A Discussions

In this section, we are going to discuss: some other types of robust MDPs appearing in existing works, including d-rectangular robust linear MDPs [29] and RMDPs with $\mathcal{S}$-rectangular robust sets [61], see Section A.1 and A.2 respectively.

A.1 d-rectangular robust linear MDPs

Recently [29] proposed the d-rectangular robust linear MDP to study offline robust RL with linear structures. We use the following example to show how a d-rectangular robust linear MDP is represented by our general framework of RMDP.

Example A.1 (d-rectangular robust linear MDP [29])**.** *A d-rectangular robust linear MDP is equipped with d-rectangular robust sets. Linear MDP is an MDP that enjoys a d-dimensional linear decomposition of its reward function and transition kernel [15]. We define the model space $\mathcal{P}_{\mathrm{M}}$ as*

$$\mathcal{P}_{\mathrm{M}} = \Big\{ P(s'|s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}(s') : \mu_i(\cdot) \in \Delta(\mathcal{S}), \forall i \in [d] \Big\},$$

where $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is a known feature mapping satisfying that

$$\sum_{i=1}^{d} \phi_i(s,a) = 1, \quad \phi_i(s,a) \geq 0, \quad \forall i \in [d].$$

We then assume that $P_h^\star(s'|s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}^\star(s') \in \mathcal{P}_{\mathrm{M}}$, and $R_h(s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h$ for some $\boldsymbol{\theta}_h \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d}$. We define the mapping $\boldsymbol{\Phi}$ as

$$\boldsymbol{\Phi}(P) = \left\{ \sum_{i=1}^{d} \phi_i(s,a) \widetilde{\mu}_i(s') : \widetilde{\mu}_i(\cdot) \in \Delta(\mathcal{S}), D(\widetilde{\mu}(\cdot) \| \mu_i(\cdot)) \leq \rho, \forall i \in [d] \right\}.$$

This is called a d-rectangular robust set and is first considered by [29]. As is argued in [29], d-rectangular robust set is not so conservative as $\mathcal{S} \times \mathcal{A}$-rectangular robust set in certain cases, which is more natural for linear MDPs due to the special linear structure.

While not satisfying Assumption 2.2 ($\mathcal{S} \times \mathcal{A}$-rectangular robust sets), it can be proved that RMDP in Example A.1 also satisfies the robust Bellman equation in Proposition 2.3 (similar to the proof in Appendix B for $\mathcal{S} \times \mathcal{A}$-rectangular robust MDPs). Our algorithm `P`^{2}`MPO` (Algorithm 1) can also be applied to offline solve robust RL with RMDP in Example A.1, under certain partial coverage assumption (Assumption A.2).

Model estimation. In the following, we give a specific implementation of the model estimation step for RMDPs in Example A.1, and we provide theoretical guarantees for this specification of our algorithm `P`^{2}`MPO`. Suppose we are given a function class $\mathcal{V} \subseteq \{v : \mathcal{S} \mapsto [0,1]\}$ which depends on the choice of distance $D(\cdot\|\cdot)$ of the robust set. Then, we define that

$$\widehat{\mathcal{P}}_h = \left\{ P \in \mathcal{P}_{\mathrm{M}} : \sup_{v \in \mathcal{V}} \frac{1}{n} \sum_{\tau=1}^{n} \left| \int_{\mathcal{S}} P(\mathrm{d}s'|s_h^\tau, a_h^\tau) v(s') - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \widehat{\boldsymbol{\theta}}_v \right|^2 \leq \xi \right\}, \tag{A.1}$$

where $\xi > 0$ is a tuning parameter that controls the size of the confidence region, and the vector $\widehat{\boldsymbol{\theta}}_v$ depends on the specific function $v \in \mathcal{V}$, given by

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_v &= \mathop{\arg\min}_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \sum_{\tau=1}^{n} \big(\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{\theta} - v(s_{h+1}^\tau) \big)^2 + \frac{\alpha}{n} \cdot \|\boldsymbol{\theta}\|_2^2 \\ &= \boldsymbol{\Lambda}_{h,\alpha}^{-1} \left(\frac{1}{n} \sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) v(s_{h+1}^\tau) \right), \end{aligned} \tag{A.2}$$

for some tuning parameter $\alpha > 0$, where $\boldsymbol{\Lambda}_{h,\alpha}$ is the regularized covariance matrix, defined as

$$\boldsymbol{\Lambda}_{h,\alpha} = \frac{1}{n} \sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \frac{\alpha}{n} \cdot \boldsymbol{I}_d.$$

Similar constructions for standard linear MDPs are also considered by [51, 34, 54]. We will specify the choice of the function class $\mathcal{V}$ in the theoretical guarantees of this implementation.

Suboptimality analysis. In the following, we provide suboptimality bounds for the above implementation of `P`^{2}`MPO` for d-rectangular robust linear MDP. Regarding the offline data, we impose the following robust partial coverage assumption.

Assumption A.2 (Robust partial coverage covariance matrix)**.** *We assume that for some constant* $c^\dagger > 0$,

$$\boldsymbol{\Lambda}_{h,\alpha} \succeq \frac{\alpha}{n} \cdot \boldsymbol{I}_d + c^\dagger \cdot \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{P,h}}[(\phi_i(s_h,a_h)\mathbf{1}_i)(\phi_i(s_h,a_h)\mathbf{1}_i)^\top] \tag{A.3}$$

for any $i \in [d]$*,* $h \in [H]$*, and* $P_h \in \boldsymbol{\Phi}(P_h^\star)$*.*

Theorem A.3 (Suboptimality of `P`^{2}`MPO`: d-rectangular robust linear MDP)**.** *Suppose that the RMDP is* d*-rectangular robust linear MDP in Example A.1 with* $D(\cdot\|\cdot)$ *being KL-divergence or TV-distance and that Assumption A.2 holds, choosing the tuning parameter* $\alpha = 1$ *in (A.2).*

♠ *when* $D(\cdot\|\cdot)$ *is KL-divergence and Assumption F.1 holds with parameter* $\underline{\lambda}$*, then by setting*

$$\mathcal{V} = \left\{ v(s) = \exp\left(-\max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \boldsymbol{w}/\lambda \right) : \|\boldsymbol{w}\|_2 \le H\sqrt{d}, \lambda \in [\underline{\lambda}, H/\rho] \right\},$$

and

$$\xi = \frac{C_1 d^2 \big(\log(1 + C_2 nH/\delta) + \log(1 + C_3 ndH/(\rho\underline{\lambda}^2)) \big)}{n},$$

for some constants $C_1, C_2, C_3 > 0$*, it holds with probability at least* $1 - 2\delta$ *that,*

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \le \frac{d^2 H^2 \exp(H/\underline{\lambda})}{c^\dagger \rho} \cdot \sqrt{\frac{C_1' \big(\log(1 + C_2' nH/\delta) + \log(1 + C_3' ndH/(\rho\underline{\lambda}^2)) \big)}{n}}.$$

♠ *when* $D(\cdot\|\cdot)$ *is TV-distance, then by setting*

$$\mathcal{V} = \left\{ v(s) = \left(\lambda - \max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \boldsymbol{w} \right)_+ : \|\boldsymbol{w}\|_2 \le H\sqrt{d}, \lambda \in [0, H] \right\},$$

and

$$\xi = \frac{C_1 d^2 H^2 \log(C_2 ndH/\delta)}{n},$$

for some constants $C_1, C_2 > 0$*, it holds with probability at least* $1 - 2\delta$ *that,*

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \le \frac{d^2 H^2}{c^\dagger} \cdot \sqrt{\frac{C_1' \log(C_2' ndH/\delta)}{n}}.$$

Here $\underline{c}$ *is from Assumption A.2 and* $C_1', C_2', C_3' > 0$ *are universal constants.*

Proof of Theorem A.3. See Appendix F for a detailed proof. ☐

A.2 RMDPs with $\mathcal{S}$-rectangular robust sets

Besides $\mathcal{S} \times \mathcal{A}$-rectangular, there exists another type of generic rectangular assumption on robust sets called $\mathcal{S}$-rectangular [61, 67]. See the following assumption.

Assumption A.4 ($\mathcal{S}$-rectangular robust sets [61])**.** *An* $\mathcal{S}$*-rectangular robust MDP is equipped with* $\mathcal{S}$*-rectangular robust sets. The mapping* $\boldsymbol{\Phi}$ *is defined as, for* $\forall P \in \mathcal{P}_{\mathrm{M}}$*,*

$$\boldsymbol{\Phi}(P) = \bigotimes_{s\in\mathcal{S}} \mathcal{P}_\rho(s; P), \quad \mathcal{P}_\rho(s; P) = \left\{ \widetilde{P}(\cdot|\cdot) : \mathcal{A} \mapsto \Delta(\mathcal{S}) : \sum_{a\in\mathcal{A}} D(\widetilde{P}(\cdot|a)\|P(\cdot|s,a)) \le \rho|\mathcal{A}| \right\},$$

for some (pseudo-)distance $D(\cdot\|\cdot)$ *on* $\Delta(\mathcal{S})$ *and some real number* $\rho \in \mathbb{R}_+$*.*

RMDP with $\mathcal{S}$-rectangular robust sets (Assumption A.4) also satisfies Proposition 2.3 [61]. Unfortunately, our algorithm framework is unable to deal with this kind of rectangular robust sets in the context of partial coverage data due to some technical problems in applying the partial coverage coefficient (Assumption 3.3) under this kind of robust sets. To our best knowledge, how to design provably efficient algorithms for $\mathcal{S}$-rectangular RMDP with partial coverage data is still unknown. It is an exciting future work to fill this gap for robust offline reinforcement learning.

B Proof of Robust Bellman Equation

Proof of Proposition 2.3 for $\mathcal{S}\times\mathcal{A}$-rectangular robust MDP. Instead of directly proving the robust Bellman equation (2.5), we prove the following stronger results via induction from step $h=H$ to 1: *there exists a set of transition kernels $P^{\pi,\dagger}=\{P_h^{\pi,\dagger}\}_{h=1}^H$ with $P_h^{\pi,\dagger}\in\mathbf{\Phi}(P_h)$ such that*

1. *Robust Bellman equation holds, i.e.,*

$$\begin{aligned} V_{h,P,\mathbf{\Phi}}^{\pi}(s) &= \mathbb{E}_{a\sim\pi_h(\cdot|s)}[Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a)], \\ Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a) &= R_h(s,a) + \inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s,a)}[V_{h+1,P,\mathbf{\Phi}}^{\pi}(s')]. \end{aligned}$$

2. *The following expressions for robust value functions hold,*

$$\begin{aligned} V_{h,P,\mathbf{\Phi}}^{\pi}(s) &= V_h^{\pi}(s;\{P_i^{\pi,\dagger}\}_{i=h}^H), \\ Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a) &= Q_h^{\pi}(s,a;\{P_i^{\pi,\dagger}\}_{i=h}^H). \end{aligned}$$

Firstly, for step $h=H$, the conclusion 1. and 2. hold directly because no transitions are involved. Now supposing that the conclusion 1. and 2. hold for some step $h+1$, which means that there exist transition kernels $\{P_i^{\pi,\dagger}\}_{i=h+1}^H$ such that the following condition hold for any $s\in\mathcal{S}$,

$$V_{h+1,P,\mathbf{\Phi}}^{\pi}(s) = V_{h+1}^{\pi}(s;\{P_i^{\pi,\dagger}\}_{i=h+1}^H). \tag{B.1}$$

By the definition of robust value function $Q_{h,P,\mathbf{\Phi}}^{\pi}$ in (2.2), we can derive that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned} Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a) &= \inf_{\widetilde{P}_i\in\mathbf{\Phi}(P_i), h\le i\le H} \mathbb{E}_{\{\widetilde{P}_i\}_{i=h}^H,\pi}\left[\sum_{i=h}^{H} R_i(s_i,a_i)\middle| s_h=s, a_h=a\right] \\ &= R_h(s,a) + \inf_{\widetilde{P}_i\in\mathbf{\Phi}(P_i), h\le i\le H}\int_{\mathcal{S}} \widetilde{P}_h(\mathrm{d}s'|s,a)\mathbb{E}_{\{\widetilde{P}_i\}_{i=h+1}^H,\pi}\left[\sum_{i=h+1}^{H} R_i(s_i,a_i)\middle| s_{h+1}=s'\right] \\ &\le R_h(s,a) + \inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}\int_{\mathcal{S}} \widetilde{P}_h(\mathrm{d}s'|s,a)\mathbb{E}_{\{P_i^{\pi,\dagger}\}_{i=h+1}^H,\pi}\left[\sum_{i=h+1}^{H} R_i(s_i,a_i)\middle| s_{h+1}=s'\right]. \end{aligned} \tag{B.2}$$

On the one hand, for $\mathcal{S}\times\mathcal{A}$-rectangular robust MDP, the robust set $\mathbf{\Phi}(P_h)$ is decoupled for different (s,a) pairs, i.e.,

$$\mathbf{\Phi}(P_h) = \bigotimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{\rho}(s,a;P_h),$$

and therefore we can find a *single* transition kernel $P_h^{\pi,\dagger}$ such that for *any* $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$P_h^{\pi,\dagger}(\cdot|s,a) = \mathop{\mathrm{arginf}}_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}\int_{\mathcal{S}} \widetilde{P}(\mathrm{d}s'|s,a)\mathbb{E}_{\{P_i^{\pi,\dagger}\}_{i=h+1}^H,\pi}\left[\sum_{i=h+1}^{H} R_i(s_i,a_i)\middle| s_{h+1}=s'\right]. \tag{B.3}$$

On the other hand, using condition (B.1) and the definition of (robust) value function $V_{h,P,\mathbf{\Phi}}^{\pi}$ and V_h^{π} in (2.1) and (2.3), we can also deduce that,

$$\begin{aligned} Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a) &\le R_h(s,a) + \inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}\int_{\mathcal{S}} \widetilde{P}_h(\mathrm{d}s'|s,a)V_{h+1}^{\pi}(s';\{P_i^{\pi,\dagger}\}_{i=h+1}^H) \\ &= R_h(s,a) + \inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}\int_{\mathcal{S}} \widetilde{P}_h(\mathrm{d}s'|s,a)V_{h+1,P,\mathbf{\Phi}}^{\pi}(s') \qquad \text{(B.4)} \\ &= R_h(s,a) + \inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}\int_{\mathcal{S}} \widetilde{P}_h(\mathrm{d}s'|s,a)\inf_{\widetilde{P}_i\in\mathbf{\Phi}(P_i), h+1\le i\le H} V_{h+1}^{\pi}(s';\{\widetilde{P}_i\}_{i=h+1}^H) \\ &\le R_h(s,a) + \inf_{\widetilde{P}_i\in\mathbf{\Phi}(P_i), h\le i\le H}\int_{\mathcal{S}} \widetilde{P}_h(\mathrm{d}s'|s,a)V_{h+1}^{\pi}(s';\{\widetilde{P}_i\}_{i=h+1}^H), \qquad \text{(B.5)} \end{aligned}$$

where the first inequality follows from inequality (B.2) and the definition of V_{h+1}^{π} in (2.3), the first equality follows from condition (B.1), and the second equality follows from the definition of $V_{h+1,P,\mathbf{\Phi}}^{\pi}$ in (2.1). Note that the right hand side of (B.5) equals to $Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a)$. Therefore, all the inequalities are actually equalities. On the one hand, from (B.4), we can know that,

$$Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a) = R_h(s,a) + \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \int_{\mathcal{S}} \widetilde{P}_h(\mathrm{d}s'|s,a) V_{h+1,P,\mathbf{\Phi}}^{\pi}(s').$$

This proves the $Q_{h,P,\mathbf{\Phi}}^{\pi}$ part of the conclusion 1. for step h. On the other hand, by combining (B.3) and (B.2), one can further obtain that,

$$Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a) = \mathbb{E}_{\{P_i^{\pi,\dagger}\}_{i=h}^{H},\pi}\left[\sum_{i=h}^{H} R_i(s_i,a_i)\middle| s_h = s, a_h = a\right] = Q_h^{\pi}(s,a;\{P_i^{\pi,\dagger}\}_{i=h}^{H}). \tag{B.6}$$

This proves the existence of $\{P_i^{\pi,\dagger}\}_{i=h}^{H}$ in the conclusion 2. for step h and $Q_{h,P,\mathbf{\Phi}}^{\pi}$. The remaining of the proof is to prove the $V_{h,P,\mathbf{\Phi}}^{\pi}$ part of the conclusion 1. and 2. for step h using $\{P_i^{\pi,\dagger}\}_{i=h}^{H}$ found in the previous proof. Specifically, by the definition of $V_{h,P,\mathbf{\Phi}}^{\pi}$ in (2.1), we have that,

$$\begin{aligned} V_{h,P,\mathbf{\Phi}}^{\pi}(s) &= \inf_{\widetilde{P}_i \in \mathbf{\Phi}(P_i), h\le i\le H} \mathbb{E}_{\{\widetilde{P}_i\}_{i=h}^{H},\pi}\left[\sum_{i=h}^{H} R_i(s_i,a_i)\middle| s_h = s\right] \\ &= \inf_{\widetilde{P}_i \in \mathbf{\Phi}(P_i), h\le i\le H} \sum_{a\in\mathcal{A}} \pi_h(a|s) \mathbb{E}_{\{\widetilde{P}_i\}_{i=h}^{H},\pi}\left[\sum_{i=h}^{H} R_i(s_i,a_i)\middle| s_h = s, a_h = a\right] \\ &\le \sum_{a\in\mathcal{A}} \pi_h(a|s) \mathbb{E}_{\{P_i^{\pi,\dagger}\}_{i=h}^{H},\pi}\left[\sum_{i=h}^{H} R_i(s_i,a_i)\middle| s_h = s, a_h = a\right]. \end{aligned} \tag{B.7}$$

Now applying (B.6) to (B.7), we can further obtain that

$$\begin{aligned} V_{h,P,\mathbf{\Phi}}^{\pi}(s) &\le \sum_{a\in\mathcal{A}} \pi_h(a|s) Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a) \qquad \text{(B.8)} \\ &= \sum_{a\in\mathcal{A}} \pi_h(a|s) \inf_{\widetilde{P}_i \in \mathbf{\Phi}(P_i), h\le i\le H} \mathbb{E}_{\{\widetilde{P}_i\}_{i=h}^{H},\pi}\left[\sum_{i=h}^{H} R_i(s_i,a_i)\middle| s_h = s, a_h = a\right] \\ &\le \inf_{\widetilde{P}_i \in \mathbf{\Phi}(P_i), h\le i\le H} \sum_{a\in\mathcal{A}} \pi_h(a|s) \mathbb{E}_{\{\widetilde{P}_i\}_{i=h}^{H},\pi}\left[\sum_{i=h}^{H} R_i(s_i,a_i)\middle| s_h = s, a_h = a\right], \qquad \text{(B.9)} \end{aligned}$$

where the equality follows from the definition of $Q_{h,P,\mathbf{\Phi}}^{\pi}$ in (2.2). Now note that the right hand side of (B.9) equals to $V_{h,P,\mathbf{\Phi}}^{\pi}$. Therefore, all the inequalities are actually equalities. On the one hand, by (B.8), we know that,

$$V_{h,P,\mathbf{\Phi}}^{\pi}(s) = \sum_{a\in\mathcal{A}} \pi_h(a|s) Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a). \tag{B.10}$$

This proves the $V_{h,P,\mathbf{\Phi}}^{\pi}$ part of the conclusion 1. for step h. On the other hand, by combining (B.10) with (B.6), we can further deduce that,

$$V_{h,P,\mathbf{\Phi}}^{\pi}(s) = \mathbb{E}_{\{P_i^{\pi,\dagger}\}_{i=h}^{H},\pi}\left[\sum_{i=h}^{H} R_i(s_i,a_i)\middle| s_h = s\right].$$

This proves the $V_{h,P,\mathbf{\Phi}}^{\pi}$ part of the conclusion 2. for step h. Finally, by using an induction argument, we can finish the proof of the conclusion 1. and 2.

