and $1/\bar{c} \le n_1/n_2 \le \bar{c}$ for an absolute constant $\bar{c} > 0$. For the policy $\pi = \{\pi_h\}_{h=1}^H = \mathtt{Algo}(\mathcal{D})$ constructed by any offline RL algorithm, recall the test function $\psi_{\mathtt{Algo}}(\mathcal{D})$ defined in Equation (C.17), which is constructed for the hypothesis testing problem defined in Equation (C.16). By Equation (C.18), we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\big[\pi_1(b_1 \,|\, x_0)\big] \\ &\quad = \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\big[\mathbb{1}\{\psi_{\mathtt{Algo}}(\mathcal{D}) = 1\}\big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\big[\mathbb{1}\{\psi_{\mathtt{Algo}}(\mathcal{D}) = 0\}\big] \\ &\quad \ge 1 - \mathrm{TV}(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1}, \mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}) \ge 1 - \sqrt{\mathrm{KL}(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1} \,\|\, \mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2})/2},\end{aligned} \tag{F.14}$$

where the first inequality follows from the definition of the total variation distance, while the second inequality follows from Pinsker's inequality. Here for each $\ell \in \{1, 2\}$, we use $\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_\ell}$ is with respect to the randomness of $\mathcal{D}$ when $\mathcal{D}$ is compliant with $\mathcal{M}_\ell$ for all $\ell \in \{1, 2\}$. Also, we use TV and KL to denote the total variation and the Kullback-Leibler (KL-) divergence, respectively.

Recall the mapping of the rewards $\{r_2^\tau\}_{\tau\in[N]}$ into the relabeled rewards $\{\kappa_j^i\}_{i,j=1}^{n_j, A}$ in Equation (C.12). Also, recall that the actions $\{a_h^\tau\}_{\tau\in[K], h\ge 2}$ are chosen arbitrarily but fixed in the dataset $\mathcal{D}$. Since $x_1, x_2 \in \mathcal{S}$ are absorbing states, we have $\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_\ell}(\mathcal{D}) = \mathbb{P}_{\mathcal{D}\sim\mathcal{M}_\ell}(\mathcal{D}_1)$ for all $\ell \in \{1, 2\}$, where $\mathcal{D}_1 = \{(x_1^\tau, a_1^\tau, x_2^\tau, r_2^\tau)\}_{\tau\in[K]}$ is the reduced dataset. By Equation (C.14), the probabilities of observing $\mathcal{D}_1$ under $\mathcal{M}_1$ and $\mathcal{M}_2$ take the form

$$\begin{aligned}\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1}(\mathcal{D}_1) &= (p^*)^{\sum_{i=1}^{n_1}\kappa_1^i} \cdot (1-p^*)^{n_1 - \sum_{i=1}^{n_1}\kappa_1^i} \cdot \prod_{j=2}^{A}\Big(p^{\sum_{i=1}^{n_j}\kappa_j^i} \cdot (1-p)^{n_j - \sum_{i=1}^{n_j}\kappa_j^i}\Big), \\ \mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}(\mathcal{D}_2) &= (p^*)^{\sum_{i=1}^{n_2}\kappa_2^i} \cdot (1-p^*)^{n_2 - \sum_{i=1}^{n_2}\kappa_2^i} \cdot \prod_{\substack{j\in[A] \\ j\neq 2}}\Big(p^{\sum_{i=1}^{n_j}\kappa_j^i} \cdot (1-p)^{n_j - \sum_{i=1}^{n_j}\kappa_j^i}\Big),\end{aligned} \tag{F.15}$$

respectively. Here we use the fact that $p_1 = p^*, p_2 = p$ in $\mathcal{M}_1 = M(p^*, p, p)$, while $p_1 = p, p_2 = p^*$ in $\mathcal{M}_2 = M(p, p^*, p)$, where $p^* > p$. By Equation (F.15), we have