Now according to the conclusion 1., we have that

$$V_{h,P,\mathbf{\Phi}}^{\pi}(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}[R_h(s,a)] + \mathbb{E}_{a\sim\pi_h(\cdot|s)}\left[\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s,a)}[V_{h+1,P,\mathbf{\Phi}}^{\pi}(s')]\right]. \tag{B.11}$$

By the conclusion 2. and the definition of $P_h^{\pi,\dagger}$ in (B.3), we can obtain from (B.11) that

$$\begin{aligned} V^{\pi}_{h,P,\boldsymbol{\Phi}}(s) &= \mathbb{E}_{a\sim\pi_h(\cdot|s)}[R_h(s,a)] + \mathbb{E}_{a\sim\pi_h(\cdot|s),s'\sim P_h^{\pi,\dagger}(\cdot|s,a)}[V^{\pi}_{h+1,P,\boldsymbol{\Phi}}(s')] \\ &= \mathbb{E}_{a\sim\pi_h(\cdot|s)}[R_h(s,a)] + \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h)} \mathbb{E}_{a\sim\pi_h(\cdot|s),s'\sim \widetilde{P}_h(\cdot|s,a)}[V^{\pi}_{h+1,P,\boldsymbol{\Phi}}(s')]. \end{aligned}$$

This finishes the proof of Proposition 2.3 under Assumption 2.2. $\square$

C Proof of Main Theoretical Result (Theorem 3.4)

In this section, we prove Theorem 3.4. Let $\mathcal{E}^{\dagger}$ denote the event that both Condition 3.1 and 3.2 hold, which happens with probability at least $1-2\delta$. In the following, we always assume that $\mathcal{E}^{\dagger}$ holds.

Proof of Theorem 3.4. By the definition of $\mathrm{SubOpt}(\widehat{\pi};s)$ in (2.11), we have that

$$\begin{aligned} \mathrm{SubOpt}(\widehat{\pi};s_1) &= V^{\pi^\star}_{1,P^\star,\boldsymbol{\Phi}}(s_1) - V^{\widehat{\pi}}_{1,P^\star,\boldsymbol{\Phi}}(s_1) \\ &= V^{\pi^\star}_{1,P^\star,\boldsymbol{\Phi}}(s_1) - \inf_{P\in\widehat{\mathcal{P}}} V^{\pi^\star}_{1,P,\boldsymbol{\Phi}}(s_1) + \inf_{P\in\widehat{\mathcal{P}}} V^{\pi^\star}_{1,P,\boldsymbol{\Phi}}(s_1) - V^{\widehat{\pi}}_{1,P^\star,\boldsymbol{\Phi}}(s_1) \\ &\le V^{\pi^\star}_{1,P^\star,\boldsymbol{\Phi}}(s_1) - \inf_{P\in\widehat{\mathcal{P}}} V^{\pi^\star}_{1,P,\boldsymbol{\Phi}}(s_1) + \inf_{P\in\widehat{\mathcal{P}}} V^{\widehat{\pi}}_{1,P,\boldsymbol{\Phi}}(s_1) - V^{\widehat{\pi}}_{1,P^\star,\boldsymbol{\Phi}}(s_1) \qquad (C.1) \\ &\le V^{\pi^\star}_{1,P^\star,\boldsymbol{\Phi}}(s_1) - \inf_{P\in\widehat{\mathcal{P}}} V^{\pi^\star}_{1,P,\boldsymbol{\Phi}}(s_1) \qquad (C.2) \\ &= \sup_{P\in\widehat{\mathcal{P}}} \Big\{ V^{\pi^\star}_{1,P^\star,\boldsymbol{\Phi}}(s_1) - V^{\pi^\star}_{1,P,\boldsymbol{\Phi}}(s_1) \Big\}. \qquad (C.3) \end{aligned}$$

Here (C.1) follows from our choice of $\widehat{\pi}$ in (3.2), and (C.2) follows from Condition 3.1. In the sequel, we present the upper bound on the right hand side of (C.3). For notational simplicity, for any P in the confidence region $\widehat{\mathcal{P}}$ and any step $h\in[H]$, we denote that

$$\Delta_{h,P,\boldsymbol{\Phi}}(s_h,a_h) = Q^{\pi^\star}_{h,P^\star,\boldsymbol{\Phi}}(s_h,a_h) - Q^{\pi^\star}_{h,P,\boldsymbol{\Phi}}(s_h,a_h). \qquad (C.4)$$

Using the robust Bellman equation in Proposition 2.3, we can derive that

$$\begin{aligned} &\Delta_{h,P,\boldsymbol{\Phi}}(s_h,a_h) \\ &\quad= \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P^\star,\boldsymbol{\Phi}}(s')] - \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s') \\ &\quad= \underbrace{\inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P^\star,\boldsymbol{\Phi}}(s')] - \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')]}_{\text{Term (i)}} \\ &\qquad\quad+ \underbrace{\inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')] - \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')]}_{\text{Term (ii)}}. \end{aligned}$$

Term (i). For the term (i), considering denote that

$$P_h^{\pi^\star,\dagger} = \mathop{\mathrm{arginf}}_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s,a)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')], \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A}. \qquad (C.5)$$

This notation is consistent with the notation of $P_h^{\pi,\dagger}$ in (B.3) in the proof of Proposition 2.3 (robust Bellman equation). It is because Assumption 2.2 ($\mathcal{S}\times\mathcal{A}$-rectangular robust set) that we can choose a *single* transition kernel $P_h^{\pi^\star,\dagger}$ that satisfies (C.5) for each (s,a)-pair. Using the definition of $P_h^{\pi^\star,\dagger}$, we observe that the following two relationships hold for any state $(s_h,a_h)\in\mathcal{S}$,

$$\begin{aligned} \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P^\star,\boldsymbol{\Phi}}(s')] &\le \mathbb{E}_{s'\sim P_h^{\pi^\star,\dagger}(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P^\star,\boldsymbol{\Phi}}(s')], \\ \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')] &= \mathbb{E}_{s'\sim P_h^{\pi^\star,\dagger}(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')]. \end{aligned}$$

Using these two observations, we can upper bound the term (i) as

$$\begin{aligned}\text{Term (i)} &\le \mathbb{E}_{s'\sim P_h^{\pi^\star,\dagger}(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P^\star,\boldsymbol{\Phi}}(s')] - \mathbb{E}_{s'\sim P_h^{\pi^\star,\dagger}(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')] \\ &= \mathbb{E}_{s'\sim P_h^{\pi^\star,\dagger}(\cdot|s_h,a_h),a'\sim\pi^\star_{h+1}(\cdot|s')}[\Delta_{h+1,P,\boldsymbol{\Phi}}(s',a')], \end{aligned} \tag{C.6}$$

where in the equality we use the robust Bellman equation (Proposition 2.3).

Term (ii). For the term (ii), currently we simply denote this term by $\Delta^{(\mathrm{ii})}_{h,P,\boldsymbol{\Phi}}(s_h,a_h)$. Combining this with (C.6), we can derive that,

$$\begin{aligned}\Delta_{h,P,\boldsymbol{\Phi}}(s_h,a_h) &= \text{Term (i)} + \text{Term (ii)} \\ &\le \mathbb{E}_{s'\sim P_h^{\pi^\star,\dagger}(\cdot|s_h,a_h),a'\sim\pi^\star_{h+1}(\cdot|s')}[\Delta_{h+1,P,\boldsymbol{\Phi}}(s',a')] + \Delta^{(\mathrm{ii})}_{h,P,\boldsymbol{\Phi}}(s_h,a_h). \end{aligned} \tag{C.7}$$

By recursively applying (C.7) and then plugging in the definition of $\Delta^{(\mathrm{ii})}_{h,P,\boldsymbol{\Phi}}$, we can obtain that

$$\begin{aligned}\mathbb{E}_{a_1\sim\pi_1^\star(\cdot|s_1)}[\Delta_{1,P,\boldsymbol{\Phi}}(s_1,a_1)] &\le \sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{P^{\pi^\star,\dagger},h}}[\Delta^{(\mathrm{ii})}_{h,P,\boldsymbol{\Phi}}(s_h,a_h)] \\ &= \sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{P^{\pi^\star,\dagger},h}}\Bigg[\inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')] \\ &\qquad - \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')]\Bigg], \end{aligned} \tag{C.8}$$

where $d^{\pi^\star}_{P^{\pi^\star,\dagger},h}$ is the state action visitation distribution induced by the transition kernels $P^{\pi^\star,\dagger} = \{P_h^{\pi^\star,\dagger}\}_{h=1}^H$ and the policy $\pi^\star$. Now we bound the right hand side of (C.8) using Condition 3.2. By Cauchy-Schwartz inequality, we have that for each $h\in[H]$,

$$\begin{aligned}&\mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{P^{\pi^\star,\dagger},h}}\Bigg[\inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')] - \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')]\Bigg] \\ &\quad = \mathbb{E}_{(s_h,a_h)\sim d^{\pi^{\mathrm{b}}}_{P^\star,h}}\Bigg[\frac{d^{\pi^\star}_{P^{\pi^\star,\dagger},h}(s_h,a_h)}{d^{\pi^{\mathrm{b}}}_{P^\star,h}(s_h,a_h)}\cdot\bigg(\inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')] \\ &\qquad - \inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')]\bigg)\Bigg] \\ &\quad \le \sqrt{\mathbb{E}_{(s_h,a_h)\sim d^{\pi^{\mathrm{b}}}_{P^\star,h}}\left[\left(\frac{d^{\pi^\star}_{P^{\pi^\star,\dagger},h}(s_h,a_h)}{d^{\pi^{\mathrm{b}}}_{P^\star,h}(s_h,a_h)}\right)^2\right]}\cdot\sqrt{\mathrm{Err}_h^{\boldsymbol{\Phi}}(n)}, \end{aligned} \tag{C.9}$$

where the last inequality follows from Condition 3.2. Furthermore, by Assumption 3.3, we know that

$$\begin{aligned}\mathbb{E}_{(s_h,a_h)\sim d^{\pi^{\mathrm{b}}}_{P^\star,h}}\left[\left(\frac{d^{\pi^\star}_{P^{\pi^\star,\dagger},h}(s_h,a_h)}{d^{\pi^{\mathrm{b}}}_{P^\star,h}(s_h,a_h)}\right)^2\right] &\le \sup_{P=\{P_h\}_{h=1}^H,P_h\in\boldsymbol{\Phi}(P_h^\star)}\mathbb{E}_{(s_h,a_h)\sim d^{\pi^{\mathrm{b}}}_{P^\star,h}}\left[\left(\frac{d^{\pi^\star}_{P,h}(s_h,a_h)}{d^{\pi^{\mathrm{b}}}_{P^\star,h}(s_h,a_h)}\right)^2\right] \\ &\le C^\star_{P^\star,\boldsymbol{\Phi}}, \end{aligned}$$

where $C^\star_{P^\star,\boldsymbol{\Phi}}$ is defined in Assumption 3.3. Applying this to (C.8) and (C.9), we can derive that

$$\sup_{P\in\widehat{\mathcal{P}}}\left\{V^{\pi^\star}_{1,P^\star,\boldsymbol{\Phi}}(s_1) - V^{\pi^\star}_{1,P,\boldsymbol{\Phi}}(s_1)\right\} = \sup_{P\in\widehat{\mathcal{P}}}\{\mathbb{E}_{a_1\sim\pi^\star(\cdot|s_1)}[\Delta_{1,P,\boldsymbol{\Phi}}(s_1,a_1)]\} \le \sqrt{C^\star_{P^\star,\boldsymbol{\Phi}}}\cdot\sum_{h=1}^{H}\sqrt{\mathrm{Err}_h^{\boldsymbol{\Phi}}(n)}.$$

Finally, by inequality (C.3), we finish the proof of Theorem 3.4. □

D Proofs for General RMDPs with $\mathcal{S}\times\mathcal{A}$-rectangular Robust Sets

Proof of Corollary 4.1. We first introduce the following proposition, which shows that the model estimation step (4.2) satisfies Condition 3.1 and Condition 3.2.

Proposition D.1 (Guarantees for model estimation)**.** *Under Assumption 2.2, choosing the (pseudo) distance $D(\cdot\|\cdot)$ as KL-divergence or TV-distance, setting the tuning parameter ξ as*

$$\xi = \frac{C_1\log(C_2 H\mathcal{N}_{[]}(1/n^2,\mathcal{P}_{\mathrm{M}},\|\cdot\|_{1,\infty})/\delta)}{n},$$

for some constants $C_1, C_2 > 0$, then Condition 3.1 and 3.2 are satisfied respectively by,

♠ *when $D(\cdot\|\cdot)$ is KL-divergence and Assumption D.3 (See Appendix D.1) holds with parameter $\underline{\lambda}$, $\mathrm{Err}_h^{\mathbf{\Phi}}(n,\delta)$ is given by*

$$\sqrt{\mathrm{Err}^{\mathbf{\Phi}}_{h,\mathrm{KL}}(n,\delta)} = \frac{H\exp(H/\underline{\lambda})}{\rho}\cdot\sqrt{\frac{C_1'\log(C_2' H\mathcal{N}_{[]}(1/n^2,\mathcal{P}_{\mathrm{M}},\|\cdot\|_{1,\infty})/\delta)}{n}}.$$

♠ *when $D(\cdot\|\cdot)$ is TV-distance, $\mathrm{Err}_h^{\mathbf{\Phi}}(n,\delta)$ is given by*

$$\sqrt{\mathrm{Err}^{\mathbf{\Phi}}_{h,\mathrm{TV}}(n,\delta)} = H\cdot\sqrt{\frac{C_1'\log(C_2' H\mathcal{N}_{[]}(1/n^2,\mathcal{P}_{\mathrm{M}},\|\cdot\|_{1,\infty})/\delta)}{n}}.$$

Here c, C_1', $C_2' > 0$ stand for three universal constants.

Proof of Proposition D.1. See Appendix D.1 for a detailed proof. ☐

By Combing Proposition D.1 and Theorem 3.4, we can obtain Corollary 4.1. ☐

D.1 Proof of Proposition D.1

Lemma D.2 (Duality for KL-robust set)**.** *The following duality for KL-robust set holds,*

$$\inf_{Q(\cdot):D_{\mathrm{KL}}(Q(\cdot)\|Q^\star(\cdot))\le\sigma}\int f(x)Q(\mathrm{d}x) = \sup_{\lambda\in\mathbb{R}_+}\left\{-\lambda\log\left(\int\exp\{-f(x)/\lambda\}\,Q^\star(\mathrm{d}x)\right)-\lambda\sigma\right\}.$$

Proof of Lemma D.2. See [12, 68] for a detailed proof. ☐

Assumption D.3 (Regularity of KL-divergence duality variable)**.** *We assume that the optimal dual variable $\lambda^\star$ for the following optimization problem*

$$\sup_{\lambda\in\mathbb{R}_+}\left\{-\lambda\log\left(\mathbb{E}_{s'\sim P_h(\cdot|s_h,a_h)}\left[\exp\left\{-V^{\pi^\star}_{h+1,Q,\mathbf{\Phi}}(s')/\lambda\right\}\right]\right)-\lambda\rho\right\},$$

is lower bounded by $\underline{\lambda} > 0$ for any transition kernels $P_h\in\mathcal{P}_{\mathrm{M}}$, $Q=\{Q_h\}_{h=1}^H\subseteq\mathcal{P}_{\mathrm{M}}$, and step $h\in[H]$.

Lemma D.4 (Duality for TV-robust set)**.** *The following duality for TV-robust set holds,*

$$\inf_{Q(\cdot):D_{\mathrm{TV}}(Q(\cdot)\|Q^\star(\cdot))\le\sigma}\int f(x)Q(\mathrm{d}x) = \sup_{\lambda\in\mathbb{R}}\left\{-\int(\lambda-f(x))_+Q^\star(\mathrm{d}x)-\frac{\sigma}{2}(\lambda-\inf_x f(x))_+ +\lambda\right\}.$$

Proof of Lemma D.4. See [68] for a detailed proof. ☐

Proof of Proposition D.1 with KL-divergence. Firstly, by invoking the first conclusion of Lemma G.1, we know that the Condition 3.1 holds. In the following, we prove the Condition 3.2. By applying the dual formulation of the KL-robust set (Lemma D.2), we can derive that

$$\begin{aligned}
&\inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h^\star)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')]-\inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')]\\
&\quad=\sup_{\lambda\ge0}\left\{-\lambda\log\left(\mathbb{E}_{s'\sim P_h^\star(\cdot|s_h,a_h)}\left[\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda\right\}\right]\right)-\lambda\rho\right\}\\
&\qquad-\sup_{\lambda\ge0}\left\{-\lambda\log\left(\mathbb{E}_{s'\sim P_h(\cdot|s_h,a_h)}\left[\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda\right\}\right]\right)-\lambda\rho\right\}.
\end{aligned}\tag{D.1}$$

By Assumption D.3 and Lemma H.7, we know that the optimal value of λ for both two optimization problems in (D.1) lies in $[\underline{\lambda}, H/\rho]$ for some $\underline{\lambda} > 0$. Thus we can further upper bound the right hand side of (D.1) as

$$\begin{aligned}
\text{(D.1)} = & \sup_{\underline{\lambda} \le \lambda \le H/\rho} \left\{ -\lambda \log \left(\mathbb{E}_{s' \sim P_h^\star(\cdot|s_h,a_h)} \left[\exp \left\{ -V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')/\lambda \right\} \right] \right) - \lambda\rho \right\} \\
& - \sup_{\underline{\lambda} \le \lambda \le H/\rho} \left\{ -\lambda \log \left(\mathbb{E}_{s' \sim P_h(\cdot|s_h,a_h)} \left[\exp \left\{ -V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')/\lambda \right\} \right] \right) - \lambda\rho \right\} \\
\le & \sup_{\underline{\lambda} \le \lambda \le H/\rho} \left\{ \lambda \log \left(\frac{\mathbb{E}_{s' \sim P_h(\cdot|s_h,a_h)} \left[\exp \left\{ -V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')/\lambda \right\} \right]}{\mathbb{E}_{s' \sim P_h^\star(\cdot|s_h,a_h)} \left[\exp \left\{ -V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')/\lambda \right\} \right]} \right) \right\},
\end{aligned} \tag{D.2}$$

where in the second inequality we use the basic fact that $\sup_x f(x) - \sup_x g(x) \le \sup_x \{f(x) - g(x)\}$. Now we work on the right hand side of (D.2) and obtain that

$$\begin{aligned}
\text{(D.2)} = & \sup_{\underline{\lambda} \le \lambda \le H/\rho} \left\{ \lambda \log \left(1 + \frac{\left(\mathbb{E}_{s' \sim P_h(\cdot|s_h,a_h)} - \mathbb{E}_{s' \sim P_h^\star(\cdot|s_h,a_h)} \right) \left[\exp \left\{ -V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')/\lambda \right\} \right]}{\mathbb{E}_{s' \sim P_h^\star(\cdot|s_h,a_h)} \left[\exp \left\{ -V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')/\lambda \right\} \right]} \right) \right\} \\
\le & \sup_{\underline{\lambda} \le \lambda \le H/\rho} \left\{ \lambda \cdot \frac{\left(\mathbb{E}_{s' \sim P_h(\cdot|s_h,a_h)} - \mathbb{E}_{s' \sim P_h^\star(\cdot|s_h,a_h)} \right) \left[\exp \left\{ -V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')/\lambda \right\} \right]}{\mathbb{E}_{s' \sim P_h^\star(\cdot|s_h,a_h)} \left[\exp \left\{ -V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')/\lambda \right\} \right]} \right\},
\end{aligned} \tag{D.3}$$

where we use the fact of $\log(1 + x) \le x$ in the second inequality. Now we can further bound the right hand side of (D.3) by

$$\begin{aligned}
\text{(D.3)} \le & \frac{H \exp(H/\underline{\lambda})}{\rho} \cdot \left| \left(\mathbb{E}_{s' \sim P_h(\cdot|s_h,a_h)} - \mathbb{E}_{s' \sim P_h^\star(\cdot|s_h,a_h)} \right) \left[\exp \left\{ -V_{h+1,P,\mathbf{\Phi}}^{\pi}(s')/\lambda \right\} \right] \right| \\
\le & \frac{H \exp(H/\underline{\lambda})}{\rho} \cdot \int_{\mathcal{S}} |P_h(\mathrm{d}s'|s_h,a_h) - P_h^\star(\mathrm{d}s'|s_h,a_h)| \\
= & \frac{H \exp(H/\underline{\lambda})}{\rho} \cdot \|P_h(\cdot|s_h,a_h) - P_h^\star(\cdot|s_h,a_h)\|_{\mathrm{TV}}.
\end{aligned} \tag{D.4}$$

Thus by combining (D.1), (D.2), (D.3), and (D.4) we obtain that

$$\begin{aligned}
& \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)} [V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)} [V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] \\
& \quad \le \frac{H \exp(H/\underline{\lambda})}{\rho} \cdot \|P_h(\cdot|s_h,a_h) - P_h^\star(\cdot|s_h,a_h)\|_{\mathrm{TV}}.
\end{aligned} \tag{D.5}$$

By using a same argument for deriving (D.5), we can also obtain that

$$\begin{aligned}
& \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)} [V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)} [V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] \\
& \quad \le \frac{H \exp(H/\underline{\lambda})}{\rho} \cdot \|P_h(\cdot|s_h,a_h) - P_h^\star(\cdot|s_h,a_h)\|_{\mathrm{TV}}.
\end{aligned} \tag{D.6}$$

Therefore, due to (D.5) and (D.6), we can finally arrive at the following upper bound,

$$\begin{aligned}
& \mathbb{E}_{(s_h,a_h) \sim d_{P^\star,h}^{\pi^{\mathrm{b}}}} \left[\left(\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)} [V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)} [V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] \right)^2 \right] \\
& \quad \le \frac{H^2 \exp(2H/\underline{\lambda})}{\rho^2} \cdot \mathbb{E}_{(s_h,a_h) \sim d_{P^\star,h}^{\pi^{\mathrm{b}}}} [\|P_h(\cdot|s_h,a_h) - P_h^\star(\cdot|s_h,a_h)\|_{\mathrm{TV}}^2].
\end{aligned} \tag{D.7}$$

By invoking the second conclusion of Lemma G.1, we have that with probability at least $1 - \delta$,

$$\mathbb{E}_{(s_h,a_h) \sim d_{P^\star,h}^{\pi^{\mathrm{b}}}} [\|P_h(\cdot|s_h,a_h) - P_h^\star(\cdot|s_h,a_h)\|_{\mathrm{TV}}^2] \le \frac{C_1' \log(C_2' H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}, \tag{D.8}$$

for some absolute constant $C_1', C_2' > 0$. Now combining (D.7) and (D.8), we have that

$$\sqrt{\mathrm{Err}_{h,\mathrm{KL}}^{\boldsymbol{\Phi}}(n)} = \frac{H\exp(H/\underline{\lambda})}{\rho} \cdot \sqrt{\frac{C_1' \log(C_2' H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}}.$$

This finishes the proof of Proposition D.1 under KL-divergence. ☐

Proof of Proposition D.1 with TV-distance. Firstly, by invoking the first conclusion of Lemma G.1, we know that the Condition 3.1 holds. In the following, we prove the Condition 3.2. By applying the dual formulation of the TV-robust set (Lemma D.4), we can similarly derive that

$$\begin{aligned}
&\left| \inf_{\widetilde{P}_h \in \boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \boldsymbol{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')] \right| \\
&\quad = \left| \sup_{\lambda \in \mathbb{R}} \left\{ -\mathbb{E}_{s' \sim P_h^\star(\cdot|s_h,a_h)} \left[\left(\lambda - V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s') \right)_+ \right] - \frac{\rho}{2} \left(\lambda - \inf_{s'' \in \mathcal{S}} V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s'') \right) + \lambda \right\} \right. \\
&\qquad \left. - \sup_{\lambda \in \mathbb{R}} \left\{ -\mathbb{E}_{s' \sim P_h(\cdot|s_h,a_h)} \left[\left(\lambda - V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s') \right)_+ \right] - \frac{\rho}{2} \left(\lambda - \inf_{s'' \in \mathcal{S}} V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s'') \right) + \lambda \right\} \right|
\end{aligned} \tag{D.9}$$

$$\leq \left| \sup_{\lambda \in \mathbb{R}} \left\{ \left(\mathbb{E}_{s' \sim P_h^\star(\cdot|s_h,a_h)} - \mathbb{E}_{s' \sim P_h(\cdot|s_h,a_h)} \right) \left[\left(\lambda - V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s') \right)_+ \right] \right\} \right| \tag{D.10}$$

As is shown in Lemma H.8, the optimal value of λ for both two optimization problems in (D.9) lies in $[0, H]$. Thus we can further upper bound the right hand side of (D.10) as

$$\text{(D.10)} \leq H \cdot \|P_h(\cdot|s_h,a_h) - P_h^\star(\cdot|s_h,a_h)\|_{\mathrm{TV}}. \tag{D.11}$$

By applying the second conclusion of Lemma G.1, we conclude that with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_{(s_h,a_h)\sim d_{P^\star,h}^{\pi^{\mathrm{b}}}} \left[\left(\inf_{\widetilde{P}_h \in \boldsymbol{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')] \right)^2 \right] \\
&\quad \leq H^2 \cdot \mathbb{E}_{(s_h,a_h)\sim d_{P^\star,h}^{\pi^{\mathrm{b}}}} [\|P_h(\cdot|s_h,a_h) - P_h^\star(\cdot|s_h,a_h)\|_{\mathrm{TV}}^2] \\
&\quad \leq \frac{C_1' H^2 \log(C_2' H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}.
\end{aligned} \tag{D.12}$$

Therefore, it suffices to choose $\mathrm{Err}_{h,\mathrm{TV}}^{\boldsymbol{\Phi}}(n)$ as

$$\sqrt{\mathrm{Err}_{h,\mathrm{TV}}^{\boldsymbol{\Phi}}(n)} = H \cdot \sqrt{\frac{C_1' \log(C_2' H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}}.$$

This finishes the proof of Proposition D.1 under TV-distance. ☐

D.2 Proofs for $\mathcal{S} \times \mathcal{A}$-rectangular Robust Tabular MDP (Equation (4.4))

The model class $\mathcal{P}_{\mathrm{M}}$ can be considered as a subspace of $\mathcal{F} = \{f(s,a,s') : \|f\|_\infty \leq 1\}$ with finite $\mathcal{S}$ and $\mathcal{A}$. Consider the collection of brackets $\mathcal{B}$ containing brackets in the form of $[g, g+1/n^2]$, where $g(s,a,s') \in \{0, 1/n^2, 2/n^2, \cdots, (n^2-1)/n^2\}$. Then we can see that $\mathcal{B}$ is actually a $1/n^2$-bracket of $\mathcal{F}$. Thus we know that the bracket number of $\mathcal{P}_{\mathrm{M}}$ is bounded by,

$$\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}) \leq \mathcal{N}_{[]}(1/n^2, \mathcal{F}_{\mathrm{M}}, \|\cdot\|_\infty) \leq |\mathcal{B}| \leq n^{2|\mathcal{S}|^2|\mathcal{A}|}.$$

This finishes the proof of (4.4).

D.3 $\mathcal{S} \times \mathcal{A}$-rectangular Robust MDPs with Kernel Function Approximations

D.3.1 A Basic Review of Reproducing Kernel Hilbert Space

We briefly review the basic knowledge of a reproducing kernel Hilbert space (RKHS). We say $\mathcal{H}$ is a RKHS on a set $\mathcal{Y}$ with the reproducing kernel $\mathcal{K} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ if its inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ satisfies, for any $f \in \mathcal{H}$ and $y \in \mathcal{Y}$, we have that $f(y) = \langle f, \mathcal{K}(y, \cdot) \rangle_{\mathcal{H}}$. The mapping $\mathcal{K}(y, \cdot) : \mathcal{Y} \mapsto \mathcal{H}$ is called the feature mapping of $\mathcal{H}$, denoted by $\boldsymbol{\psi}(y) : \mathcal{Y} \mapsto \mathcal{H}$.

When the reproducing kernel $\mathcal{K}$ is continuous, symmetric, and positive definite, Mercer's theorem [50] says that $\mathcal{K}$ has the following representation,

$$\mathcal{K}(x, y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(y), \quad \forall x, y \in \mathcal{Y},$$

where $\psi_j : \mathcal{Y} \mapsto \mathbb{R}$ and $\{\sqrt{\lambda_j} \cdot \psi_j\}_{j=1}^{\infty}$ forms an orthonormal basis of $\mathcal{H}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Also, the feature mapping $\boldsymbol{\psi}(y)$ can be represented as

$$\boldsymbol{\psi}(y) = \sum_{j=1}^{+\infty} \lambda_j \psi_j(y) \psi_j, \quad \forall y \in \mathcal{Y}.$$

D.3.2 Bracket Number of Kernel Function Model Class and Suboptimality of Algorithm 1

For kernel function approximations via RKHS, our theoretical results rely on the following regularity assumptions on the RKHS involved in Example 2.7, which is commonly adopted in kernel function approximation literature for RL [70, 4, 23]. Specifically, the kernel $\mathcal{K}$ can be decomposed as $\mathcal{K}(x, y) = \sum_{j=1}^{+\infty} \lambda_j \psi_j(x) \psi_j(y)$ for some $\{\lambda_j\}_{j=1}^{+\infty} \subseteq \mathbb{R}$ and $\{\psi_j : \mathcal{X} \mapsto \mathbb{R}\}_{j=1}^{+\infty}$ with $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ (See Appendix D.3 for details). Our assumption on $\mathcal{K}$ is summarized in the following.

Assumption D.5 (Regularity of RKHS)**.** *We assume that the kernel $\mathcal{K}$ of the RKHS satisfies that:*

1. *(Boundedness) It holds that $|\mathcal{K}(x, y)| \leq 1$, $|\psi_j(x)| \leq 1$, and $|\lambda_j| \leq 1$ for any $j \in \mathbb{N}_+$, $x, y \in \mathcal{X}$.*
2. *(Eigenvalue decay) There exists some $\gamma \in (0, 1/2)$, $C_1, C_2 > 0$ such that $|\lambda_j| \leq C_1 \exp(-C_2 j^{\gamma})$ for any $j \in \mathbb{N}_+$.*

Under Assumption D.5, we can upper bound the bracket number $\mathcal{N}_{[]}$ of the realizable model space $\mathcal{P}_{\mathrm{M}}$ defined in (2.7) as (see Appendix D.3.3 for a proof),

$$\log(\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})) \leq C_{\mathrm{K}} \cdot 1/\gamma \cdot \log^2(1/\gamma) \cdot \log^{1+1/\gamma}(n\texttt{Vol}(\mathcal{S}) B_{\mathrm{K}}), \tag{D.13}$$

where $C_{\mathrm{K}} > 0$ is an absolute constant, $\texttt{Vol}(\mathcal{S})$ is the measure of the state space $\mathcal{S}$, and B_{K} is defined in Example 2.7. Combining (D.13) and Corollary 4.1, we can conclude that: i) under TV-distance the suboptimality of $\texttt{P}^2\texttt{MPO}$ for $\mathcal{S} \times \mathcal{A}$-rectangular robust MDPs with kernel function approximations is,

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \leq \mathcal{O}\left(H^2 \log(1/\gamma) \sqrt{C^{\star}_{P^{\star}, \boldsymbol{\Phi}}/\gamma \cdot \log^{1+1/\gamma}(nH\texttt{Vol}(\mathcal{S})/\delta)/n}\right), \tag{D.14}$$

and ii) under KL-divergence the suboptimality of $\texttt{P}^2\texttt{MPO}$ for $\mathcal{S} \times \mathcal{A}$-rectangular robust MDPs with kernel function approximations is,

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \leq \mathcal{O}\left(H^2 \exp(H/\underline{\lambda}) \log(1/\gamma)/\rho \sqrt{C^{\star}_{P^{\star}, \boldsymbol{\Phi}}/\gamma \cdot \log^{1+1/\gamma}(nH\texttt{Vol}(\mathcal{S})/\delta)/n}\right). \tag{D.15}$$

D.3.3 Proof of Equation (D.13)

We invoke the following lemma to bound the bracket number of $\mathcal{P}_{\mathrm{M}}$ in Example 2.7.

Lemma D.6 (Bracket number of kernel function class [25])**.** *Under Assumption D.5, the bracket number of $\mathcal{P}_{\mathrm{M}}$ given by*

$$\mathcal{P}_{\mathrm{M}} = \left\{P(s'|s, a) = \langle \boldsymbol{\psi}(s, a, s'), \boldsymbol{f} \rangle_{\mathcal{H}} : \boldsymbol{f} \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq B_{\mathrm{K}}\right\}$$

is bounded by, for any $\epsilon > 0$,

$$\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})) \leq C_{\mathrm{K}} \cdot 1/\gamma \cdot \log^2(1/\gamma) \cdot \log^{1+1/\gamma}(\texttt{Vol}(\mathcal{S}) B_{\mathrm{K}}/\epsilon).$$

Proof of Lemma D.6. We refer to Lemma B.11 in [25] for a detailed proof. □

By taking $\epsilon = 1/n^2$ in Lemma D.6, we can finish the proof of (D.13).

D.4 $\mathcal{S} \times \mathcal{A}$-rectangular Robust MDPs with Neural Function Approximations

For neural function approximations, we borrow the tool of neural tangent kernel (NTK [14]), which relates overparameterized neural networks (2.8) to kernel function approximations.

To this end, given the neural network (2.8), we define its NTK $\mathcal{K}_{\mathrm{NTK}} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ as

$$\mathcal{K}_{\mathrm{NTK}}(x, y) := \nabla_{\mathbf{W}} \mathrm{NN}(x, \mathbf{W}^0, \mathbf{a}^0)^\top \nabla_{\mathbf{W}} \mathrm{NN}(y, \mathbf{W}^0, \mathbf{a}^0), \quad \forall x, y \in \mathcal{X}. \tag{D.16}$$

Assumption D.7 (Regularity of Neural Tangent Kernel)**.** *We assume that the neural tangent kernel $\mathcal{K}_{\mathrm{NTK}}$ defined in* (D.16) *satisfies Assumption D.5 with constant $\gamma_{\mathrm{N}} \in (0, 1/2)$.*

This assumption on the spectral perspective of NTK is justified by [67]. As we prove in Appendix D.4.1, when the number of hidden units is large enough, i.e., overparameterized, the neural network is well approximated by its linear expansion at initialization (Lemma D.8), where we can apply the tool of NTK. Under Assumption D.7, the bracket number $\mathcal{N}_{[]}$ of $\mathcal{P}_{\mathrm{M}}$ defined in (2.9) is bounded by (see Appendix D.4.2 for a proof), for number of hidden units $m \geq d_{\mathcal{X}} n^4 B_{\mathrm{N}}^4$,

$$\log(\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})) \leq C_{\mathrm{N}} \cdot 1/\gamma_{\mathrm{N}} \cdot \log^2(1/\gamma_{\mathrm{N}}) \cdot \log^{1+1/\gamma_{\mathrm{N}}}(n \mathtt{Vol}(\mathcal{S}) B_{\mathrm{N}}), \tag{D.17}$$

where $C_{\mathrm{N}} > 0$ is an absolute constant, $\gamma_{\mathrm{N}} \in (0, 1/2)$ is specified in Assumption D.7, and B_{N} is defined in Example 2.8. Combining (D.17) and Corollary 4.1, we can conclude that, in the overparameterized paradigm, i.e., $m \geq d_{\mathcal{X}} n^4 B_{\mathrm{N}}^4$: i) under TV-distance the suboptimality of $\mathtt{P^2MPO}$ for $\mathcal{S} \times \mathcal{A}$-rectangular robust MDPs with neural function approximations is,

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \leq \mathcal{O}\left(H^2 \log(1/\gamma_{\mathrm{N}}) \sqrt{C^\star_{P^\star, \boldsymbol{\Phi}}/\gamma_{\mathrm{N}} \cdot \log^{1+1/\gamma_{\mathrm{N}}}(nH\mathtt{Vol}(\mathcal{S})/\delta)/n}\right), \tag{D.18}$$

and ii) under KL-divergence the suboptimality of $\mathtt{P^2MPO}$ for $\mathcal{S} \times \mathcal{A}$-rectangular robust MDPs with kernel function approximations is,

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \leq \mathcal{O}\left(H^2 \exp(H/\underline{\lambda}) \log(1/\gamma_{\mathrm{N}})/\rho \sqrt{C^\star_{P^\star, \boldsymbol{\Phi}}/\gamma_{\mathrm{N}} \cdot \log^{1+1/\gamma_{\mathrm{N}}}(nH\mathtt{Vol}(\mathcal{S})/\delta)/n}\right). \tag{D.19}$$

D.4.1 Neural Tangent Kernel and Implicit Linearization

We consider the overparameterized paradigm of the neural network (2.8) in the sense that the neural network is very wide, i.e., the number of hidden units m is large. The following lemma shows that in this paradigm, neural networks in $\mathcal{P}_{\mathrm{M}}$ are well approximated by a linear expansion at initialization.

Lemma D.8 (Implicit Linearization [4])**.** *Consider the two-layer neural network* NN *defined in* (2.8). *Assuming that the activation function $\sigma(\cdot)$ is* 1*-Lipschitz continuous and the input space $\mathcal{X}$ is normalized via $\|\mathbf{x}\|_2 \leq 1$ for any $\mathbf{x} \in \mathcal{X}$. Then it holds that*

$$\sup_{\mathbf{x} \in \mathcal{X}, \mathrm{NN}(\cdot; \mathbf{W}, \mathbf{a}^0) \in \mathcal{P}_{\mathrm{M}}} \left|\mathrm{NN}(\mathbf{x}; \mathbf{W}, \mathbf{a}^0) - \nabla_{\mathbf{W}} \mathrm{NN}(\mathbf{x}; \mathbf{W}^0, \mathbf{a}^0)^\top (\mathbf{W} - \mathbf{W}^0)\right| \leq d_{\mathcal{X}}^{1/2} B_{\mathrm{N}}^2 m^{-1/2}.$$

Proof of Lemma D.8. See the proof of Lemma 4.5 in [4] for a detailed proof. □

In view of Lemma D.8, we can study the linearization of the neural networks in $\mathcal{P}_{\mathrm{M}}$ as a surrogate. To this end, we introduce the neural tangent kernel $\mathcal{K}_{\mathrm{NTK}}$ of NN as

$$\mathcal{K}_{\mathrm{NTK}}(x, y) := \nabla_{\mathbf{W}} \mathrm{NN}(x, \mathbf{W}^0, \mathbf{a}^0)^\top \nabla_{\mathbf{W}} \mathrm{NN}(y, \mathbf{W}^0, \mathbf{a}^0), \quad \forall x, y \in \mathcal{X}.$$

The idea is to approximate the functions in $\mathcal{P}_{\mathrm{M}}$ via the RKHS induced by the kernel $\mathcal{K}_{\mathrm{NTK}}$. According to Lemma D.8, when the width of the neural network is large enough, i.e., $m \to \infty$, the approximation error is negligible. See the following Section D.4.2 for detailed proofs.

D.4.2 Proof of Equation (D.17)

Now we use Lemma D.8 to bound the bracket number of $\mathcal{P}_{\mathrm{M}}$ in Example 2.8.