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1} \,\|\, \mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}) &= \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\bigg[\Big(\sum_{i=1}^{n_1}\kappa_1^i - \sum_{i=1}^{n_2}\kappa_2^i\Big)\cdot\log\frac{p^*\cdot(1-p)}{p\cdot(1-p^*)} + (n_1 - n_2)\cdot\log\frac{1-p^*}{1-p}\bigg] \\ &= (n_1 p^* - n_2 p)\cdot\log\frac{p^*\cdot(1-p)}{p\cdot(1-p^*)} + (n_1 - n_2)\cdot\log\frac{1-p^*}{1-p} \\ &= n_1\cdot\Big(p^*\cdot\log\frac{p^*}{p} + (1-p^*)\cdot\log\frac{1-p^*}{1-p}\Big) + n_2\cdot\Big(p\cdot\log\frac{p}{p^*} + (1-p)\cdot\log\frac{1-p}{1-p^*}\Big),\end{aligned}$$

where the second equality follows from Equation (C.13). Note that for all $x \in (-1, 1)$, it holds that $\log(1+x) \le x$. Hence, when $p^* - p < \min\{p, 1-p\}$, we have

$$\begin{aligned}p^*\cdot\log\frac{p^*}{p} + (1-p^*)\cdot\log\frac{1-p^*}{1-p} &= p^*\cdot\log\Big(1 + \frac{p^*-p}{p}\Big) + (1-p^*)\cdot\log\Big(1 + \frac{p-p^*}{1-p}\Big) \\ &\le p^*\cdot\frac{p^*-p}{p} + (1-p^*)\cdot\frac{p-p^*}{1-p} = \frac{(p^*-p)^2}{p\cdot(1-p)}.\end{aligned}$$

Similarly, when $p^* - p < \min\{p^*, 1-p^*\}$, we have

$$p \cdot \log \frac{p}{p^*} + (1-p) \cdot \log \frac{1-p}{1-p^*} \leq \frac{(p^*-p)^2}{p^* \cdot (1-p^*)}.$$

Recall that $n_1, n_2 \geq 4$ and $1/\bar{c} \leq n_1/n_2 \leq \bar{c}$ for an absolute constant $\bar{c} > 0$. We set

$$p^* = \frac{1}{2} + \frac{1}{8} \cdot \sqrt{\frac{3}{2 \cdot (n_1+n_2)}}, \qquad p = \frac{1}{2} - \frac{1}{8} \cdot \sqrt{\frac{3}{2 \cdot (n_1+n_2)}} \tag{F.16}$$

such that $p^*, p \in [1/4, 3/4]$, $0 \leq p^* - p \leq 1/4$ and $p^* - p < \min\{p, 1-p, p^*, 1-p^*\}$. Hence, the KL-divergence is upper bounded as

$$\mathrm{KL}(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1} \,\|\, \mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}) \leq \frac{n_1 \cdot (p^*-p)^2}{p \cdot (1-p)} + \frac{n_2 \cdot (p^*-p)^2}{p^* \cdot (1-p^*)} \leq 16/3 \cdot (n_1+n_2) \cdot (p^*-p)^2 \leq 1/2, \tag{F.17}$$

where the second inequality follows from the fact that $p, p^* \in [1/4, 3/4]$ and the last inequality follows from Equation (F.16). By Equations (F.14) and (F.17), we have

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\big[\pi_1(b_1 \,|\, x_0)\big] \geq 1 - \sqrt{\mathrm{KL}(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1} \,\|\, \mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2})/2} \geq 1/2. \tag{F.18}$$

Combining Equations (F.16) and (F.18), for the output $\pi = \{\pi_h\}_{h=1}^H = \mathtt{Algo}(\mathcal{D})$ of any offline RL algorithm, we have

$$\begin{aligned}
&\max_{\ell \in \{1,2\}} \sqrt{n_\ell} \cdot \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Big[\mathrm{SubOpt}\big(\mathcal{M}_\ell, \mathtt{Algo}(\mathcal{D}); x_0\big)\Big] \\
&\quad \geq \frac{\sqrt{n_1 n_2}}{\sqrt{n_1} + \sqrt{n_2}} \cdot (p^* - p) \cdot (H-1) \cdot \Big(\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\big[\pi_1(b_1 \,|\, x_0)\big]\Big) \\
&\quad \geq \frac{\sqrt{n_1 n_2}}{\sqrt{n_1} + \sqrt{n_2}} \cdot \frac{1}{4} \cdot \sqrt{\frac{3}{2 \cdot (n_1+n_2)}} \cdot \frac{H-1}{2} \\
&\quad = \frac{\sqrt{n_1/n_2}}{(\sqrt{n_1/n_2} + 1) \cdot \sqrt{1 + n_1/n_2}} \cdot \frac{\sqrt{3}}{8\sqrt{2}} \cdot (H-1) \geq C' \cdot (H-1)
\end{aligned} \tag{F.19}$$

for an absolute constant

$$C' = \frac{\sqrt{3}}{8\sqrt{2}} \cdot \frac{1}{(\sqrt{\bar{c}} + 1) \cdot \sqrt{\bar{c} \cdot (\bar{c}+1)}} > 0. \tag{F.20}$$

Here the first inequality follows from Lemma C.3, while the last inequality follows from the fact that $1/\bar{c} \leq n_1/n_2 \leq \bar{c}$.

By the definition of $\mathcal{M}_1$ and $\mathcal{M}_2$ in Equation (C.15), the optimal policy $\pi^{*,1}$ for $\mathcal{M}_1$ always chooses the action b_1, while the optimal policy $\pi^{*,1}$ for $\mathcal{M}_2$ always chooses the action b_2. Recall that $n_j = \sum_{\tau=1}^K \mathbb{1}\{a_1^\tau = b_j\}$ for all $j \in [A]$ and $m_j = \sum_{\tau=1}^K \mathbb{1}\{x_2^\tau = x_j\}$ for all $j \in \{1, 2\}$. Also, recall that $j^* = \mathrm{argmax}_{j \in [A]}\, p_j = 1$ for $\mathcal{M}_1$, while $j^* = \mathrm{argmax}_{j \in [A]}\, p_j = 2$ for $\mathcal{M}_2$. Therefore, $n_{j^*} = \ell$ on $\mathcal{M}_\ell$ for all $\ell \in \{1, 2\}$. By Lemma F.1, for all $\ell \in \{1, 2\}$, we have

$$\sum_{h=1}^H \mathbb{E}_{\pi^{*,\ell}, \mathcal{M}_\ell}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x_0\Big] = \frac{1}{\sqrt{1 + n_\ell}} + (H-1) \cdot \Big(\frac{p^*}{\sqrt{1+m_1}} + \frac{1-p^*}{\sqrt{1+m_2}}\Big),$$

where $\mathbb{E}_{\pi^{*,\ell}, \mathcal{M}_\ell}$ is taken with repect to the trajectory induced by $\pi^{*,\ell}$ in $\mathcal{M}_\ell$ for all $\ell \in \{1, 2\}$. Given the actions $\{a_1^\tau\}_{\tau=1}^K$, m_1 and m_2 are the sums of K independent Bernoulli random variables. We define the event

$$\overline{\mathcal{E}} = \{m_1 \geq K/8,\ m_2 \geq K/8\}. \tag{F.21}$$

On $\overline{\mathcal{E}}$, it holds that

$$\begin{aligned}
&\sum_{h=1}^H \mathbb{E}_{\pi^{*,\ell}, \mathcal{M}_\ell}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x_0\Big] \\
&\quad \leq \frac{1}{\sqrt{n_\ell}} + (H-1) \cdot \Big(\frac{3/4}{\sqrt{m_1}} + \frac{3/4}{\sqrt{m_2}}\Big) \leq 6(H-1)/\sqrt{n_\ell}.
\end{aligned} \tag{F.22}$$

Hence, we have

$$\begin{aligned}
&\max_{\ell\in\{1,2\}} \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Biggl[\frac{\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)}{\sum_{h=1}^H \mathbb{E}_{\pi^*,\ell,\mathcal{M}_\ell}\Bigl[\bigl(\phi(s_h,a_h)^\top\Lambda_h^{-1}\phi(s_h,a_h)\bigr)^{1/2}\,\Big|\, s_1 = x_0\Bigr]}\Biggr] \\
&\quad\geq \max_{\ell\in\{1,2\}} \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Biggl[\frac{\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)}{\sum_{h=1}^H \mathbb{E}_{\pi^*,\ell,\mathcal{M}_\ell}\Bigl[\bigl(\phi(s_h,a_h)^\top\Lambda_h^{-1}\phi(s_h,a_h)\bigr)^{1/2}\,\Big|\, s_1 = x_0\Bigr]}\cdot \mathbb{1}_{\overline{\mathcal{E}}}\Biggr] \\
&\quad\geq \max_{\ell\in\{1,2\}}\biggl\{\frac{\sqrt{n_\ell}}{6(H-1)}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\cdot\mathbb{1}_{\overline{\mathcal{E}}}\Bigr]\biggr\}, \qquad \text{(F.23)}
\end{aligned}$$