Lemma D.9 (Bracket number of neural function class)**.** *Under Assumption D.7, for the number of hidden units $m \geq d_{\mathcal{X}} B_{\mathrm{N}}^4/\epsilon^2$, the bracket number of $\mathcal{P}_{\mathrm{M}}$ given by*

$$\mathcal{P}_{\mathrm{M}} = \left\{P(s'|s,a) = \mathrm{NN}((s,a,s'); \mathbf{W}, \mathbf{a}^0) : \|\mathbf{W} - \mathbf{W}^0\|_2 \leq B_{\mathrm{N}}\right\},$$

is bounded by, for any $\epsilon > 0$,

$$\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})) \leq C_{\mathrm{N}} \cdot 1/\gamma_{\mathrm{N}} \cdot \log^2(1/\gamma_{\mathrm{N}}) \cdot \log^{1+1/\gamma_{\mathrm{N}}}(\texttt{Vol}(\mathcal{S}) B_{\mathrm{K}}/\epsilon).$$

Proof of Lemma D.9. We denote the RKHS induced by the neural tangent kernel $\mathcal{K}_{\mathrm{NTK}}$ as $\mathcal{P}_{\mathrm{NTK}}$

$$\mathcal{P}_{\mathrm{NTK}} = \left\{\bar{P}(\mathbf{x}) = \nabla_{\mathbf{W}} \mathrm{NN}(\mathbf{x}; \mathbf{W}^0, \mathbf{a}^0)^\top (\mathbf{W} - \mathbf{W}^0) : \|\mathbf{W} - \mathbf{W}^0\|_2 \leq B_{\mathrm{N}}\right\}. \tag{D.20}$$

For any $\mathrm{NN}(\cdot; \mathbf{W}, \mathbf{a}^0) \in \mathcal{P}_{\mathrm{M}}$, we denote its linear expansion at initialization as $\overline{\mathrm{NN}}(\cdot; \mathbf{W}, \mathrm{a}^0) \in \mathcal{P}_{\mathrm{NTK}}$. Here we use the fact that for $\mathrm{NN}(\cdot; \mathbf{W}, \mathbf{a}^0) \in \mathcal{P}_{\mathrm{M}}$, $\|\mathbf{W} - \mathbf{W}^0\|_2 \leq B_{\mathrm{N}}$. Now according to Lemma D.6 and Assumption D.7, we know that the bracket number of $\mathcal{P}_{\mathrm{NTK}}$ is bounded by

$$\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{NTK}}, \|\cdot\|_{1,\infty})) \leq C \cdot 1/\gamma_{\mathrm{N}} \cdot \log^2(1/\gamma_{\mathrm{N}}) \cdot \log^{1+1/\gamma_{\mathrm{N}}}(\texttt{Vol}(\mathcal{S}) B_{\mathrm{N}}/\epsilon), \tag{D.21}$$

for some constant $C > 0$. Therefore, we can find a collect of brackets $\mathcal{B}_0 = \{[g_j^{\mathrm{l}}, g_j^{\mathrm{u}}]\}_{j \in [\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{NTK}}, \|\cdot\|_{1,\infty})]}$ such that for any $\bar{P} \in \mathcal{P}_{\mathrm{NTK}}$, there exists a bracket $[g_j^{\mathrm{l}}, g_j^{\mathrm{u}}] \in \mathcal{B}_0$ such that $g_j^{\mathrm{l}}(\mathbf{x}) \leq \bar{P}(\mathbf{x}) \leq g_j^{\mathrm{u}}(\mathbf{x})$ and $\|g_j^{\mathrm{l}} - g_j^{\mathrm{u}}\|_{1,\infty} \leq \epsilon$. Now for any $P = \mathrm{NN}(\cdot; \mathbf{W}, \mathbf{a}^0) \in \mathcal{P}_{\mathrm{M}}$, by Lemma D.8, we have that

$$\overline{\mathrm{NN}}(\mathbf{x}; \mathbf{W}, \mathbf{a}^0) - \epsilon_{\mathrm{N}} \leq \mathrm{NN}(\mathbf{x}; \mathbf{W}, \mathbf{a}^0) \leq \overline{\mathrm{NN}}(\mathbf{x}; \mathbf{W}, \mathbf{a}^0) + \epsilon_{\mathrm{N}},$$

where $\epsilon_{\mathrm{N}} = d_{\mathcal{X}}^{1/2} B_{\mathrm{N}}^2 m^{-1/2}$. By previous arguments, there exists a bracket $[g_j^{\mathrm{l}}, g_j^{\mathrm{u}}] \in \mathcal{B}_0$ such that

$$g_j^{\mathrm{l}}(\mathbf{x}) - \epsilon_{\mathrm{N}} \leq \mathrm{NN}(\mathbf{x}; \mathbf{W}, \mathbf{a}^0) \leq g_j^{\mathrm{u}}(\mathbf{x}) + \epsilon_{\mathrm{N}}.$$

Now it suffices to define a new collect of brackets $\mathcal{B} = \{[g_j^{\mathrm{l}} - \epsilon_{\mathrm{N}}, g_j^{\mathrm{u}} + \epsilon_{\mathrm{N}}]\}_{j \in [\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{NTK}}, \|\cdot\|_{1,\infty})]}$. For any $P = \mathrm{NN}(\cdot; \mathbf{W}, \mathbf{a}^0) \in \mathcal{P}_{\mathrm{M}}$, there exists a bracket $[\widetilde{g}_j^{\mathrm{l}}, \widetilde{g}_j^{\mathrm{u}}] \in \mathcal{B}$ such that $\widetilde{g}_j^{\mathrm{l}}(\mathbf{x}) \leq P(\mathbf{x}) \leq \widetilde{g}_j^{\mathrm{u}}(\mathbf{x})$, and

$$\|\widetilde{g}_j^{\mathrm{l}}(\mathbf{x}) - \widetilde{g}_j^{\mathrm{u}}(\mathbf{x})\|_{1,\infty} \leq \|g_j^{\mathrm{l}}(\mathbf{x}) - g_j^{\mathrm{u}}(\mathbf{x})\|_\infty + 2\epsilon_{\mathrm{N}} \leq \epsilon + 2\epsilon_{\mathrm{N}}.$$

By taking $m \geq d_{\mathcal{X}} B_{\mathrm{N}}^4/\epsilon^2$, we obtain that $\|\widetilde{g}_j^{\mathrm{l}}(\mathbf{x}) - \widetilde{g}_j^{\mathrm{u}}(\mathbf{x})\|_{1,\infty} \leq 3\epsilon$. Therefore, we can conclude that the bracket number of $\mathcal{P}_{\mathrm{M}}$ is bounded by,

$$\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}) = \mathcal{N}_{[]}(\epsilon/3, \mathcal{P}_{\mathrm{NTK}}, \|\cdot\|_{1,\infty}). \tag{D.22}$$

Finally, by combining (D.21) and (D.22), we have that, for $m \geq d_{\mathcal{X}} B_{\mathrm{N}}^4/\epsilon^2$,

$$\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})) \leq C_{\mathrm{N}} \cdot 1/\gamma_{\mathrm{N}} \cdot \log^2(1/\gamma_{\mathrm{N}}) \cdot \log^{1+1/\gamma_{\mathrm{N}}}(\texttt{Vol}(\mathcal{S}) B_{\mathrm{N}}/\epsilon),$$

for some constant $C_{\mathrm{N}} > 0$. This finishes the proof of Lemma D.9. ☐

Now by taking $\epsilon = 1/n^2$, i.e., $m \geq d_{\mathrm{X}} n^4 B_{\mathrm{N}}^4$, we can derive the desired result in (D.17).

E Proofs for $\mathcal{S} \times \mathcal{A}$-rectangular Robust Factored MDPs

Proof of Corollary 4.3. We first introduce the following proposition, which shows that te model estimation step (4.6) satisfies Condition 3.1 and Condition 3.2.

Proposition E.1 (Guarantees for model estimation)**.** *Suppose the RMDP is the $\mathcal{S} \times \mathcal{A}$-rectangular robust factored MDP in Example 2.9 with $D(\cdot\|\cdot)$ being KL-divergence or TV-distance. By choosing the tuning parameter ξ_i defined in (4.6) as*

$$\xi_i = \frac{C_1 |\mathcal{O}|^{1+|\mathrm{pa}_i|} |\mathcal{A}| \log(C_2 n d H/\delta)}{n}$$

for constants $C_1, C_2 > 0$ and each $i \in [d]$, then Condition 3.1 and 3.2 are satisfied respectively by,

♣ *when $D(\cdot\|\cdot)$ is KL-divergence and Assumption E.2 (See Appendix E.1) holds with parameter $\underline{\lambda}$, then $\mathrm{Err}_h^{\mathbf{\Phi}}(n,\delta)$ is given by*

$$\sqrt{\mathrm{Err}_{h,\mathrm{KL}}^{\mathbf{\Phi}}(n,\delta)} = \frac{H\exp(H/\underline{\lambda})}{\rho_{\min}} \cdot \sqrt{\frac{dC_1' \sum_{i=1}^d |\mathcal{O}|^{1+|\mathrm{pa}_i|}|\mathcal{A}|\log(C_2'nd/\delta)}{n}},$$

where $\rho_{\min} = \min_{i\in[d]}\rho_i$.

♣ *when $D(\cdot\|\cdot)$ is TV-distance, then $\mathrm{Err}_h^{\mathbf{\Phi}}(n,\delta)$ is given by*

$$\sqrt{\mathrm{Err}_{h,\mathrm{KL}}^{\mathbf{\Phi}}(n,\delta)} = H\sqrt{\frac{dC_1' \sum_{i=1}^d |\mathcal{O}|^{1+|\mathrm{pa}_i|}|\mathcal{A}|\log(C_2'nd/\delta)}{n}}.$$

Here c, C_1', $C_2' > 0$ stand for three universal constants.

Proof of Proposition E.1. See Appendix E.1 for a detailed proof. ☐

By Combing Proposition E.1 and Theorem 3.4, we can obtain Corollary 4.3. ☐

E.1 Proof of Proposition E.1

Assumption E.2 (Regularity of KL-divergence duality variable)**.** *We assume that the optimal dual variable $\lambda^\star$ for the following optimization problem*

$$\sup_{\lambda\in\mathbb{R}_+}\left\{-\lambda\log\left(\mathbb{E}_{s'[j]\sim P_{h,j}(\cdot|s_h[\mathrm{pa}_j],a_h)}\left[\exp\left\{-v_{h,T,Q,\mathbf{\Phi}}^j(s'[j])/\lambda\right\}\right]\right)-\lambda\rho\right\},$$

is lower bounded by $\underline{\lambda} > 0$ for any transition kernel $P_h \in \mathcal{P}_{\mathrm{M}}$, $T = \{T_h\}_{h=1}^H \subseteq \mathcal{P}_{\mathrm{M}}$, $Q = \{Q_h\}_{h=1}^H \subseteq \mathcal{P}_{\mathrm{M}}$, step $h\in[H]$, and factor $j\in[d]$. Here the function $v_{h,T,Q,\mathbf{\Phi}}^j(s'[j])$ is defined as

$$v_{h,T,Q,\mathbf{\Phi}}^j(s'[j]) = \int_{\mathcal{O}^{d-1}}\prod_{\substack{i=1\\ i\neq j}}^d T_{h,i}(\mathrm{d}s'[i])V_{h+1,Q,\mathbf{\Phi}}^{\pi^\star}(s'[1],\cdots,s'[j-1],s[j],s'[j+1],\cdots,s'[d]).$$

Proof of Proposition E.1 with KL-divergence. Firstly, by invoking the first conclusion of Lemma G.2, we know that the Condition 3.1 holds. In the following, we prove the Condition 3.2. By the definition of robust set in Example 2.9,

$$\begin{aligned}
&\inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h^\star)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')]\\
&\quad= \inf_{\widetilde{P}_{h,i}\in\Delta(\mathcal{O}):D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot)\|P_{h,i}(\cdot|s_h[\mathrm{pa}_i],a_h))\le\rho_i,i\in[d]}\int_{\mathcal{O}^d}\prod_{i=1}^d\widetilde{P}_{h,i}(\mathrm{d}s'[i])V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')\\
&\qquad- \inf_{\widetilde{P}_{h,i}\in\Delta(\mathcal{O}):D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot)\|P_{h,i}^\star(\cdot|s_h[\mathrm{pa}_i],a_h))\le\rho_i,i\in[d]}\int_{\mathcal{O}^d}\prod_{i=1}^d\widetilde{P}_{h,i}(\mathrm{d}s'[i])V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s').
\end{aligned}\tag{E.1}$$

Consider the following decomposition of the right hand side of (E.1),

$$\begin{aligned}
(\mathrm{E.1}) = \sum_{j=1}^d &\inf_{\substack{\widetilde{P}_{h,i}\in\Delta(\mathcal{O}):D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot)\|P_{h,i}(\cdot|s_h[\mathrm{pa}_i],a_h))\le\rho_i,1\le i\le j\\ \widetilde{P}_{h,i}\in\Delta(\mathcal{O}):D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot)\|P_{h,i}^\star(\cdot|s_h[\mathrm{pa}_i],a_h))\le\rho_i,j+1\le i\le d}}\int_{\mathcal{O}^d}\prod_{i=1}^d\widetilde{P}_{h,i}(\mathrm{d}s'[i])V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')\\
&- \inf_{\substack{\widetilde{P}_{h,i}\in\Delta(\mathcal{O}):D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot)\|P_{h,i}(\cdot|s_h[\mathrm{pa}_i],a_h))\le\rho_i,1\le i\le j-1\\ \widetilde{P}_{h,i}\in\Delta(\mathcal{O}):D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot)\|P_{h,i}^\star(\cdot|s_h[\mathrm{pa}_i],a_h))\le\rho_i,j\le i\le d}}\int_{\mathcal{O}^d}\prod_{i=1}^d\widetilde{P}_{h,i}(\mathrm{d}s'[i])V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s').
\end{aligned}$$

For each $1 \leq j \leq d$, we denote that

$$(\widetilde{P}_{h,1}^{*,j}, \cdots, \widetilde{P}_{h,d}^{*,j}) = \mathop{\mathrm{arginf}}_{\substack{\widetilde{P}_{h,i} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot) \| P_{h,i}(\cdot | s_h[\mathrm{pa}_i], a_h)) \leq \rho_i, 1 \leq i \leq j-1 \\ \widetilde{P}_{h,i} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot) \| P_{h,i}^{\star}(\cdot | s_h[\mathrm{pa}_i], a_h)) \leq \rho_i, j \leq i \leq d}} \int_{\mathcal{O}^d} \prod_{i=1}^{d} \widetilde{P}_{h,i}(\mathrm{d}s'[i]) V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')$$

By the definition of taking infimum over d variables, we can conclude that

$$\begin{aligned} &\inf_{\substack{\widetilde{P}_{h,i} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot) \| P_{h,i}(\cdot | s_h[\mathrm{pa}_i], a_h)) \leq \rho_i, 1 \leq i \leq j-1 \\ \widetilde{P}_{h,i} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot) \| P_{h,i}^{\star}(\cdot | s_h[\mathrm{pa}_i], a_h)) \leq \rho_i, j \leq i \leq d}} \int_{\mathcal{O}^d} \prod_{i=1}^{d} \widetilde{P}_{h,i}(\mathrm{d}s'[i]) V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \\ &\qquad = \inf_{\widetilde{P}_{h,j} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,j}(\cdot) \| P_{h,j}^{\star}(\cdot | s_h[\mathrm{pa}_j], a_h)) \leq \rho_j} \int_{\mathcal{O}^d} \widetilde{P}_{h,j}(\mathrm{d}s'[j]) \prod_{\substack{i=1 \\ i \neq j}}^{d} \widetilde{P}_{h,i}^{*,j}(\mathrm{d}s'[i]) V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s'). \end{aligned} \tag{E.2}$$

Meanwhile, it naturally holds that for each $1 \leq j \leq d$,

$$\begin{aligned} &\inf_{\substack{\widetilde{P}_{h,i} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot) \| P_{h,i}(\cdot | s_h[\mathrm{pa}_i], a_h)) \leq \rho_i, 1 \leq i \leq j \\ \widetilde{P}_{h,i} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,i}(\cdot) \| P_{h,i}^{\star}(\cdot | s_h[\mathrm{pa}_i], a_h)) \leq \rho_i, j+1 \leq i \leq d}} \int_{\mathcal{O}^d} \prod_{i=1}^{d} \widetilde{P}_{h,i}(\mathrm{d}s'[i]) V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \\ &\qquad \leq \inf_{\widetilde{P}_{h,j} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,j}(\cdot) \| P_{h,j}(\cdot | s_h[\mathrm{pa}_j], a_h)) \leq \rho_j} \int_{\mathcal{O}^d} \widetilde{P}_{h,j}(\mathrm{d}s'[j]) \prod_{\substack{i=1 \\ i \neq j}}^{d} \widetilde{P}_{h,i}^{*,j}(\mathrm{d}s'[i]) V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s'). \end{aligned} \tag{E.3}$$

Thus by combining (E.2) and (E.3), we have that

$$\begin{aligned} \text{(E.1)} \leq \sum_{j=1}^{d} & \inf_{\widetilde{P}_{h,j} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,j}(\cdot) \| P_{h,j}(\cdot | s_h[\mathrm{pa}_j], a_h)) \leq \rho_j} \int_{\mathcal{O}^d} \widetilde{P}_{h,j}(\mathrm{d}s'[j]) \prod_{\substack{i=1 \\ i \neq j}}^{d} \widetilde{P}_{h,i}^{*,j}(\mathrm{d}s'[i]) V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \\ & - \inf_{\widetilde{P}_{h,j} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,j}(\cdot) \| P_{h,j}^{\star}(\cdot | s_h[\mathrm{pa}_j], a_h)) \leq \rho_j} \int_{\mathcal{O}^d} \widetilde{P}_{h,j}(\mathrm{d}s'[j]) \prod_{\substack{i=1 \\ i \neq j}}^{d} \widetilde{P}_{h,i}^{*,j}(\mathrm{d}s'[i]) V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s'). \end{aligned} \tag{E.4}$$

Now for simplicity, for each $1 \leq j \leq d$, we denote a function $\upsilon_h^j(s'[j]) : \mathcal{O} \mapsto \mathbb{R}$ as

$$\upsilon_h^j(s'[j]) = \int_{\mathcal{O}^{d-1}} \prod_{\substack{i=1 \\ i \neq j}}^{d} \widetilde{P}_{h,i}^{*,j}(\mathrm{d}s'[i]) V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s'[1], \cdots, s'[j-1], s[j], s'[j+1], \cdots, s'[d]), \tag{E.5}$$

which satisfies $0 \leq \upsilon_h^j \leq H$. For each $1 \leq j \leq d$, we can then upper bound

$$\begin{aligned} \Delta_h^j(s_h, a_h) = & \inf_{\widetilde{P}_{h,j} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,j}(\cdot) \| P_{h,j}(\cdot | s_h[\mathrm{pa}_j], a_h)) \leq \rho_j} \int_{\mathcal{O}} \widetilde{P}_{h,j}(\mathrm{d}s'[j]) \upsilon_h^j(s'[j]) \\ & - \inf_{\widetilde{P}_{h,j} \in \Delta(\mathcal{O}) : D_{\mathrm{KL}}(\widetilde{P}_{h,j}(\cdot) \| P_{h,j}^{\star}(\cdot | s_h[\mathrm{pa}_j], a_h)) \leq \rho_j} \int_{\mathcal{O}} \widetilde{P}_{h,j}(\mathrm{d}s'[j]) \upsilon_h^j(s'[j]) \end{aligned} \tag{E.6}$$

using the same argument as in the proof of Proposition D.1 under KL-divergence in Appendix D.1, in which we apply Assumption E.2 and Lemma H.7. The corresponding result is given by

$$\Delta_h^j(s_h, a_h) \leq \frac{H \exp(H/\underline{\lambda})}{\rho_j} \cdot \| P_{h,j}(\cdot | s_h[\mathrm{pa}_j], a_h) - P_{h,j}^{\star}(\cdot | s_h[\mathrm{pa}_j], a_h) \|_{\mathrm{TV}}. \tag{E.7}$$

Thus plugging (E.7) into (E.4) and (E.1), we can arrive at

$$\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P^\star_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')]$$

$$\leq \sum_{j=1}^{d} \frac{H \exp(H/\underline{\lambda})}{\rho_j} \cdot \|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P^\star_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h)\|_{\mathrm{TV}}. \tag{E.8}$$

By using the same argument for deriving (E.8), we can also obtain that

$$\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P^\star_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')]$$

$$\leq \sum_{j=1}^{d} \frac{H \exp(H/\underline{\lambda})}{\rho_j} \cdot \|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P^\star_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h)\|_{\mathrm{TV}}. \tag{E.9}$$

Therefore, due to (E.8) and (E.9), we can finally arrive at the following upper bound,

$$\mathbb{E}_{(s_h,a_h)\sim d^{\pi^{\mathrm{b}}}_{P^\star,h}}\left[\left(\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P^\star_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')]\right)^2\right]$$

$$\leq \mathbb{E}_{(s_h,a_h)\sim d^{\pi^{\mathrm{b}}}_{P^\star,h}}\left[\left(\sum_{j=1}^{d} \frac{H \exp(H/\underline{\lambda})}{\rho_j} \cdot \|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P^\star_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h)\|_{\mathrm{TV}}\right)^2\right]$$

$$\leq \frac{dH^2 \exp(2H/\underline{\lambda})}{\rho_{\min}} \cdot \sum_{j=1}^{d} \mathbb{E}_{(s_h[\mathrm{pa}_j],a_h)\sim d^{\pi^{\mathrm{b}}}_{P^\star,h}}\left[\|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P^\star_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h)\|^2_{\mathrm{TV}}\right], \tag{E.10}$$

where the last inequality is from Cauchy-Schwartz inequality and $\rho_{\min} = \min_{i\in[d]} \rho_i$. Now invoking the second conclusion of Lemma G.2, we have that with probability at least $1-\delta$,

$$\mathbb{E}_{(s_h[\mathrm{pa}_j],a_h)\sim d^{\pi^{\mathrm{b}}}_{P^\star,h}}[\|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P^\star_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h)\|^2_{\mathrm{TV}}] \leq \frac{C'_1|\mathcal{O}|^{1+|\mathrm{pa}_j|}|\mathcal{A}|\log(C'_2 ndH/\delta)}{n}, \tag{E.11}$$

for some absolute constant $C'_1, C'_2 > 0$ and each $j \in [d]$. Combining (E.10) and (E.11), we have that

$$\sqrt{\mathrm{Err}^{\mathbf{\Phi}}_{h,\mathrm{KL}}(n)} = \frac{H \exp(H/\underline{\lambda})}{\rho_{\min}} \cdot \sqrt{\frac{dC'_1 \sum_{i=1}^{d} |\mathcal{O}|^{1+|\mathrm{pa}_i|}|\mathcal{A}|\log(C'_2 ndH/\delta)}{n}}.$$