where the last inequality follows from Equation (F.22) and the definition of $\overline{\mathcal{E}}$ in Equation (F.21). We use $\overline{\mathcal{E}}^c$ to denote the complement of $\overline{\mathcal{E}}$. By Equation (C.8), we have $r_h \in [0,1]$ for all $h \in [H]$ and $r_1(x_0, a) = 0$ for all $a \in \mathcal{A}$. Hence, the suboptimality of any policy is upper bounded by $H-1$. For all $\ell \in \{1,2\}$, we have

$$\begin{aligned}
&\sqrt{n_\ell}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\cdot\mathbb{1}_{\overline{\mathcal{E}}}\Bigr] \\
&\quad= \sqrt{n_\ell}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\Bigr] - \sqrt{n_\ell}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\cdot\mathbb{1}_{\overline{\mathcal{E}}^c}\Bigr] \\
&\quad\geq \sqrt{n_\ell}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\Bigr] - (H-1)\cdot\sqrt{n_\ell}\cdot\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_\ell}(\overline{\mathcal{E}}^c). \qquad \text{(F.24)}
\end{aligned}$$

We invoke the same argument as in Equations (F.10) and (F.11). For any $\delta > 0$ and any $\ell \in \{1,2\}$, since we have $p, p^* \geq 1/4$, when K is sufficiently large so that $K \geq 32\cdot\log(4/\delta)$, we have $\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_\ell}(\overline{\mathcal{E}}) \geq 1-\delta$. Hence, setting $\delta = 1/K^2$, when K is sufficiently large so that $K \geq 32\cdot\log(4K^2) = 64\cdot\log(2K)$, we have $\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_\ell}(\overline{\mathcal{E}}) \geq 1 - 1/K^2$ for all $\ell \in \{1,2\}$. By Equation (F.24), for all $\ell \in \{1,2\}$, it holds that

$$\begin{aligned}
&\sqrt{n_\ell}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\cdot\mathbb{1}_{\overline{\mathcal{E}}}\Bigr] \\
&\quad\geq \sqrt{n_\ell}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\Bigr] - (H-1)\cdot\sqrt{n_\ell}/K^2. \qquad \text{(F.25)}
\end{aligned}$$

By Equations (F.19), (F.23), and (F.25), we have

$$\begin{aligned}
&\max_{\ell\in\{1,2\}} \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Biggl[\frac{\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)}{\sum_{h=1}^H \mathbb{E}_{\pi^*,\ell,\mathcal{M}_\ell}\bigl[\bigl(\phi(s_h,a_h)^\top\Lambda_h^{-1}\phi(s_h,a_h)\bigr)^{1/2}\,\big|\, s_1 = x_0\bigr]}\Biggr] \\
&\quad\geq \max_{\ell\in\{1,2\}}\biggl\{\frac{\sqrt{n_\ell}}{6(H-1)}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\cdot\mathbb{1}_{\overline{\mathcal{E}}}\Bigr]\biggr\} \\
&\quad\geq \max_{\ell\in\{1,2\}}\biggl\{\frac{\sqrt{n_\ell}}{6(H-1)}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\Bigr] - \frac{n_\ell}{6K^2}\biggr\} \\
&\quad\geq \frac{1}{6(H-1)}\sup_{\ell\in\{1,2\}}\Bigl\{\sqrt{n_\ell}\cdot\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_\ell}\Bigl[\text{SubOpt}\bigl(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\bigr)\Bigr]\Bigr\} - \frac{1}{6K} \\
&\quad\geq C'/6 - 1/(6K) \geq C'/12 \qquad \text{(F.26)}
\end{aligned}$$

when K is sufficiently large so that $K \geq 64\cdot\log(2K)$ and $K \geq 2/C'$ for C' defined in Equation (F.20). Here the first inequality follows from Equation (F.23), the second inequality follows from Equation (F.25), the third inequality follows from the fact that $n_\ell \leq K$ for all $\ell \in \{1,2\}$, and the fourth inequality follows from Equation (F.19).

Since $\mathfrak{M}$ defined in Equation (C.6) is a subclass of linear MDPs, Equation (F.26) implies the lower bound in Equation (4.13) with $c = C'/12$ for K sufficiently large so that $K \geq 64\cdot\log(2K)$ and $K \geq 2/C'$. Therefore, we conclude the proof of Theorem 4.6. □

F.4. Locally Refined Upper Bounds

Since $\mathfrak{M}$ is a class of linear MDPs, a direct application of Theorem 4.4 yields an upper bound on the suboptimality of $\texttt{Pess}(\mathcal{D})$ constructed by Algorithm 2, which is minimax optimal up to β and absolute constant. In the sequel, focusing on $\mathfrak{M}$, we prove that, with a different choice of the ξ-uncertainty quantifier designed specifically for $\mathcal{M}$, PEVI achieves a more refined local minimax optimality.

Specifically, for any $\mathcal{M} = M(p_1, p_2, p_3) \in \mathfrak{M}$, recall that both x_1 and x_2 are absorbing states. By the construction of the reward function of $\mathcal{M}$, for any value function V and any $h \geq 2$, the Bellman operator $\mathbb{B}_h$ is defined by

$$(\mathbb{B}_h V)(x_1, a) = r_h(x_1, a) + V(x_1) = V(x_1) + 1, \qquad (\mathbb{B}_h V)(x_2, a) = r_h(x_2, a) + V(x_2) = V(x_2),$$

for all $a \in \mathcal{A}$. Besides, recall that the initial state is fixed to x_0. For any $j \in [A]$, we have

$$(\mathbb{B}_1 V)(x_0, b_j) = p_j \cdot V(x_1) + (1 - p_j) \cdot V(x_2), \tag{F.27}$$

where we let $p_j = p_3$ for all $j \geq 3$. Then, based on any dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$ that is compliant with $\mathcal{M}$, we construct an estimated Bellman operator $\widehat{B}_h$ and value function $\widehat{V}_h$ as follows. Starting from $\widehat{V}_{H+1}$ being a zero function, for any $h \geq 2$, we define $\widehat{V}_h$ by letting $\widehat{V}_h(x_1) = H - h + 1$ and $\widehat{V}_h(x_2) = 0$. For any $a \in \mathcal{A}$, we define $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$ by

$$\begin{aligned} (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x_1, a) &= (\mathbb{B}_h \widehat{V}_{h+1})(x_1, a) = \widehat{V}_{h+1}(x_1) + 1 = H - h + 1, \\ (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x_2, a) &= (\mathbb{B}_h \widehat{V}_{h+1})(x_2, a) = \widehat{V}_{h+1}(x_2) = 0. \end{aligned} \tag{F.28}$$

Furthermore, for $h = 1$, we define the empirical Bellman update by replacing Equation (F.27) with its empirical estimator. Specifically, since $r_1(x_0, a) = 0$ for all $a \in \mathcal{A}$, by Equation (F.28), for all $j \in [A]$ with $n_j > 0$, we define

$$\begin{aligned} (\widehat{\mathbb{B}}_1 \widehat{V}_2)(x_0, b_j) &= \frac{1}{n_j} \sum_{\tau=1}^{K} \mathbb{1}\{a_1^\tau = b_j\} \cdot \widehat{V}_2(x_2^\tau) \\ &= \frac{H-1}{n_j} \sum_{\tau=1}^{K} \mathbb{1}\{(a_1^\tau, x_2^\tau) = (b_j, x_1)\} = \frac{H-1}{n_j} \sum_{\tau=1}^{K} \mathbb{1}\{a_1^\tau = b_j\} \cdot r_2^\tau, \end{aligned} \tag{F.29}$$

while for $j \in [A]$ with $n_j = 0$, we simply set $(\widehat{\mathbb{B}}_1 \widehat{V}_2)(x_0, b_j) = 0$. Furthermore, for any $\xi > 0$, we define

$$\Gamma_1^\xi(x_0, b_j) = (H-1) \cdot \sqrt{\log(2A/\xi)/(1 + n_j)}, \quad \forall j \in [A], \qquad \Gamma_h^\xi(\cdot, \cdot) \equiv 0, \quad \forall h \geq 2. \tag{F.30}$$

Thus, employing the empirical Bellman update $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$ given in Equations (F.28) and (F.29), and function Γ_h^ξ defined in Equation (F.30), we obtain an instantiation of PEVI that is specified in Algorithm 1. We let $\texttt{Pess}^*(\mathcal{D})$ denote the output policy, whose suboptimality is established in the following proposition.