This finishes the proof of Proposition E.1 under KL-divergence. ☐

Proof of Proposition E.1 with TV-distance. Firstly, by invoking the first conclusion of Lemma G.2, we know that the Condition 3.1 holds. In the following, we prove the Condition 3.2. Using the same argument as in the proof of Proposition E.1 under KL-divergence, we can derive that

$$\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P^\star_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')] \leq \sum_{j=1}^{d} \Delta^j_h(s_h,a_h), \tag{E.12}$$

where $\Delta^j_h(s_h,a_h)$ is defined in (E.6). Now applying the same argument as in the proof of Proposition D.1 under TV-divergence, we can derive that

$$\Delta^j_h(s_h,a_h) \leq H \cdot \|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P^\star_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h)\|_{\mathrm{TV}}, \tag{E.13}$$

where we have applied Lemma H.8. Therefore, by combining (E.12) and (E.13), we can derive that

$$\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P^\star_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')]$$

$$\leq H \cdot \sum_{j=1}^{d} \|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P^\star_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h)\|_{\mathrm{TV}}. \tag{E.14}$$

By the same argument as in deriving (E.14), we can also obtain that,

$$\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')]$$

$$\leq H \cdot \sum_{j=1}^{d} \|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P_{h,j}^\star(\cdot|s_h[\mathrm{pa}_j], a_h)\|_{\mathrm{TV}}. \tag{E.15}$$

Now by combining (E.14) and (E.15), we can derive the following upper bound,

$$\mathbb{E}_{(s_h,a_h)\sim d_{P^\star,h}^{\pi^{\mathrm{b}}}}\left[\left(\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')]\right)^2\right]$$

$$\leq \mathbb{E}_{(s_h,a_h)\sim d_{P^\star,h}^{\pi^{\mathrm{b}}}}\left[\left(H \cdot \sum_{j=1}^{d} \|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P_{h,j}^\star(\cdot|s_h[\mathrm{pa}_j], a_h)\|_{\mathrm{TV}}\right)^2\right]$$

$$\leq dH^2 \cdot \sum_{j=1}^{d} \mathbb{E}_{(s_h[\mathrm{pa}_j],a_h)\sim d_{P^\star,h}^{\pi^{\mathrm{b}}}}\left[\|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P_{h,j}^\star(\cdot|s_h[\mathrm{pa}_j], a_h)\|_{\mathrm{TV}}^2\right], \tag{E.16}$$

where the last inequality follows from Cauchy-Schwartz inequality. Now invoking the second conclusion of Lemma G.2, we have that with probability at least $1-\delta$,

$$\mathbb{E}_{(s_h[\mathrm{pa}_j],a_h)\sim d_{P^\star,h}^{\pi^{\mathrm{b}}}}[\|P_{h,j}(\cdot|s_h[\mathrm{pa}_j], a_h) - P_{h,j}^\star(\cdot|s_h[\mathrm{pa}_j], a_h)\|_{\mathrm{TV}}^2] \leq \frac{C_1'|\mathcal{O}|^{1+|\mathrm{pa}_j|}|\mathcal{A}|\log(C_2' ndH/\delta)}{n}, \tag{E.17}$$

for some absolute constant $C_1', C_2' > 0$ and each $j \in [d]$. Combining (E.16) and (E.17), we have that

$$\sqrt{\mathrm{Err}_{h,\mathrm{KL}}^{\mathbf{\Phi}}(n)} = H \cdot \sqrt{\frac{dC_1' \sum_{i=1}^{d} |\mathcal{O}|^{1+|\mathrm{pa}_i|}|\mathcal{A}|\log(C_2' ndH/\delta)}{n}}.$$

This finishes the proof of Proposition E.1 under TV-distance. ☐

F Proofs for d-rectangular Robust Linear MDP

Assumption F.1 (Regularity of KL-divergence duality variable)**.** *We assume that the optimal dual variable $\lambda^\star$ for the following optimization problem*

$$\sup_{\lambda \in \mathbb{R}_+} \left\{-\lambda \log\left(\mathbb{E}_{s'\sim\mu(\cdot)}\left[\exp\left\{-V_{h+1,Q,\mathbf{\Phi}}^{\pi^\star}(s')/\lambda\right\}\right]\right) - \lambda\rho\right\},$$

is lower bounded by $\underline{\lambda} > 0$ for any distribution $\mu \in \Delta(\mathcal{S})$, transition kernels $Q = \{Q_h\}_{h=1}^H \subseteq \mathcal{P}_{\mathrm{M}}$, and step $h \in [H]$.

Proof of Theorem A.3 with KL-divergence. Recall that we consider the following definition of $\mathcal{V}$,

$$\mathcal{V} = \left\{v(s) = \exp\left(-\max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \boldsymbol{w}/\lambda\right) : \|\boldsymbol{w}\|_2 \leq H\sqrt{d}, \lambda \in [\underline{\lambda}, H/\rho]\right\}. \tag{F.1}$$

Following the Section 7 of [54] as well as the Section 8 of [1], we introduce the notion $\widehat{P}_h$ that satisfies for any $v \in \mathcal{V}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\int_{\mathcal{S}} \widehat{P}_h(\mathrm{d}s'|s,a)v(s') = \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{\theta}}_v, \tag{F.2}$$

where $\widehat{\boldsymbol{\theta}}_v$ is defined in (A.2). Actually $\widehat{P}_h$ takes the following closed form,

$$\widehat{P}_h(\mathrm{d}s'|s,a) = \boldsymbol{\phi}(s,a)^\top \frac{1}{n}\sum_{\tau=1}^{n} \boldsymbol{\Lambda}_{h,\alpha}^{-1}\boldsymbol{\phi}(s_h^\tau, a_h^\tau)\delta_{s_{h+1}^\tau}(\mathrm{d}s'), \tag{F.3}$$

where $\delta_s(\cdot)$ is the dirac measure centering at s. Regarding the estimator $\widehat{P}_h$, we have the following.

Lemma F.2. *Setting $\alpha = 1$ and choosing the function class $\mathcal{V}$ as (F.1), then the estimator $\widehat{P}_h$ defined in (F.3) satisfies that, with probability at least $1-\delta$,*

$$\begin{aligned}\sup_{v\in\mathcal{V}}\left|\int_{\mathcal{S}}\big(P_h^\star(\mathrm{d}s'|s,a)-\widehat{P}_h(\mathrm{d}s'|s,a)\big)v(s')\right|^2 \\ \le C_1\cdot\|\boldsymbol{\phi}(s,a)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\cdot\frac{d\big(\log(1+C_2nH/\delta)+\log(1+C_3ndH/(\rho\underline{\lambda}^2))\big)}{n},\end{aligned}$$

for any step $h\in[H]$, where $C_1, C_2, C_3 > 0$ are three constants.

Proof of Lemma F.2. See Appendix F.1 for a detailed proof. ☐

With Lemma F.2, we can further derive that, with probability at least $1-\delta$, for any $h\in[H]$,

$$\begin{aligned}\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{\tau=1}^{n}\left|\int_{\mathcal{S}}\big(P_h^\star(\mathrm{d}s'|s_h^\tau,a_h^\tau)-\widehat{P}_h(\mathrm{d}s'|s_h^\tau,a_h^\tau)\big)v(s')\right|^2 \\ \le\frac{1}{n}\sum_{\tau=1}^{n}\|\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\cdot\frac{C_1d\big(\log(1+C_2nH/\delta)+\log(1+C_3ndH/(\rho\underline{\lambda}^2))\big)}{n}.\end{aligned}$$

In the right hand side of the above inequality, it holds that,

$$\begin{aligned}\frac{1}{n}\sum_{\tau=1}^{n}\|\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}\left(\boldsymbol{\phi}(s_h^\tau,a_h^\tau)^\top\boldsymbol{\Lambda}_{h,\alpha}^{-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\right) \\ &= \mathrm{Tr}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\boldsymbol{\phi}(s_h^\tau,a_h^\tau)^\top\boldsymbol{\Lambda}_{h,\alpha}^{-1}\right) \\ &\le \mathrm{Tr}\left(\boldsymbol{\Lambda}_{h,\alpha}\boldsymbol{\Lambda}_{h,\alpha}^{-1}\right)=d.\end{aligned}\tag{F.4}$$

Thus, we have that with probability at least $1-\delta$, for each step $h\in[H]$,

$$\begin{aligned}\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{\tau=1}^{n}\left|\int_{\mathcal{S}}\big(P_h^\star(\mathrm{d}s'|s_h^\tau,a_h^\tau)-\widehat{P}_h(\mathrm{d}s'|s_h^\tau,a_h^\tau)\big)v(s')\right|^2 \\ \le\frac{C_1d^2\big(\log(1+C_2nH/\delta)+\log(1+C_3ndH/(\rho\underline{\lambda}^2))\big)}{n}=\xi.\end{aligned}$$

This proves Condition 3.1 in Section 3.2. In the following, we prove Theorem A.3 given Condition 3.1 holds. Using the definition of robust set $\boldsymbol{\Phi}(\cdot)$ in Example A.1, we can derive that

$$\begin{aligned}&\inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')]-\inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')] \\ &\quad=\inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h^\star)}\sum_{i=1}^{d}\phi_i(s_h,a_h)\int_{\mathcal{S}}\widetilde{\mu}_i(\mathrm{d}s')V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')-\inf_{\widetilde{P}_h\in\boldsymbol{\Phi}(P_h)}\sum_{i=1}^{d}\phi_i(s,a)\int_{\mathcal{S}}\widetilde{\mu}_i(\mathrm{d}s')V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s') \\ &\quad=\sum_{i=1}^{d}\phi_i(s_h,a_h)\inf_{\widetilde{\mu}_{h,i}\in\Delta(\mathcal{S}):D(\widetilde{\mu}_{h,i}(\cdot)\|\mu^\star_{h,i}(\cdot))\le\rho}\int_{\mathcal{S}}\widetilde{\mu}_{h,i}(\mathrm{d}s')V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s') \\ &\qquad-\sum_{i=1}^{d}\phi_i(s_h,a_h)\inf_{\widetilde{\mu}_{h,i}\in\Delta(\mathcal{S}):D(\widetilde{\mu}_{h,i}(\cdot)\|\mu_{h,i}(\cdot))\le\rho}\int_{\mathcal{S}}\widetilde{\mu}_{h,i}(\mathrm{d}s')V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s'),\end{aligned}\tag{F.5}$$

where the last equality follows from $\phi(s,a)\ge 0$ for any $i\in[d]$. Now invoking the dual formulation of KL-divergence in Lemma D.2, we can derive that

$$\begin{aligned}(\text{F.5})=\sum_{i=1}^{d}\phi_i(s_h,a_h)\cdot\Bigg[&\sup_{\lambda_i\ge0}\left\{-\lambda_i\log\left(\mathbb{E}_{s'\sim\mu^\star_{h,i}(\cdot)}\left[\exp\left\{-V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')/\lambda_i\right\}\right]\right)-\lambda_i\rho\right\} \\ &-\sup_{\lambda_i\ge0}\left\{-\lambda_i\log\left(\mathbb{E}_{s'\sim\mu_{h,i}(\cdot)}\left[\exp\left\{-V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')/\lambda_i\right\}\right]\right)-\lambda_i\rho\right\}\Bigg]\end{aligned}\tag{F.6}$$

Following the same argument in the proof of Proposition D.1 (derivation of (D.3)), during which we invoke Assumption F.1 and Lemma H.7 to bound the optimal dual variable λ, we can derive that

$$
\begin{aligned}
(\text{F.6}) &\le \sum_{i=1}^{d} \phi_i(s_h,a_h)\cdot \sup_{\underline{\lambda}\le\lambda_i\le H/\rho}\left\{ g(\lambda_i,\mu^\star_{h,i}) \int_{\mathcal{S}} \left(\mu^\star_{h,i}(\mathrm{d}s') - \mu_{h,i}(\mathrm{d}s')\right) \exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\right\},\\
&= \sum_{i=1}^{d} \sup_{\underline{\lambda}\le\lambda_i\le H/\rho}\left\{ g(\lambda_i,\mu^\star_{h,i})\phi_i(s_h,a_h) \int_{\mathcal{S}} \left(\mu^\star_{h,i}(\mathrm{d}s') - \mu_{h,i}(\mathrm{d}s')\right) \exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\right\},
\end{aligned}
\tag{F.7}
$$

where we have defined $g(\lambda_i,\mu_{h,i}) = \lambda_i/(\int_{\mathcal{S}}\mu_{h,i}(\mathrm{d}s')\exp\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\})$ for simplicity, and in the equality we have used the fact that $\phi_i(s,a)\ge 0$. To go ahead, we rewrite the summand in (F.7) for each $i\in[d]$. To be specific, recall the regularized covariance matrix $\mathbf{\Lambda}_{h,\alpha}$ of the feature $\boldsymbol{\phi}$,

$$
\mathbf{\Lambda}_{h,\alpha} = \frac{1}{n}\sum_{\tau=1}^{n}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\boldsymbol{\phi}(s_h^\tau,a_h^\tau)^\top + \frac{\alpha}{n}\cdot\boldsymbol{I}_d.
$$

Then, by denoting $\mathbf{1}_i = (0,\cdots,0,1,0,\cdots,0)^\top$ where 1 is at the i-th coordinate, we have the following,

$$
\begin{aligned}
&\phi_i(s_h,a_h)\int_{\mathcal{S}}\left(\mu^\star_{h,i}(\mathrm{d}s') - \mu_{h,i}(\mathrm{d}s')\right)\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\\
&\quad = \phi_i(s_h,a_h)\mathbf{1}_i^\top\mathbf{\Lambda}_{h,\alpha}^{-1/2}\mathbf{\Lambda}_{h,\alpha}^{1/2}\int_{\mathcal{S}}\left(\boldsymbol{\mu}^\star_h(\mathrm{d}s') - \boldsymbol{\mu}_h(\mathrm{d}s')\right)\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\\
&\quad \le \underbrace{\|\phi_i(s_h,a_h)\mathbf{1}_i\|_{\mathbf{\Lambda}_{h,\alpha}^{-1}}}_{\text{Term (i)}}\cdot\underbrace{\left\|\int_{\mathcal{S}}\left(\boldsymbol{\mu}^\star_h(\mathrm{d}s') - \boldsymbol{\mu}_h(\mathrm{d}s')\right)\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\right\|_{\mathbf{\Lambda}_{h,\alpha}}}_{\text{Term (ii)}}.
\end{aligned}
\tag{F.8}
$$

For the term (ii) in (F.8), by the definition of $\mathbf{\Lambda}_{h,\alpha}$, we have that,

$$
\begin{aligned}
\text{Term (ii)}^2 &= \frac{1}{n}\sum_{\tau=1}^{n}\left|\boldsymbol{\phi}(s_h^\tau,a_h^\tau)^\top\int_{\mathcal{S}}\left(\boldsymbol{\mu}^\star_h(\mathrm{d}s') - \boldsymbol{\mu}_h(\mathrm{d}s')\right)\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\right|^2\\
&\qquad + \frac{\alpha}{n}\cdot\left\|\int_{\mathcal{S}}\left(\boldsymbol{\mu}^\star_h(\mathrm{d}s') - \boldsymbol{\mu}_h(\mathrm{d}s')\right)\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\right\|_2^2\\
&= \frac{1}{n}\sum_{\tau=1}^{n}\left|\int_{\mathcal{S}}\left(P^\star_h(\mathrm{d}s'|s_h^\tau,a_h^\tau) - P_h(\mathrm{d}s'|s_h^\tau,a_h^\tau)\right)\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\right|^2\\
&\qquad + \frac{\alpha}{n}\cdot\left\|\int_{\mathcal{S}}\left(\boldsymbol{\mu}^\star_h(\mathrm{d}s') - \boldsymbol{\mu}_h(\mathrm{d}s')\right)\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\right\|_2^2.
\end{aligned}
\tag{F.9}
$$

In the following, we upper bound the right hand side of (F.9). On the one hand, we have that

$$
\begin{aligned}
&\frac{1}{n}\sum_{\tau=1}^{n}\left|\int_{\mathcal{S}}\left(P^\star_h(\mathrm{d}s'|s_h^\tau,a_h^\tau) - P_h(\mathrm{d}s'|s_h^\tau,a_h^\tau)\right)\exp\left\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\right\}\right|^2\\
&\quad\le \sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{\tau=1}^{n}\left|\int_{\mathcal{S}}\left(P^\star_h(\mathrm{d}s'|s_h^\tau,a_h^\tau) - \widehat{P}_h(\mathrm{d}s'|s_h^\tau,a_h^\tau)\right)v(s')\right|^2\\
&\qquad + \sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{\tau=1}^{n}\left|\int_{\mathcal{S}}\left(\widehat{P}_h(\mathrm{d}s'|s_h^\tau,a_h^\tau) - P_h(\mathrm{d}s'|s_h^\tau,a_h^\tau)\right)v(s')\right|^2\\
&\quad\le 2\xi,
\end{aligned}
\tag{F.10}
$$

with probability at least $1-\delta$, where the first inequality holds since $\exp\{-V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')/\lambda_i\}\in\mathcal{V}$, and the last inequality follows from the fact that Condition 3.1 holds and the fact that $P_h\in\widehat{\mathcal{P}}_h$. On the other hand, by setting the regularization parameter $\alpha = 1$ we have that

$$\begin{aligned}
&\frac{\alpha}{n} \cdot \left\| \int_{\mathcal{S}} (\boldsymbol{\mu}_h^\star(\mathrm{d}s') - \boldsymbol{\mu}_h(\mathrm{d}s')) \exp\left\{-V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')/\lambda_i\right\} \right\|_2^2 \\
&\quad = \frac{1}{n} \cdot \sum_{i=1}^{d} \left| \int_{\mathcal{S}} \left(\mu_{h,i}^\star(\mathrm{d}s') - \mu_{h,i}(\mathrm{d}s')\right) \exp\left\{-V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')/\lambda_i\right\} \right|^2 \\
&\quad \le \frac{1}{n} \cdot \sum_{i=1}^{d} \|\mu_{h,i}^\star(\cdot) - \mu_{h,i}(\cdot)\|_{\mathrm{TV}}^2 \le \frac{2d}{n}. \qquad \text{(F.11)}
\end{aligned}$$

By combining (F.9), (F.10) and (F.11), we can conclude that with probability at least $1-\delta$,

$$\text{Term (ii)}^2 \le 2\xi + \frac{2d}{n} \le 3\xi. \qquad \text{(F.12)}$$

Now by combining (F.7), (F.8), (F.12), we can conclude that with probability at least $1-\delta$,

$$\begin{aligned}
&\inf_{\widetilde{P}_h \in \boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \boldsymbol{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')] \\
&\quad \le \sum_{i=1}^{d} \sup_{\underline{\lambda} \le \lambda_i \le H/\rho} \left\{ \|\phi_i(s_h,a_h)\mathbf{1}_i\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}} \cdot g(\lambda_i, \mu_{h,i}^\star) \cdot \sqrt{3\xi} \right\} \\
&\quad \le \frac{2\sqrt{\xi} \cdot H \exp(H/\underline{\lambda})}{\rho} \cdot \sum_{i=1}^{d} \|\phi_i(s_h,a_h)\mathbf{1}_i\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}, \qquad \text{(F.13)}
\end{aligned}$$

for any step $h \in [H]$, $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$, and $P_h \in \widehat{\mathcal{P}}_h$, where we apply the definition of $g(\lambda_i, \mu_i)$. Now using the same argument as in the proof of Theorem 3.4, using Condition 3.1, we can derive that

$$\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}; s_1) &\le \sup_{P \in \widehat{\mathcal{P}}} \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h) \sim d_{P^{\pi^\star,\dagger},h}^{\pi^\star}} \Bigg[\inf_{\widetilde{P}_h \in \boldsymbol{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')] \\
&\qquad\qquad - \inf_{\widetilde{P}_h \in \boldsymbol{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\boldsymbol{\Phi}}^{\pi^\star}(s')] \Bigg] \\
&\le \frac{2\sqrt{\xi} \cdot H \exp(H/\underline{\lambda})}{\rho} \cdot \sum_{h=1}^{H} \sum_{i=1}^{d} \mathbb{E}_{(s_h,a_h) \sim d_{P^{\pi^\star,\dagger},h}^{\pi^\star}} \left[\|\phi_i(s_h,a_h)\mathbf{1}_i\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}} \right], \qquad \text{(F.14)}
\end{aligned}$$

where we have used (F.13). Here $P_h^{\pi^\star,\dagger}$ is some transition kernel chosen from $\boldsymbol{\Phi}(P_h^\star)$. Now we upper bound the right hand side of (F.14) using Assumption A.2. Consider that

$$\begin{aligned}
&\sum_{i=1}^{d} \mathbb{E}_{(s_h,a_h) \sim d_{P^{\pi^\star,\dagger},h}^{\pi^\star}} \left[\|\phi_i(s_h,a_h)\mathbf{1}_i\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}} \right] \\
&\quad = \sum_{i=1}^{d} \mathbb{E}_{(s_h,a_h) \sim d_{P^{\pi^\star,\dagger},h}^{\pi^\star}} \left[\sqrt{\mathrm{Tr}\left((\phi_i(s_h,a_h)\mathbf{1}_i)(\phi_i(s_h,a_h)\mathbf{1}_i)^\top \boldsymbol{\Lambda}_{h,\alpha}^{-1} \right)} \right] \\
&\quad \le \sum_{i=1}^{d} \sqrt{\mathrm{Tr}\left(\mathbb{E}_{(s_h,a_h) \sim d_{P^{\pi^\star,\dagger},h}^{\pi^\star}} \left[(\phi_i(s_h,a_h)\mathbf{1}_i)(\phi_i(s_h,a_h)\mathbf{1}_i)^\top \right] \boldsymbol{\Lambda}_{h,\alpha}^{-1} \right)}. \qquad \text{(F.15)}
\end{aligned}$$