Proposition F.2 (Local Optimality of PEVI)**.** For any $\mathcal{M} \in \mathfrak{M}$ and any dataset $\mathcal{D}$ that is compliant with $\mathcal{M}$, we assume that $n_{j^*} = \sum_{\tau=1}^{K} \mathbb{1}\{a_1^\tau = b_j^*\} \geq 1$ for the optimal action b_{j^*}, where $j^* = \mathrm{argmax}_{j \in \{1,2\}}\{p_j\}$. Then the following statements hold: (i) $\{\Gamma_h^\xi\}_{h=1}^H$ defined in Equation (F.30) are ξ-uncertainty quantifiers satisfying Equation (4.1), and hence (ii) we have

$$\text{SubOpt}\big(\texttt{Pess}^*(\mathcal{D}); x_0\big) \leq c \cdot H\sqrt{\log(A/\xi)} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x_0\Big]$$

with probability at least least $1 - \xi$ with respect to $\mathbb{P}_\mathcal{D}$, where $c > 0$ is an absolute constant. Here $\mathbb{E}_{\pi^*}$ is taken with respect to the trajectory induced by the optimal policy π^* under the MDP $\mathcal{M}$, and Λ_h is defined as in Equation (4.6) with $\lambda = 1$.

We remark that by choosing ξ-uncertainty quantifiers designed specifically for linear MDPs in $\mathfrak{M}$, Proposition F.2 establishes a tighter upper bound compared to that in Theorem 4.4. Specifically, Equation (F.4) shows that directly applying Theorem 4.4 yields an $\widetilde{\mathcal{O}}(H^2 A/\sqrt{n_{j^*}})$ suboptimality upper bound, where $\widetilde{\mathcal{O}}(\cdot)$ omits logarithmic terms and absolute constants. In contrast, as shown in Equation (F.33), $\texttt{Pess}^*(\mathcal{D})$ achieves an improved $\widetilde{\mathcal{O}}(H/\sqrt{n_{j^*}})$ suboptimality upper bound. Thus, neglecting logarithmic terms and absolute constants, although both being minimax optimal algorithms, $\texttt{Pess}^*(\mathcal{D})$ is superior over $\texttt{Pess}(\mathcal{D})$ given in Algorithm 2 by a factor of HA, and $\texttt{Pess}^*(\mathcal{D})$ is minimax optimal up to a factor of H.

Proof of Proposition F.2. The proof consists of two steps. In the first step, we prove that $\{\Gamma_h^\xi\}_{h=1}^H$ given in Equation (F.30) is are valid ξ-uncertainty quantifiers. In the second step, we apply Theorem 4.2 and establish the final upper bound.

Step I. In the following, we show that $\{\Gamma_h^\xi\}_{h=1}^H$ are valid ξ-uncertainty quantifiers. That is, for the empirical Bellman operators given in Equations (F.28) and (F.29), we have

$$\mathbb{P}_\mathcal{D}\Big(\big|(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x, a) - (\mathbb{B}_h \widehat{V}_{h+1})(x, a)\big| \leq \Gamma_h^\xi(x, a), \ \ \forall (x, a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H]\Big) \geq 1 - \xi. \tag{F.31}$$

Combining Equations (F.28) and (F.30), we directly have

$$\big|(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x,a) - (\mathbb{B}_h \widehat{V}_{h+1})(x,a)\big| = 0 = \Gamma_h^\xi(x,a)$$

for all $(x,a) \in \mathcal{S} \times \mathcal{A}$ and $h \geq 2$.

Thus, it suffices to only focus on the case where $h = 1$. Recall that we define $n_j = \sum_{\tau=1}^K \mathbb{1}\{a_1^\tau = b_j\}$. For any $j \in [A]$, we consider the cases where $n_j = 0$ and $n_j \geq 1$ separately. For the former case, recall that we define $(\widehat{\mathbb{B}}_1 \widehat{V}_2)(x_0, b_j) = 0$. By Equation (F.27), we have $(\mathbb{B}_1 \widehat{V}_2)(x_0, b_j) = p_j \cdot (H-1)$. Besides, Equation (F.30) implies that

$$\Gamma_1^\xi(x_0, b_j) = (H-1) \cdot \sqrt{2\log(2A/\xi)} \geq (H-1) \geq (\mathbb{B}_1 \widehat{V}_2)(x_0, b_j).$$

Thus, we have $|(\widehat{\mathbb{B}}_1 \widehat{V}_2)(x_0, b_j) - (\mathbb{B}_1 \widehat{V}_2)(x_0, b_j)| \leq \Gamma_1^\xi(x_0, b_j)$.

Meanwhile, when $n_j \geq 1$, consider the σ-algebra $\mathcal{F}_\tau = \sigma(\{\kappa_j^i\}_{i=1}^\tau)$ for all $\tau \in [n_j]$, where we denote $\{\kappa_j^i\}_{i=1}^{n_j} = \{r_2^\tau \colon a_1^\tau = b_j, \tau \in [K]\}$ as introduced in Section C.3.1. Since $\mathcal{D}$ is compliant with $\mathcal{M}$, by Equation (C.13), $\{\kappa_j^i - p_j\}_{i=1}^{n_j}$ is a martingale difference sequence that is adapted to filtration $\{\mathcal{F}_i\}_{i=1}^{n_j}$. Applying Azuma-Hoeffding's inequality, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}}\bigg(\Big|\frac{1}{n_j}\sum_{i=1}^{n_j}(\kappa_j^i - p_j)\Big| \geq \sqrt{\log(2A/\xi)/(1+n_j)}\bigg) \\
&\qquad \leq 2\exp\big(-2n_j/(1+n_j) \cdot \log(2A/\xi)\big) \leq 2 \cdot (2A/\xi)^{-2n_j/(1+n_j)} \leq \xi/A,
\end{aligned} \tag{F.32}$$

where we utilize the fact that $n_j \geq 1$. Meanwhile, combining Equations (F.27), (F.28), (F.29), (F.30) and (F.32), for any fixed $j \in [A]$, we have

$$\big|(\widehat{\mathbb{B}}_1 \widehat{V}_2)(x_0, b_j) - (\mathbb{B}_1 \widehat{V}_2)(x_0, b_j)\big| = (H-1) \cdot \bigg|\frac{1}{n_j}\sum_{i=1}^{n_j}(\kappa_j^i - p_j)\bigg| \leq \Gamma_h^\xi(x, b_j)$$

with probability at least $1 - \xi/A$. Taking the union bound over all $j \in [A]$ yields the desired result in Equation (F.31). Thus, we have shown that $\{\Gamma_h^\xi\}_{h=1}^H$ given in Equation (F.30) are ξ-uncertainty quantifiers.

Step II. In the sequel, we apply Theorem 4.2 to $\texttt{Pess}^*(\mathcal{D})$ and establish the suboptimality upper bound in Proposition F.2. Specifically, Theorem 4.2 shows that, with probability at least $1 - \xi$ with respect to $\mathbb{P}_{\mathcal{D}}$, we have

$$\text{SubOpt}\big(\texttt{Pess}^*(\mathcal{D}); x_0\big) \leq 2\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\Gamma_h^\xi(s_h, a_h) \,\big|\, s_1 = x_0\big] = 2(H-1) \cdot \sqrt{\frac{\log(2A/\xi)}{1 + n_{j^*}}}, \tag{F.33}$$

where the last equality follows from Equation (F.30) and $j^* = \text{argmax}_{j \in \{1,2\}}\{p_j\}$. Meanwhile, by Equation (F.7) with $\lambda = 1$, it holds that

$$\sum_{h=1}^H \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x_0\Big] \geq \frac{1}{\sqrt{1 + n_{j^*}}}. \tag{F.34}$$