For notational simplicity, in the sequel, we denote by

$$\boldsymbol{\Sigma}_{P,h,i} = \mathbb{E}_{(s_h,a_h) \sim d_{P,h}^{\pi^\star}} \left[(\phi_i(s_h,a_h)\mathbf{1}_i)(\phi_i(s_h,a_h)\mathbf{1}_i)^\top \right]$$

Note that the matrix $\boldsymbol{\Sigma}_{P,h,i}$ has non-zero element only at $(\boldsymbol{\Sigma}_{P,h,i})_{(i,i)}$, which equals to $\phi_i(s,a)^2$. Under Assumption A.2 and the fact that $P_h^{\pi^\star,\dagger} \in \boldsymbol{\Phi}(P_h^\star)$, we have that

$$\boldsymbol{\Lambda}_{h,\alpha} \succeq \frac{\alpha}{n} \cdot \boldsymbol{I}_d + c^\dagger \cdot \boldsymbol{\Sigma}_{P^{\pi^\star,\dagger},h,i}.$$

Thus, using (F.15) and under $\alpha = 1$, we have that,

$$\begin{aligned}\sum_{i=1}^{d} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{P^{\pi^\star},\dagger,h}} \left[\|\phi_i(s_h,a_h)\mathbf{1}_i\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\right] &\leq \sum_{i=1}^{d} \sqrt{\mathrm{Tr}\left(\boldsymbol{\Sigma}_{P^{\pi^\star},h,i}\left(\frac{\alpha}{n}\cdot \boldsymbol{I}_d + c^\dagger \cdot \boldsymbol{\Sigma}_{P^{\pi^\star},h,i}\right)^{-1}\right)} \\ &= \sum_{i=1}^{d} \sqrt{\frac{\phi_i(s,a)^2}{n^{-1} + c^\dagger \cdot \phi_i(s,a)^2}} \leq \frac{d}{c^\dagger}. \end{aligned} \tag{F.16}$$

Therefore, by combining (F.14) and (F.16), we have that with probability at least $1-\delta$,

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \leq \frac{2\sqrt{\xi}\cdot H\exp(H/\underline{\lambda})}{\rho}\cdot\sum_{h=1}^{H}\frac{d}{c^\dagger} = \frac{2d\sqrt{\xi}\cdot H^2\exp(H/\underline{\lambda})}{c^\dagger\rho}.$$

Using the definition of ξ, we can finally derive that with probability at least $1-\delta$,

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \leq \frac{d^2H^2\exp(H/\underline{\lambda})}{c^\dagger\rho}\cdot\sqrt{\frac{C_1'\big(\log(1+C_2'nH/\delta)+\log(1+C_3'ndH/(\rho\underline{\lambda}^2))\big)}{n}}.$$

This finishes the proof of Theorem A.3 under KL-divergence. ☐

Proof of Theorem A.3 with TV-divergence. We use the same notation of $\widehat{P}_h$ introduced in the proof of KL-divergence case, which satisfies (F.2) with $\mathcal{V}$ defined as

$$\mathcal{V} = \left\{ v(s) = \left(\lambda - \max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \boldsymbol{w}\right)_+ : \|\boldsymbol{w}\|_2 \leq H\sqrt{d}, \lambda \in [0,H] \right\}. \tag{F.17}$$

Regarding the estimator $\widehat{P}_h$ with $\mathcal{V}$ defined in (F.17), we have the following.

Lemma F.3. *Setting $\alpha = 1$ and choosing the function class $\mathcal{V}$ as (F.17), then the estimator $\widehat{P}_h$ defined in (F.3) satisfies that, with probability at least $1-\delta$,*

$$\begin{aligned}&\sup_{v\in\mathcal{V}}\left|\int_{\mathcal{S}}\big(P_h^\star(\mathrm{d}s'|s,a) - \widehat{P}_h(\mathrm{d}s'|s,a)\big)v(s')\right|^2 \\ &\qquad \leq C_1\cdot\|\boldsymbol{\phi}(s,a)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\cdot\frac{dH^2\log(C_2ndH/\delta)}{n},\end{aligned}$$

for any step $h\in[H]$, where $C_1, C_2 > 0$ are two constants.

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

With Lemma F.3, we can further derive that, with probability at least $1-\delta$, for any $h\in[H]$,

$$\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{\tau=1}^{n}\left|\int_{\mathcal{S}}\big(P_h^\star(\mathrm{d}s'|s_h^\tau,a_h^\tau) - \widehat{P}_h(\mathrm{d}s'|s_h^\tau,a_h^\tau)\big)v(s')\right|^2 \leq \frac{1}{n}\sum_{\tau=1}^{n}\|\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\cdot\frac{C_1dH^2\log(C_2ndH/\delta)}{n}.$$

In the right hand side of the above inequality, it holds that,

$$\frac{1}{n}\sum_{\tau=1}^{n}\|\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}\left(\boldsymbol{\phi}(s_h^\tau,a_h^\tau)^\top\boldsymbol{\Lambda}_{h,\alpha}^{-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\right) \leq \mathrm{Tr}\left(\boldsymbol{\Lambda}_{h,\alpha}\boldsymbol{\Lambda}_{h,\alpha}^{-1}\right) = d. \tag{F.18}$$

Thus, we have that with probability at least $1-\delta$, for each step $h\in[H]$,

$$\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{\tau=1}^{n}\left|\int_{\mathcal{S}}\big(P_h^\star(\mathrm{d}s'|s_h^\tau,a_h^\tau) - \widehat{P}_h(\mathrm{d}s'|s_h^\tau,a_h^\tau)\big)v(s')\right|^2 \leq \frac{C_1d^2H^2\log(C_2ndH/\delta)}{n} = \xi.$$

This proves Condition 3.1 in Section 3.2. In the following, we prove Theorem A.3 given Condition 3.1 holds. Using the definition of robust set $\mathbf{\Phi}(\cdot)$ in Example A.1, following the same argument as (F.5), we have that,

$$
\begin{aligned}
&\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] \\
&\qquad = \sum_{i=1}^{d} \phi_i(s_h,a_h) \inf_{\widetilde{\mu}_{h,i} \in \Delta(\mathcal{S}) : D(\widetilde{\mu}_{h,i}(\cdot) \| \mu_{h,i}^\star(\cdot)) \le \rho} \int_{\mathcal{S}} \widetilde{\mu}_{h,i}(\mathrm{d}s') V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \\
&\qquad\qquad - \sum_{i=1}^{d} \phi_i(s_h,a_h) \inf_{\widetilde{\mu}_{h,i} \in \Delta(\mathcal{S}) : D(\widetilde{\mu}_{h,i}(\cdot) \| \mu_{h,i}(\cdot)) \le \rho} \int_{\mathcal{S}} \widetilde{\mu}_{h,i}(\mathrm{d}s') V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s').
\end{aligned} \tag{F.19}
$$

Now invoking the dual formulation of TV-distance in Lemma D.4, we can further derive that

$$
\begin{aligned}
(\text{F.19}) &= \sum_{i=1}^{d} \phi_i(s_h,a_h) \cdot \Bigg[\sup_{\lambda \in \mathbb{R}} \Bigg\{ -\mathbb{E}_{s' \sim \mu_{h,i}^\star(\cdot)} \bigg[\Big(\lambda - V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \Big)_+ \bigg] - \frac{\rho}{2} \Big(\lambda - \inf_{s'' \in \mathcal{S}} V_{h+1,P,\mathbf{\Phi}}^{\pi}(s'') \Big) + \lambda \Bigg\} \\
&\qquad - \sup_{\lambda \in \mathbb{R}} \Bigg\{ -\mathbb{E}_{s' \sim \mu_{h,i}(\cdot)} \bigg[\Big(\lambda - V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \Big)_+ \bigg] - \frac{\rho}{2} \Big(\lambda - \inf_{s'' \in \mathcal{S}} V_{h+1,P,\mathbf{\Phi}}^{\pi}(s'') \Big) + \lambda \Bigg\} \Bigg] \\
&\le \sum_{i=1}^{d} \phi_i(s_h,a_h) \cdot \sup_{\lambda \in [0,H]} \Bigg\{ \Big(\mathbb{E}_{s' \sim \mu_{h,i}^\star(\cdot)} - \mathbb{E}_{s' \sim \mu_{h,i}(\cdot)} \Big) \bigg[\Big(\lambda - V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \Big)_+ \bigg] \Bigg\} \\
&= \sum_{i=1}^{d} \sup_{\lambda \in [0,H]} \Bigg\{ \phi_i(s_h,a_h) \int_{\mathcal{S}} \big(\mu_{h,i}^\star(\mathrm{d}s') - \mu_{h,i}(\mathrm{d}s') \big) \Big(\lambda - V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \Big)_+ \Bigg\}.
\end{aligned} \tag{F.20}
$$

where in the first inequality we use Lemma H.8 to bound $\lambda \in [0,H]$. Now we consider each summand $i \in [d]$ in the right hand side of (F.20). We rewrite it as

$$
\begin{aligned}
&\phi_i(s_h,a_h) \int_{\mathcal{S}} \big(\mu_{h,i}^\star(\mathrm{d}s') - \mu_{h,i}(\mathrm{d}s') \big) \Big(\lambda - V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \Big)_+ \\
&\qquad = \phi_i(s_h,a_h) \mathbf{1}_i^\top \mathbf{\Lambda}_{h,\alpha}^{-1/2} \mathbf{\Lambda}_{h,\alpha}^{1/2} \int_{\mathcal{S}} \big(\boldsymbol{\mu}_h^\star(\mathrm{d}s') - \boldsymbol{\mu}_h(\mathrm{d}s') \big) \Big(\lambda - V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \Big)_+ \\
&\qquad \le \underbrace{\| \phi_i(s_h,a_h) \mathbf{1}_i \|_{\mathbf{\Lambda}_{h,\alpha}^{-1}}}_{\text{Term (i)}} \cdot \underbrace{\bigg\| \int_{\mathcal{S}} \big(\boldsymbol{\mu}_h^\star(\mathrm{d}s') - \boldsymbol{\mu}_h(\mathrm{d}s') \big) \Big(\lambda - V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s') \Big)_+ \bigg\|_{\mathbf{\Lambda}_{h,\alpha}}}_{\text{Term (ii)}}.
\end{aligned} \tag{F.21}
$$

Following the same argument as (F.9), (F.10), and (F.11), using the fact that $(\lambda - V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s'))_+ \in \mathcal{V}$ with $\mathcal{V}$ in (F.17), we can derive that with probability at least $1-\delta$,

$$
\mathrm{Term(ii)}^2 \le 3\xi \tag{F.22}
$$

Now by combining (F.19), (F.21), (F.22), we can conclude that with probability at least $1-\delta$,

$$
\begin{aligned}
&\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] \\
&\qquad \le \sum_{i=1}^{d} \sup_{0 \le \lambda_i H} \Big\{ \| \phi_i(s_h,a_h) \mathbf{1}_i \|_{\mathbf{\Lambda}_{h,\alpha}^{-1}} \cdot \sqrt{3\xi} \Big\} \le 2\sqrt{\xi} \cdot \sum_{i=1}^{d} \| \phi_i(s_h,a_h) \mathbf{1}_i \|_{\mathbf{\Lambda}_{h,\alpha}^{-1}},
\end{aligned} \tag{F.23}
$$

for any step $h \in [H]$, $(s_h,a_h) \in \mathcal{S} \times \mathcal{A}$, and $P_h \in \widehat{\mathcal{P}}_h$. Now using the same argument as in the proof of Theorem 3.4, using Condition 3.1, we can derive that with probability at least $1-\delta$,

$$
\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}; s_1) &\le \sup_{P \in \widehat{\mathcal{P}}} \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h) \sim d_{P^{\pi^\star},\dagger,h}^{\pi^\star}} \Bigg[\inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h^\star)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] \\
&\qquad - \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s' \sim \widetilde{P}_h(\cdot|s_h,a_h)}[V_{h+1,P,\mathbf{\Phi}}^{\pi^\star}(s')] \Bigg] \\
&\le 2\sqrt{\xi} \cdot \sum_{h=1}^{H} \sum_{i=1}^{d} \mathbb{E}_{(s_h,a_h) \sim d_{P^{\pi^\star},\dagger,h}^{\pi^\star}} \Big[\| \phi_i(s_h,a_h) \mathbf{1}_i \|_{\mathbf{\Lambda}_{h,\alpha}^{-1}} \Big],
\end{aligned} \tag{F.24}
$$

where in the last inequality we apply (F.23). Here $P_h^{\pi^\star,\dagger}$ is some transition kernel chosen from $\mathbf{\Phi}(P_h^\star)$. Now we use the same argument as (F.15) and (F.16) to upper bound the right hand side of (F.24) using Assumption A.2, which gives that,

$$\sum_{i=1}^{d} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{P^{\pi^\star,\dagger},h}} \left[\|\phi_i(s_h,a_h)\mathbf{1}_i\|_{\mathbf{\Lambda}_{h,\alpha}^{-1}} \right] \leq \frac{d}{c^\dagger}. \tag{F.25}$$

Therefore, by combining (F.24) and (F.25), we have that with probability at least $1-\delta$,

$$\mathrm{SubOpt}(\widehat{\pi};s_1) \leq 2\sqrt{\xi} \cdot \sum_{h=1}^{H} \frac{d}{c^\dagger} = \frac{2d\sqrt{\xi}\cdot H}{c^\dagger}.$$

Using the definition of ξ, we can finally derive that with probability at least $1-\delta$,

$$\mathrm{SubOpt}(\widehat{\pi};s_1) \leq \frac{d^2H^2}{c^\dagger} \cdot \sqrt{\frac{C_1' \log(C_2' ndH/\delta)}{n}}.$$

This finishes the proof of Theorem A.3 under TV-distance. □

F.1 Proof of Lemma F.2 and Lemma F.3

Proof of Lemma F.2. The proof of Lemma F.2 follows from the main proofs in Section 8 of [1] and the covering number of the function class $\mathcal{V}$ (Lemma F.4). Denote $\mathcal{C}_{\mathcal{V},\epsilon}$ as an ϵ-cover of the function class $\mathcal{V}$ under $\|\cdot\|_\infty$. Following the exact same argument of Lemma 8.7 in [1], we can derive that with probability at least $1-\delta$, for any h and $v \in \mathcal{C}_{\mathcal{V},\epsilon}$.

$$\begin{aligned} &\left\| \sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau,a_h^\tau) \left(\int_{\mathcal{S}} P_h^\star(\mathrm{d}s'|s_h^\tau,a_h^\tau)v(s') - v(s_{h+1}^\tau) \right) \right\|_{\mathbf{\Lambda}_{h,\alpha}^{-1}}^2 \\ &\qquad \leq 9n \cdot \left(\log(H/\delta) + \log(|\mathcal{C}_{\mathcal{V},\epsilon}|) + d\log(1+N) \right), \end{aligned} \tag{F.26}$$

where we have taken $\alpha = 1$, which we will keep in the following. For any function $v \in \mathcal{V}$, take $\widehat{v} \in \mathcal{C}_{\mathcal{V},\epsilon}$ such that $\|v - \widehat{v}\|_\infty \leq \epsilon$. Then we have that

$$\begin{aligned} &\left\| \sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau,a_h^\tau) \left(\int_{\mathcal{S}} P_h^\star(\mathrm{d}s'|s_h^\tau,a_h^\tau)v(s') - v(s_{h+1}^\tau) \right) \right\|_{\mathbf{\Lambda}_{h,\alpha}^{-1}}^2 \\ &\quad \leq 2 \left\| \sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau,a_h^\tau) \left(\int_{\mathcal{S}} P_h^\star(\mathrm{d}s'|s_h^\tau,a_h^\tau)\widehat{v}(s') - \widehat{v}(s_{h+1}^\tau) \right) \right\|_{\mathbf{\Lambda}_{h,\alpha}^{-1}}^2 \\ &\qquad + 2 \left\| \sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau,a_h^\tau) \left(\int_{\mathcal{S}} P_h^\star(\mathrm{d}s'|s_h^\tau,a_h^\tau)(\widehat{v}-v)(s') - (\widehat{v}-v)(s_{h+1}^\tau) \right) \right\|_{\mathbf{\Lambda}_{h,\alpha}^{-1}}^2 \\ &\quad \leq 18n \cdot \left(\log(H/\delta) + \log(|\mathcal{C}_{\mathcal{V},\epsilon}|) + d\log(1+n) \right) + 8\epsilon^2 n^2. \end{aligned} \tag{F.27}$$

Now we apply the definition of $\widehat{P}_h$ and we can then derive that

$$
\begin{aligned}
&\left|\int_{\mathcal{S}} \big(P_h^\star(\mathrm{d}s'|s,a) - \widehat{P}_h(\mathrm{d}s'|s,a)v(s')\big)\right|^2 \\
&\quad = \left|\boldsymbol{\phi}(s,a)^\top \left(\int_{\mathcal{S}} \boldsymbol{\mu}^\star(\mathrm{d}s')v(s') - \frac{1}{n}\sum_{\tau=1}^{n} \boldsymbol{\Lambda}_{h,\alpha}^{-1}\boldsymbol{\phi}(s_h^\tau, a_h^\tau)v(s_{h+1}^\tau)\right)\right|^2 \\
&\quad = \left|\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Lambda}_{h,\alpha}^{-1}\left(\boldsymbol{\Lambda}_{h,\alpha}\int_{\mathcal{S}} \boldsymbol{\mu}^\star(\mathrm{d}s')v(s') - \frac{1}{n}\sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)v(s_{h+1}^\tau)\right)\right|^2 \\
&\quad = \left|\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Lambda}_{h,\alpha}^{-1}\left(\frac{1}{n}\int_{\mathcal{S}} \boldsymbol{\mu}_h^\star(\mathrm{d}s')v(s') + \frac{1}{n}\sum_{\tau=1}^{n} \boldsymbol{\phi}(s,a)\int_{\mathcal{S}} P_h^\star(\mathrm{d}s'|s_h^\tau, a_h^\tau)v(s') \right.\right. \\
&\qquad\qquad \left.\left. -\frac{1}{n}\sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)v(s_{h+1}^\tau)\right)\right|^2 \\
&\quad \le \frac{2}{n^2}\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \cdot \left\|\int_{\mathcal{S}} \boldsymbol{\mu}^\star(\mathrm{d}s')v(s')\right\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \\
&\qquad + \frac{2}{n^2}\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \cdot \left\|\sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\left(\int_{\mathcal{S}} P_h^\star(\mathrm{d}s'|s_h^\tau, a_h^\tau)v(s') - v(s_{h+1}^\tau)\right)\right\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 .
\end{aligned}
\tag{F.28}
$$