Also notice that $\log(2A/\xi) = \log 2 + \log(A/\xi) \leq 2\log(A/\xi)$ for $A \geq 2$. Finally, combining Equations (F.33) and (F.34), we have

$$\text{SubOpt}\big(\texttt{Pess}^*(\mathcal{D}); x_0\big) \leq c \cdot H\sqrt{\log(A/\xi)} \cdot \sum_{h=1}^H \mathbb{E}_{\pi^*}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x_0\Big]$$

with probability at least $1 - \xi$ with respect to $\mathbb{P}_{\mathcal{D}}$, where $c > 0$ is an absolute constant that can be chosen as $c = 4$. Therefore, we complete the proof of Proposition F.2. □

G. Supporting Lemmas

The following lemma characterizes the deviation of the sample mean of random matrices, which can be found in various literature, for example, see Theorem 1.6.2 of (Tropp, 2015).

Lemma G.1 (Matrix Bernstein Inequality). Assume that $\{A_k\}_{k=1}^n$ are n independent and centered random matrices in $\mathbb{R}^{d_1 \times d_2}$, that is, $\mathbb{E}[A_k] = \mathbf{0}$ for all $k \in [n]$. Besides, we assume that these random matrices are uniformly bounded and assume that each one is uniformly bounded in the sense that $\|A_k\|_{\text{op}} \leq L$ for all $k \in [n]$. Let $Z = \sum_{k=1}^n A_k$, and define

$v(Z)$ as

$$v(Z) = \max\Big\{\big\|\mathbb{E}[ZZ^\top]\big\|_{\rm op}, \big\|\mathbb{E}[Z^\top Z]\big\|_{\rm op}\Big\} = \max\Big\{\Big\|\sum_{k=1}^{n} \mathbb{E}[A_k A_k^\top]\Big\|_{\rm op}, \Big\|\sum_{k=1}^{n} \mathbb{E}[A_k^\top A_k]\Big\|_{\rm op}\Big\}.$$

Then, for any $t \geq 0$, we have

$$\mathbb{P}\big(\|Z\|_{\rm op} \geq t\big) \leq (d_1 + d_2) \cdot \exp\Big(-\frac{t^2/2}{v(Z) + L/3 \cdot t}\Big).$$

Proof. See, e.g., Theorem 1.6.2 of (Tropp, 2015) for a detailed proof. □

In addition, the following lemma, obtained from (Abbasi-Yadkori et al., 2011), establishes the concentration of self-normalized processes.

Lemma G.2 (Concentration of Self-Normalized Processes (Abbasi-Yadkori et al., 2011))**.** Let $\{\epsilon_t\}_{t=1}^\infty$ be a real-valued stochastic process that is adaptive to a filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. That is, ϵ_t is $\mathcal{F}_t$-measurable for all $t \geq 1$. Moreover, we assume that, for any $t \geq 1$, conditioning on $\mathcal{F}_{t-1}$, ϵ_t is a zero-mean and σ-subGaussian random variable such that

$$\mathbb{E}[\epsilon_t \,|\, \mathcal{F}_{t-1}] = 0 \qquad \text{and} \qquad \mathbb{E}[\exp(\lambda \epsilon_t) \,|\, \mathcal{F}_{t-1}] \leq \exp(\lambda^2 \sigma^2/2), \quad \forall \lambda \in \mathbb{R}. \tag{G.1}$$

Besides, let $\{\phi_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that ϕ_t is $\mathcal{F}_{t-1}$-measurable for all $t \geq 1$. Let $M_0 \in \mathbb{R}^{d\times d}$ be a deterministic and positive-definite matrix, and we define $M_t = M_0 + \sum_{s=1}^t \phi_s \phi_s^\top$ for all $t \geq 1$. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have for all $t \geq 1$ that

$$\Big\|\sum_{s=1}^{t} \phi_s \cdot \epsilon_s\Big\|_{M_t^{-1}}^2 \leq 2\sigma^2 \cdot \log\bigg(\frac{\det(M_t)^{1/2} \det(M_0)^{-1/2}}{\delta}\bigg).$$

Proof. See Theorem 1 of (Abbasi-Yadkori et al., 2011) for a detailed proof. □