On the one hand, the first term in the right hand side of (F.28) is bounded by

$$
\begin{aligned}
\frac{2}{n^2}\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \cdot \left\|\int_{\mathcal{S}} \boldsymbol{\mu}^\star(\mathrm{d}s')v(s')\right\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 &\le \frac{2}{n}\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \cdot \left\|\int_{\mathcal{S}} \boldsymbol{\mu}^\star(\mathrm{d}s')v(s')\right\|_{2}^2 \\
&\le \frac{2d}{n}\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 ,
\end{aligned}
\tag{F.29}
$$

where we use the fact that $\boldsymbol{\Lambda}_{h,\alpha} \succeq (1/n)\cdot \boldsymbol{I}_d$ and $\|v(\cdot)\|_\infty \le 1$ for any $v \in \mathcal{V}$. On the other hand, the second term in the right hand side of (F.28) is bounded by

$$
\begin{aligned}
&\frac{2}{n^2}\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \cdot \left\|\sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\left(\int_{\mathcal{S}} P_h^\star(\mathrm{d}s'|s_h^\tau, a_h^\tau)v(s') - v(s_{h+1}^\tau)\right)\right\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \\
&\quad \le \left(\frac{36}{n}\cdot\big(\log(H/\delta) + \log(|\mathcal{C}_{\mathcal{V},\epsilon}|) + d\log(1+n)\big) + 16\epsilon^2\right)\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 ,
\end{aligned}
$$

where we have applied (F.27). Now taking $\epsilon = 1/\sqrt{n}$, applying Lemma F.4 to bound the covering number of $\mathcal{V}$, we can further derive that,

$$
\begin{aligned}
&\frac{2}{n^2}\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \cdot \left\|\sum_{\tau=1}^{n} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\left(\int_{\mathcal{S}} P_h^\star(\mathrm{d}s'|s_h^\tau, a_h^\tau)v(s') - v(s_{h+1}^\tau)\right)\right\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \\
&\quad \le \frac{36}{n}\cdot\big(\log(H/\delta) + d\log(1+4\sqrt{n}Hd/(\underline{\lambda})) + \log(1+4\sqrt{n}Hd/(\underline{\lambda}^2\rho)) + d\log(1+n)\big)\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 \\
&\qquad + \frac{16}{n}\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 , \\
&\quad \le \frac{C_1 d\big(\log(1+C_2 nH/\delta) + \log(1+C_3 ndH/(\rho\underline{\lambda}^2))\big)}{n}\cdot \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2 ,
\end{aligned}
\tag{F.30}
$$

where $C_1, C_2, C_3 > 0$ are three constants. Finally, by combining (F.28), (F.29), and (F.30), we can conclude that with probability at least $1-\delta$, for each step $h \in [H]$,

$$\begin{aligned}&\sup_{v\in\mathcal{V}}\left|\int_{\mathcal{S}}\big(P_h^{\star}(\mathrm{d}s'|s,a)-\widehat{P}_h(\mathrm{d}s'|s,a)\big)v(s')\right|^2\\&\qquad\le C_1'\cdot\|\boldsymbol{\phi}(s,a)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\cdot\frac{d\big(\log(1+C_2nH/\delta)+\log(1+C_3ndH/(\rho\underline{\lambda}^2))\big)}{n}.\end{aligned}$$

where C_1' is another constant. This finishes the proof of Lemma F.2. ☐

Proof of Lemma F.3. The proof of Lemma F.3 follows the same argument as proof of Lemma F.2, except a different covering number of the function class $\mathcal{V}$ which we show in the following. Using the same argument as the proof of Lemma F.2, with probability at least $1-\delta$, for any $v\in\mathcal{V}$,

$$\begin{aligned}&\left|\int_{\mathcal{S}}\big(P_h^{\star}(\mathrm{d}s'|s,a)-\widehat{P}_h(\mathrm{d}s'|s,a)v(s')\big)\right|^2\\&\quad\le\frac{2H^2}{n^2}\cdot\|\boldsymbol{\phi}(s,a)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\cdot\left\|\int_{\mathcal{S}}\boldsymbol{\mu}^{\star}(\mathrm{d}s')v(s')\right\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\\&\qquad+\frac{2H^2}{n^2}\cdot\|\boldsymbol{\phi}(s,a)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\cdot\left\|\sum_{\tau=1}^{n}\boldsymbol{\phi}(s_h^{\tau},a_h^{\tau})\left(\int_{\mathcal{S}}P_h^{\star}(\mathrm{d}s'|s_h^{\tau},a_h^{\tau})v(s')-v(s_{h+1}^{\tau})\right)\right\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\\&\quad\le H^2\cdot\left(\frac{36}{n}\cdot(\log(H/\delta)+\log(|\mathcal{C}_{\mathcal{V},\epsilon}|)+d\log(1+n))+16\epsilon^2+\frac{2d}{n}\right)\cdot\|\boldsymbol{\phi}(s,a)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}},\end{aligned}\tag{F.31}$$

where $\mathcal{C}_{\mathcal{V},\epsilon}$ is an ϵ-covering of the function class $\mathcal{V}$ defined in (F.17). Now taking $\epsilon = 1/\sqrt{n}$, applying Lemma F.5 to bound the covering number of $\mathcal{V}$, we can further derive that,

$$\begin{aligned}&\sup_{v\in\mathcal{V}}\left|\int_{\mathcal{S}}\big(P_h^{\star}(\mathrm{d}s'|s,a)-\widehat{P}_h(\mathrm{d}s'|s,a)v(s')\big)\right|^2\\&\qquad\le H^2\cdot\|\boldsymbol{\phi}(s,a)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\cdot\left(\frac{36}{n}\cdot\big(\log(H/\delta)\right.\\&\qquad\qquad\left.+d\log(1+4\sqrt{n}Hd)+\log(1+4\sqrt{n}H)+d\log(1+n)\big)+\frac{16+2d}{n}\right)\\&\qquad\le C_1\cdot\|\boldsymbol{\phi}(s,a)\|^2_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}\cdot\frac{dH^2\log(C_2ndH/\delta)}{n}.\end{aligned}\tag{F.32}$$

This finishes the proof of Lemma F.3. ☐

F.2 Other Lemmas

Lemma F.4 (Covering number of $\mathcal{V}$: KL-divergence case)**.** *The ϵ-covering number of function class $\mathcal{V}$ defined in* (F.1) *under $\|\cdot\|_\infty$-norm is bounded by*

$$\log(\mathcal{N}(\epsilon,\mathcal{V},\|\cdot\|_\infty))\le d\log(1+4Hd/(\underline{\lambda}\epsilon))+\log(1+4H^2d/(\underline{\lambda}^2\rho\epsilon)).$$

Proof of Lemma F.4. Consider any two pairs of parameters $(\boldsymbol{w},\lambda)$ and $(\widehat{\boldsymbol{w}},\widehat{\lambda})$, and denote the functions they induce as v and $\widehat{v}$. Then we have that

$$|v(s)-\widehat{v}(s)|=\left|\exp\left\{-\max_{a\in\mathcal{A}}\boldsymbol{\phi}(s,a)^{\top}\boldsymbol{w}/\lambda\right\}-\exp\left\{-\max_{a\in\mathcal{A}}\boldsymbol{\phi}(s,a)^{\top}\widehat{\boldsymbol{w}}/\widehat{\lambda}\right\}\right|$$

Using the fact that, for any $x, y > 0$, $\exp(-x) - \exp(-y) = \exp(-\zeta(x, y)) \cdot (y - x)$ for some $\zeta(x, y)$ between x and y, we know that

$$
\begin{aligned}
&|v(s) - \widehat{v}(s)| \\
&\quad \le \exp\left\{-\zeta\left(\max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \boldsymbol{w}/\lambda, \max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}/\widehat{\lambda}\right)\right\} \cdot \left|\max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \boldsymbol{w}/\lambda - \max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}/\widehat{\lambda}\right| \\
&\quad \le \left|\max_{a\in\mathcal{A}} \left\{\boldsymbol{\phi}(s,a)^\top \boldsymbol{w}/\lambda - \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}/\widehat{\lambda}\right\}\right| \\
&\quad = \left|\max_{a\in\mathcal{A}} \left\{\boldsymbol{\phi}(s,a)^\top \boldsymbol{w}/\lambda - \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}/\lambda + \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}/\lambda - \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}/\widehat{\lambda}\right\}\right|.
\end{aligned}
$$

Notice that $\|\boldsymbol{\phi}(s,a)\|_2 \le \sqrt{d}$ (because $\sum_{i=1}^d \phi_i(s,a) = 1$), $\|\widehat{\boldsymbol{w}}\|_2 \le H\sqrt{d}$, and $\lambda, \widehat{\lambda} \ge \underline{\lambda}$, we have,

$$
\begin{aligned}
&\left|\boldsymbol{\phi}(s,a)^\top \boldsymbol{w}/\lambda - \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}/\lambda + \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}/\lambda - \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}/\widehat{\lambda}\right| \\
&\quad \le \left|\lambda^{-1}\boldsymbol{\phi}(s,a)^\top(\boldsymbol{w} - \widehat{\boldsymbol{w}})\right| + \left|\lambda^{-1}\widehat{\lambda}^{-1}\boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}(\lambda - \widehat{\lambda})\right| \\
&\quad \le \underline{\lambda}^{-1}\sqrt{d} \cdot \|\boldsymbol{w} - \widehat{\boldsymbol{w}}\|_2 + \underline{\lambda}^{-2} Hd \cdot |\lambda - \widehat{\lambda}|.
\end{aligned}
$$

Thus we conclude that to form an ϵ-cover of $\mathcal{V}$ under $\|\cdot\|_\infty$-norm, it suffices to consider the product of an $\underline{\lambda}\epsilon/(2\sqrt{d})$-cover of $\{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \le H\sqrt{d}\}$ under $\|\cdot\|_2$-norm and an $\underline{\lambda}^2\epsilon/(2Hd)$-cover of the interval $[\underline{\lambda}, H/\rho]$. Therefore, we can derive that

$$
\log(\mathcal{N}(\epsilon, \mathcal{V}, \|\cdot\|_\infty)) \le d\log(1 + 4Hd/(\underline{\lambda}\epsilon)) + \log(1 + 4H^2 d/(\underline{\lambda}^2\rho\epsilon)).
$$

This finishes the proof of Lemma F.4. □

Lemma F.5 (Covering number of $\mathcal{V}$: TV-distance case)**.** *The ϵ-covering number of function class $\mathcal{V}$ defined in* (F.17) *under $\|\cdot\|_\infty$-norm is bounded by*

$$
\log(\mathcal{N}(\epsilon, \mathcal{V}, \|\cdot\|_\infty)) \le d\log(1 + 4Hd/\epsilon) + \log(1 + 4H/\epsilon).
$$

Proof of Lemma F.5. Consider any two pairs of parameters $(\boldsymbol{w}, \lambda)$ and $(\widehat{\boldsymbol{w}}, \widehat{\lambda})$, and denote the functions they induce as v and $\widehat{v}$. Then we have that,

$$
\begin{aligned}
|v(s) - \widehat{v}(s)| &= \left|\left(\lambda - \max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \boldsymbol{w}\right)_+ - \left(\widehat{\lambda} - \max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}\right)_+\right| \\
&\le |\lambda - \widehat{\lambda}| + \left|\max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \boldsymbol{w} - \max_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{w}}\right| \\
&\le |\lambda - \widehat{\lambda}| + \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\boldsymbol{\phi}(s,a)\|_2 \cdot \|\boldsymbol{w} - \widehat{\boldsymbol{w}}\|_2 \\
&\le |\lambda - \widehat{\lambda}| + \sqrt{d} \cdot \|\boldsymbol{w} - \widehat{\boldsymbol{w}}\|_2
\end{aligned}
$$

Thus we conclude that to form an ϵ-cover of $\mathcal{V}$ under $\|\cdot\|_\infty$-norm, it suffices to consider the product of an $\epsilon/(2\sqrt{d})$-cover of $\{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \le H\sqrt{d}\}$ under $\|\cdot\|_2$-norm and an $\epsilon/2$-cover of the interval $[0, H]$. Therefore, we can derive that

$$
\log(\mathcal{N}(\epsilon, \mathcal{V}, \|\cdot\|_\infty)) \le d\log(1 + 4Hd/\epsilon) + \log(1 + 4H/\epsilon).
$$

This finishes the proof of Lemma F.5. □

G Analysis of Maximum Likelihood Estimator

Lemma G.1 (MLE estimator guarantee: infinite model space)**.** *The maximum likelihood estimator procedure given by* (4.1) *and* (4.2) *for $\mathcal{S}\times\mathcal{A}$-rectangular robust MDP with tuning parameter ξ given by Proposition D.1 satisfies that w.p. at least $1-\delta$,*

1. *$P_h^\star \in \widehat{\mathcal{P}}_h$ for any step $h \in [H]$.*

2. *for any step $h \in [H]$ and $P_h \in \widehat{\mathcal{P}}_h$, it holds that*

$$\mathbb{E}_{(s_h,a_h)\sim d^{\mathrm{b}}_{P^\star,h}}[\|P_h(\cdot|s_h,a_h) - P^\star_h(\cdot|s_h,a_h)\|^2_{\mathrm{TV}}]$$
$$\leq \frac{C_1 \log(C_2 H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}.$$

for some absolute constant $C_1, C_2 > 0$. Here $d^{\mathrm{b}}_{P^\star,h}$ is the state-action visitation measure induced by the behavior policy π^{b} and transition kernel $P^\star$.

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

Lemma G.2 (MLE estimator guarantee: factored model space)**.** *The maximum likelihood estimator procedure given by* (4.5) *and* (4.6) *for $\mathcal{S}\times\mathcal{A}$-rectangular robust factored MDP with tuning parameter ξ_i given by Proposition E.1 satisfies that w.p. at least $1-\delta$,*

1. *$P^\star_h \in \widehat{\mathcal{P}}_h$ for any step $h \in [H]$.*
2. *for any step $h \in [H]$, $P_h \in \widehat{\mathcal{P}}_h$, and any factor $i \in [d]$ it holds that*

$$\mathbb{E}_{(s_h[\mathrm{pa}_i],a_h)\sim d^{\mathrm{b}}_{P^\star,h}}[\|P_{h,i}(\cdot|s_h[\mathrm{pa}_i],a_h) - P^\star_{h,i}(\cdot|s_h[\mathrm{pa}_i],a_h)\|^2_{\mathrm{TV}}]$$
$$\leq \frac{C_1|\mathcal{O}|^{1+|\mathrm{pa}_i|}|\mathcal{A}|\log(C_2 n d H/\delta)}{n}.$$

for some absolute constant $C_1, C_2 > 0$. Here $d^{\mathrm{b}}_{P^\star,h}$ is the state-action visitation measure induced by the behavior policy π^{b} and transition kernel $P^\star$.

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

G.1 Proof of Lemma G.1

In this section, we establish the proof of Lemma G.1. We firstly introduce several notations. For any function $f : \mathcal{S}\times\mathcal{A}\mapsto\mathbb{R}$, we denote

$$\mathbb{E}_{\mathbb{D}_h}[f] = \frac{1}{n}\sum_{\tau=1}^{n} f(s_h^\tau, a_h^\tau).$$

Proof of Lemma G.1. We follow the proof of similar MLE guarantees in [54] and [25]. We begin with proving the first conclusion of Lemma G.1, i.e., $P^\star_h \in \widehat{\mathcal{P}}_h$ for each step $h \in [H]$. For notational simplicity, we define

$$g_h(P)(s,a) = \|P(\cdot|s,a) - P^\star_h(\cdot|s,a)\|_1^2, \quad \forall P \in \mathcal{P}_{\mathrm{M}}. \tag{G.1}$$

To prove the first conclusion, it suffices to show that

$$\mathbb{E}_{\mathbb{D}_h}[g_h(\widehat{P}_h)] \leq \xi, \quad \forall h \in [H]. \tag{G.2}$$

where $\widehat{P}_h$ is the MLE estimator given in (4.1) and the parameter ξ is given by Proposition D.1. To this end, we first invoke Lemma H.1, which gives that with probability at least $1-\delta$,

$$\mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(\widehat{P}_h)] \leq c_1\big(\zeta_h + \sqrt{\log(c_2/\delta)/n}\big)^2, \tag{G.3}$$

for some absolute constants $c_1, c_2 > 0$. Here ζ_h is a solution to the inequality $\sqrt{n}\epsilon^2 \geq c_0 G_h(\epsilon)$ w.r.t ϵ, with some carefully chosen function G_h which is specified in Lemma H.1. As proved in Lemma H.2, choosing $G_h(\epsilon) = (\epsilon - \epsilon^2/2)\sqrt{\log(\mathcal{N}_{[]}(\epsilon^4/2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))}$ and $\zeta_h = c_3\sqrt{\log(\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))/n}$ for some absolute constant $c_3 > 0$ can satisfy the inequality and the requirements on G_h. Thus we can obtain from (G.3) that, with probability at least $1-\delta$,

$$\mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(\widehat{P}_h)] \leq c_1\left(c_3\sqrt{\frac{\log(\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))}{n}} + \sqrt{\frac{\log(c_2/\delta)}{n}}\right)^2$$
$$\leq \frac{c_1'\log(c_2'\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}, \tag{G.4}$$

for some absolute constants $c_1', c_2' > 0$. Now to prove (G.2), it suffices to relate the expectation w.r.t. dataset $\mathbb{D}_h$ and the expectation w.r.t. visitation measure $d^{\mathrm{b}}_{P^\star,h}$. To bridge this gap, we invoke Lemma H.3, which is a Bernstein style concentration inequality and gives that with probability at least $1-\delta$,

$$|\mathbb{E}_{\mathbb{D}_h}[g_h(\widehat{P}_h)] - \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(\widehat{P}_h)]| \leq \frac{c_4 \log(c_5 \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}, \tag{G.5}$$

for some absolute constant $c_4 > 0$. Now combining (G.4) and (G.5), we can obtain that,

$$\begin{aligned}\mathbb{E}_{\mathbb{D}_h}[g_h(\widehat{P}_h)] &= \mathbb{E}_{\mathbb{D}_h}[g_h(\widehat{P}_h)] - \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(\widehat{P}_h)] + \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(\widehat{P}_h)] \\ &\leq \frac{c_1'' \log(c_2'' \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n},\end{aligned}$$

for some absolute constants $c_1'', c_2'' > 0$. Finally, taking a union bound over step $h \in [H]$ and rescaling δ, we obtain that, with probability at least $1-\delta/2$,

$$\mathbb{E}_{\mathbb{D}_h}[g_h(\widehat{P}_h)] \leq \frac{\widetilde{C}_1 \log(\widetilde{C}_2 H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n} = \xi, \quad \forall h \in [H], \tag{G.6}$$

for some absolute constants $\widetilde{C}_1, \widetilde{C}_2 > 0$. This finishes the proof of the first conclusion of Lemma G.1.

The following of the proof is to prove the second conclusion of Lemma G.1. With the notation of g_h, it suffices to prove that with probability at least $1-\delta/2$,

$$\sup_{h\in[H], P_h \in \widehat{\mathcal{P}}_h} \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)] \leq \frac{C_1 \log(C_2 H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n},$$

for some absolute constants $C_1, C_2 > 0$. To this end, for any step $h \in [H]$ and $P_h \in \widehat{\mathcal{P}}_h$, consider the following decomposition of $\mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)]$,

$$\mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)] = \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)] - \mathbb{E}_{\mathbb{D}_h}[g_h(P_h)] + \mathbb{E}_{\mathbb{D}_h}[g_h(P_h)]. \tag{G.7}$$

Note that the term $\mathbb{E}_{\mathbb{D}_h}[g_h(P_h)]$ in (G.7) satisfies, with probability at least $1-\delta/2$,

$$\begin{aligned}\mathbb{E}_{\mathbb{D}_h}[g_h(P_h)] &= \mathbb{E}_{\mathbb{D}_h}[\|P_h(\cdot|s,a) - P_h^\star(\cdot|s,a)\|_1^2] \\ &= \mathbb{E}_{\mathbb{D}_h}[\|P_h(\cdot|s,a) - \widehat{P}_h(\cdot|s,a) + \widehat{P}_h(\cdot|s,a) - P_h^\star(\cdot|s,a)\|_1^2] \\ &\leq 2\mathbb{E}_{\mathbb{D}_h}[\|P_h(\cdot|s,a) - \widehat{P}_h(\cdot|s,a)\|_1^2] + 2\mathbb{E}_{\mathbb{D}_h}[\|\widehat{P}_h(\cdot|s,a) - P_h^\star(\cdot|s,a)\|_1^2] \\ &\leq 4\xi,\end{aligned} \tag{G.8}$$

where the last inequality follows from the definition of confidence region $\widehat{\mathcal{P}}_h$ and the first conclusion of Lemma G.1, i.e., (G.6). Thus by taking (G.8) back into (G.7), we obtain that,

$$\mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)] \leq 4\xi + \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)] - \mathbb{E}_{\mathbb{D}_h}[g_h(P_h)]. \tag{G.9}$$

Finally, invoking another Bernstein style concentration inequality (Lemma H.4), we have that with probability at least $1-\delta$,

$$\sup_{P_h \in \widehat{\mathcal{P}}_h} |\mathbb{E}_{\mathbb{D}_h}[g_h(P_h)] - \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)]| \leq \frac{c_6 \log(c_7 \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n} \tag{G.10}$$

Thus by combining (G.9) and (G.10), taking a union bound over step $h \in [H]$, rescaling δ, and using the definition of ξ, we can conclude that with probability at least $1-\delta/2$,

$$\sup_{h\in[H], P_h \in \widehat{\mathcal{P}}_h} \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)] \leq \frac{C_1 \log(C_2 H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n},$$

for some absolute constants $C_1, C_2 > 0$. This finishes the proof of Lemma G.1. ☐

G.2 Proof of Lemma G.2

Proof of Lemma G.2. This is a direct corollary of Lemma G.1 in the finite state space case: for each factor $i \in [d]$, consider $\mathcal{O}$ as the state finite space and apply the upper bound of bracket number (4.4) for finite state space proved in Appendix D.2. This proves Lemma G.2. ☐

H Technical Lemmas

H.1 Lemmas for Maximum Likelihood Estimator

In this section, we give technical lemmas for the maximum likelihood estimator. We firstly introduce several notations which are also considered by [54] and [25], We define a localized model space $\overline{\mathcal{P}}_h(\epsilon)$ as

$$\overline{\mathcal{P}}_h(\epsilon) = \left\{ P \in \overline{\mathcal{P}}_{\mathrm{M},h} : \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[D^2_{\mathrm{Hellinger}}(P(\cdot|s,a)\|P^\star_h(\cdot|s,a))] \le \epsilon^2 \right\},$$

where $D_{\mathrm{Hellinger}}(\cdot\|\cdot)$ is the Hellinger distance between two probability measures, and $\overline{\mathcal{P}}_{\mathrm{M},h}$ is called a modified space $\mathcal{P}_{\mathrm{M}}$, defined as $\overline{\mathcal{P}}_{\mathrm{M},h} = \{(P + P^\star_h)/2 : P \in \mathcal{P}_{\mathrm{M}}\}$. Also, we define the entropy integral of $\overline{\mathcal{P}}_h(\epsilon)$ under the $\|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}}$-norm as

$$J_{\mathrm{B}}(\epsilon, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}}) = \max\left\{ \epsilon, \int_{\epsilon^2/2}^{\epsilon} \sqrt{\log(\mathcal{N}_{[]}(u, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}}))}\mathrm{d}u \right\}.$$

Lemma H.1 (MLE Gaurantee, [55])**.** *Take a function $G_h(\epsilon) : [0,1] \to \mathbb{R}$ s.t. $G_h(\epsilon) \ge J_B(\epsilon, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}})$ and $G_h(\epsilon)/\epsilon^2$ non-increasing w.r.t ϵ. Then, letting ζ_h be a solution to $\sqrt{n}\epsilon^2 \ge c_0 G_h(\epsilon)$ w.r.t ϵ, where c_0 is an absolute constant. With probability at least $1-\delta$, we have that*

$$\mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[\|\widehat{P}_h(\cdot|s,a) - P^\star_h(\cdot|s,a)\|_1^2] \le c_1\big(\zeta_h + \sqrt{\log(c_2/\delta)/n}\big)^2.$$

Proof of Lemma H.1. We refer to Theorem 7.4 in [55] for a detailed proof. □

Lemma H.2 (Choice of $G_h(\epsilon)$ and ζ_h in Lemma H.1)**.** *In Lemma H.1, we can choose $G_h(\epsilon)$ as*

$$G_h(\epsilon) = (\epsilon - \epsilon^2/2)\sqrt{\log(\mathcal{N}_{[]}(\epsilon^4/2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))},$$

In this case, $\zeta_h = c_0\sqrt{\log(\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))/n}$ solves the inequality $\sqrt{n}\epsilon^2 \ge c_0 G_h(\epsilon)$ w.r.t ϵ.

Proof of Lemma H.2. We first check the conditions that G_h should satisfy. By the choice of G_h,

$$\begin{aligned}
G_h(\epsilon) &= (\epsilon - \epsilon^2/2)\sqrt{\log(\mathcal{N}_{[]}(\epsilon^4/2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))} \\
&\ge (\epsilon - \epsilon^2/2)\sqrt{\log(\mathcal{N}_{[]}(\epsilon^2/2, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}}))} \\
&\ge \max\left\{ \epsilon, \int_{\epsilon^2/2}^{\epsilon} \sqrt{\log(\mathcal{N}_{[]}(u, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}}))}\mathrm{d}u \right\} \\
&= J_B(\epsilon, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}}),
\end{aligned}$$

where the first inequality follows from Lemma H.6, the second inequality follows from the fact that $\mathcal{N}_{[]}(u_1, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}}) \ge \mathcal{N}_{[]}(u_2, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}})$ for $u_1 \le u_2$. In the second inequality we assume without loss of generality that $\log(\mathcal{N}_{[]}(\epsilon^2/2, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}})) \ge 4$. Besides, since

$$G_h(\epsilon)/\epsilon^2 = (1/\epsilon - 1/2)\sqrt{\log(\mathcal{N}_{[]}(\epsilon^4/2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))}$$

is non-increasing w.r.t ϵ for $\epsilon \in [0,1]$, we can confirm that G_h satisfy the conditions in Lemma H.1. With this choice of G_h, the inequality $\sqrt{n}\epsilon^2 \ge c_0 G_h(\epsilon)$ reduces to

$$\sqrt{n} \ge c_0(1/\epsilon - 1/2)\sqrt{\log(\mathcal{N}_{[]}(\epsilon^4/2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))},$$

which equivalents to

$$\epsilon \ge \frac{c_0\sqrt{\log(\mathcal{N}_{[]}(\epsilon^4/2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))}}{\sqrt{n} + \frac{c_0}{2}\sqrt{\log(\mathcal{N}_{[]}(\epsilon^4/2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))}}. \tag{H.1}$$

Taking $\zeta_h = c_0\sqrt{\log(\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))/n}$, when $c_0\sqrt{\log(\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))} \geq 2^{1/4}$, we can check that ζ_h satisfies the inequality (H.1) by,

$$\zeta_h = \frac{c_0\sqrt{\log(\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))}}{\sqrt{n}} \geq \frac{c_0\sqrt{\log(\mathcal{N}_{[]}(\zeta_h^2/2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))}}{\sqrt{n} + \frac{c_0}{2}\sqrt{\log(\mathcal{N}_{[]}(\zeta_h^2/2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}))}}.$$

This finishes the proof of Lemma H.2. □

H.2 Lemmas for Concentration Inequalities and Bracket Numbers

Lemma H.3 (Bernstein inequality I). *For any step $h \in [H]$, with probability at least $1-\delta$,*

$$|\mathbb{E}_{\mathbb{D}_h}[g_h(\widehat{P}_h)] - \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(\widehat{P}_h)]| \leq \frac{c_1\log(c_2\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}.$$

Proof of Lemma H.3. Motivated by [54] and [25], to obtain a fast rate of convergence, we will utilize the localization technique in proving concentration. To this end, we first define the following localized realizable model space,

$$\mathcal{P}^{\mathrm{Loc}}_{\mathrm{M},h} = \left\{P \in \mathcal{P}_{\mathrm{M}} : \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P)] \leq \frac{c_1'\log(c_2'\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}\right\},$$

where absolute constants c_1' and c_2' are specified in (G.4). According to the proof of (G.4), we know that with probability at least $1-\delta$, the event $E_1 = \{\widehat{P}_h \in \mathcal{P}^{\mathrm{Loc}}_{\mathrm{M},h}\}$ holds. In the sequel, we will always condition on the event E_1. Now we define another function class as

$$\mathcal{F}_h = \left\{g_h(P) : P \in \mathcal{P}^{\mathrm{Loc}}_{\mathrm{M},h}\right\}.$$

Then applying Bernstein inequality with union bound (Lemma H.5) on the function class $\mathcal{F}_h$, we can obtain that with probability at least $1-\delta$, for any $P \in \mathcal{P}^{\mathrm{Loc}}_{\mathrm{M},h}$, (denote $\mathcal{M}(\epsilon) = \mathcal{N}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty)$)

$$\begin{aligned}
&|\mathbb{E}_{\mathbb{D}_h}[g_h(P)] - \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P)]| \qquad\qquad (\mathrm{H.2})\\
&\leq \sqrt{\frac{2\mathbb{V}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P)]\log(\mathcal{M}(\epsilon)/\delta)}{n}} + 8\sqrt{\frac{\epsilon\log(\mathcal{M}(\epsilon)/\delta)}{n}} + \frac{8\log(\mathcal{M}(\epsilon)/\delta)}{3n} + 2\epsilon\\
&\leq \sqrt{\frac{8\mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P)]\log(\mathcal{M}(\epsilon)/\delta)}{n}} + 8\sqrt{\frac{\epsilon\log(\mathcal{M}(\epsilon)/\delta)}{n}} + \frac{8\log(\mathcal{M}(\epsilon)/\delta)}{3n} + 2\epsilon\\
&\leq \frac{\sqrt{8c_1'\log(c_2'\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)\cdot\log(\mathcal{M}(\epsilon)/\delta)}}{n} + 8\sqrt{\frac{\epsilon\log(\mathcal{M}(\epsilon)/\delta)}{n}}\\
&\quad + \frac{8\log(\mathcal{M}(\epsilon)/\delta)}{3n} + 2\epsilon,
\end{aligned}$$

where the first inequality follows from Lemma H.5, both the first and the second inequality use the fact that $\sup_{P\in\mathcal{P}^{\mathrm{Loc}}_{\mathrm{M},h}}|g_h(P)| \leq 4$, and the last inequality uses the definition of $\mathcal{P}^{\mathrm{Loc}}_{\mathrm{M},h}$. If we denote

$$\mathcal{F}_h' = \{g_h(P) : P \in \mathcal{P}_{\mathrm{M}}\}, \tag{H.3}$$

we can upper bound the covering number $\mathcal{M}(\epsilon)$ via the following sequence of inequalities,

$$\mathcal{M}(\epsilon) = \mathcal{N}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq \mathcal{N}(\epsilon, \mathcal{F}_h', \|\cdot\|_\infty) \leq \mathcal{N}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}) \leq \mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}), \tag{H.4}$$

where the first inequality follows from $\mathcal{F}_h \subseteq \mathcal{F}_h'$, the second inequality can be easily derived from the relationship between $\mathcal{F}_h'$ and $\mathcal{P}_{\mathrm{M}}$, and the last inequality follows from the fact that covering number can be bounded by bracket number. Therefore, by combining (H.2) and (H.4), letting $\epsilon = 1/n^2$, we can derive that, conditioning on $E_1 = \{\widehat{P}_h \in \mathcal{P}^{\mathrm{Loc}}_{\mathrm{M},h}\}$, with probability at least $1-\delta$,

$$|\mathbb{E}_{\mathbb{D}_h}[g_h(\widehat{P}_h)] - \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(\widehat{P}_h)]| \leq \frac{c_1\log(c_2\mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n},$$

for some absolute constant $c_1, c_2 > 0$. Finally, since the event E_1 holds with probability at least $1-\delta$, by rescaling δ, we can finish the proof. □

Lemma H.4 (Bernstein inequality II)**.** *For any step $h \in [H]$, with probability at least $1-\delta$,*

$$|\mathbb{E}_{\mathbb{D}_h}[g_h(P_h)] - \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)]| \le \frac{c_1 \log(c_2 \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}, \quad \forall P_h \in \widehat{\mathcal{P}}_h.$$

Proof of Lemma H.4. According to the proof of (G.8), we know that the event E_2 defined as

$$E_2 = \left\{ \mathbb{E}_{\mathbb{D}_h}[g_h(P_h)] \le 4\xi, \ \forall P_h \in \widehat{\mathcal{P}}_h \right\}$$

holds with probability at least $1-\delta/2$. In the sequel, we always condition on the event E_2. Now we define a function class $\mathcal{G}_h$ as following,

$$\mathcal{G}_h = \left\{ g_h(P_h) : P_h \in \widehat{\mathcal{P}}_h \right\}.$$

Applying Bernstein inequality with union bound (Lemma H.5) on the function class $\mathcal{G}_h$, we can obtain that with probability at least $1-\delta$, for any $P_h \in \widehat{\mathcal{P}}_h$, (denote $\mathcal{M}'(\epsilon) = \mathcal{N}(\epsilon, \mathcal{G}_h, \|\cdot\|_\infty)$)

$$\begin{aligned}
&|\mathbb{E}_{\mathbb{D}_h}[g_h(P_h)] - \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)]| \\
&\quad \le \sqrt{\frac{2\mathbb{V}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)] \log(\mathcal{M}'(\epsilon)/\delta)}{n}} + 8\sqrt{\frac{\epsilon \log(\mathcal{M}'(\epsilon)/\delta)}{n}} + \frac{8\log(\mathcal{M}'(\epsilon)/\delta)}{3n} + 2\epsilon \\
&\quad \le \sqrt{\frac{8\mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)] \log(\mathcal{M}'(\epsilon)/\delta)}{n}} + 8\sqrt{\frac{\epsilon \log(\mathcal{M}'(\epsilon)/\delta)}{n}} + \frac{8\log(\mathcal{M}'(\epsilon)/\delta)}{3n} + 2\epsilon \\
&\quad \le \sqrt{\frac{8(|\mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)] - \mathbb{E}_{\mathbb{D}_h}[g_h(P_h)]| + 4\xi) \log(\mathcal{M}'(\epsilon)/\delta)}{n}} \\
&\qquad + 8\sqrt{\frac{\epsilon \log(\mathcal{M}'(\epsilon)/\delta)}{n}} + \frac{8\log(\mathcal{M}'(\epsilon)/\delta)}{3n} + 2\epsilon, \qquad \text{(H.5)}
\end{aligned}$$

where the first inequality follows from Lemma H.5, both the first and the second inequality use the fact that $\sup_{P_h \in \widehat{\mathcal{P}}_h} |g_h(P_h)| \le 4$, and the last inequality uses the definition of event E_2. By using the fact that the function class $\mathcal{G}_h \subseteq \mathcal{F}'_h$ where $\mathcal{F}'_h$ is defined in (H.3) in the proof of Lemma H.3, we can apply the same argument as (H.4) to derive that $\mathcal{M}'(\epsilon) \le \mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})$. Thus taking $\epsilon = 1/n^2$, denoting $\Delta_h(P_h) = |\mathbb{E}_{\mathbb{D}_h}[g_h(P_h)] - \mathbb{E}_{d^{\mathrm{b}}_{P^\star,h}}[g_h(P_h)]|$, we can derive from (H.5) that,

$$\begin{aligned}
\Delta_h(P_h) &\le \sqrt{\frac{8(\Delta_h(P_h) + 4\xi)\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}} \\
&\qquad + 8\sqrt{\frac{\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n^3}} + \frac{8\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{3n} + \frac{2}{n^2} \\
&\le \sqrt{\frac{8(\Delta_h(P_h) + 4\xi)\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}} + \frac{c_1' \log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n} \\
&\le \sqrt{\frac{8\Delta_h(P_h)\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}} + \frac{c_1'' \log(c_2''\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}, \qquad \text{(H.6)}
\end{aligned}$$

for some absolute constants $c_1', c_1'', c_2'' > 0$, where in the last inequality we have applied the definition of ξ. Now solving this quadratic inequality (H.6) w.r.t $\Delta_h(P_h)$, we can obtain that,

$$\Delta_h(P_h) \le \frac{c_1 \log(c_2 \mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n},$$

for some absolute constants $c_1, c_2 > 0$. Thus we obtain that when conditioning on the event E_2, with probability at least $1-\delta$, for any $P_h \in \widehat{\mathcal{P}}_h$, the desired concentration inequality holds. Finally, since E_2 holds with probability at least $1-\delta/2$, by rescaling δ, we can finish the proof of Lemma H.4. ☐

Lemma H.5 (Bernstein inequality with union bound)**.** *Consider a function class $\mathcal{F} \subset \{f : \mathcal{X} \mapsto \mathbb{R}\}$, where $\mathcal{X}$ is a probability space. If we assume that the ϵ-covering number of $\mathcal{F}$ under infinity-norm is finite, that is, $M = \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) < \infty$, and we also assume that there exists an absolute constant*

R such that $|f(X)| \leq R$, then with probability at least $1-\delta$ the following inequality holds for all $f \in \mathcal{F}$,

$$\left|\frac{1}{n}\sum_{\tau=1}^{n} f\left(X_\tau\right) - \mathbb{E}[f(X)]\right| \leq 2\epsilon + \sqrt{\frac{2\mathbb{V}[f(X)]\log(M/\delta)}{n}} + 4\sqrt{\frac{R\epsilon\log(M/\delta)}{n}} + \frac{2R\log(M/\delta)}{3n},$$

where $X, X_1, \ldots, X_n$ are i.i.d. samples on the probability space $\mathcal{X}$.

Proof of Lemma H.5. We refer to Lemma F.1 in [25] for a detailed proof. ☐

Lemma H.6 (Bracket number I)**.** *It holds for any $\epsilon \geq 0$ that*

$$\mathcal{N}_{[]}(\epsilon, \overline{P}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}}) \leq \mathcal{N}_{[]}(2\epsilon^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}).$$

Proof of Lemma H.6. We refer to Lemma G.2 in [25] for a detailed proof. ☐

H.3 Lemmas for Dual Variables

Lemma H.7 (Dual variable for KL-divergence)**.** *The optimal solution to the following optimization problem*

$$\lambda^\star = \operatorname*{argsup}_{\lambda \in \mathbb{R}_+} \left\{ -\lambda \log\left(\int \exp\left\{-f(x)/\lambda\right\} P(\mathrm{d}x) \right) - \lambda\sigma \right\},$$

with $\|f\|_\infty \leq H$ and some probability measure P satisfies that $\lambda^\star \leq H/\sigma$.

Proof of Lemma H.7. For simplicity, denote by $g(\lambda) = -\lambda \log\left(\int \exp\left\{-f(x)/\lambda\right\} P(\mathrm{d}x)\right) - \lambda\sigma$. Notice that $g(0) = 0$, and for $\lambda > H/\sigma$, due to $\|f\|_\infty \leq H$, we have that

$$g(\lambda) < -\lambda\log(\exp\{-H/(H/\sigma)\}) - \lambda\sigma = \lambda\sigma - \lambda\sigma = 0.$$

Thus we can conclude that $\lambda^\star \leq H/\sigma$. ☐

Lemma H.8 (Dual variable for TV-distance)**.** *The optimal solution to the following optimization problem*

$$\lambda^\star = \operatorname*{argsup}_{\lambda \in \mathbb{R}} \left\{ -\int (\lambda - f(x))_+ P(\mathrm{d}x) - \frac{\sigma}{2}(\lambda - \inf_x f(x))_+ + \lambda \right\}.$$

with $\|f\|_\infty \leq H$ and some probability measure P satisfies that $0 \leq \lambda^\star \leq H$.

Proof of Lemma H.8. For simplicity, denote $g(\lambda) = -\int(\lambda - f(x))_+ P(\mathrm{d}x) - \frac{\sigma}{2}(\lambda - \inf_x f(x))_+ + \lambda$. We can observe that $g(0) = 0$, and $g(\lambda) \leq 0$ for $\lambda \leq 0$. Thus we have shown that $\lambda^\star \geq 0$. Also, for $\lambda \geq H$, due to $\|f\|_\infty \leq H$, we can write $g(\lambda)$ as

$$\begin{aligned} g(\lambda) &= -\int \lambda - f(x) P(\mathrm{d}x) - \frac{\sigma}{2}(\lambda - \inf_x f(x)) + \lambda \\ &= \int f(x) P(\mathrm{d}x) + \frac{\sigma}{2}\inf_x f(x) - \frac{\sigma}{2}\lambda, \end{aligned}$$

which is a monotonically decreasing function with respect to λ. Thus we prove that $\lambda^\star \leq H$. ☐