# Double Pessimism is Provably Efficient for Distributionally Robust Offline Reinforcement Learning: Generic Algorithm and Robust Partial Coverage

**Jose Blanchet**[1]* **Miao Lu**[1]* **Tong Zhang**[2]* **Han Zhong**[3]*
[1] Department of Management Science and Engineering, Stanford University
[2] Department of Mathematics, The Hong Kong University of Science and Technology
[3] Center for Data Science, Peking University

## Abstract

We study distributionally robust offline reinforcement learning (RL), which seeks to find an optimal robust policy purely from an offline dataset that can perform well in perturbed environments. We propose a generic algorithm framework Doubly Pessimistic Model-based Policy Optimization ($\texttt{P}^2\texttt{MPO}$) for robust offline RL, which features a novel combination of a flexible model estimation subroutine and a doubly pessimistic policy optimization step. Here the *double pessimism* principle is crucial to overcome the distribution shift incurred by i) the mismatch between behavior policy and the family of target policies; and ii) the perturbation of the nominal model. Under certain accuracy assumptions on the model estimation subroutine, we show that $\texttt{P}^2\texttt{MPO}$ is provably sample-efficient with *robust partial coverage data*, which means that the offline dataset has good coverage of the distributions induced by the optimal robust policy and perturbed models around the nominal model. By tailoring specific model estimation subroutines for concrete examples including tabular Robust Markov Decision Process (RMDP), factored RMDP, and RMDP with kernel and neural function approximations, we show that $\texttt{P}^2\texttt{MPO}$ enjoys a $\widetilde{\mathcal{O}}(n^{-1/2})$ convergence rate, where $n$ is the number of trajectories in the offline dataset. Notably, these models, except for the tabular case, are first identified and proven tractable by this paper. To the best of our knowledge, we first propose a general learning principle — double pessimism — for robust offline RL and show that it is provably efficient in the context of general function approximations.

## 1 Introduction

Reinforcement learning (RL) [52] aims to learn an optimal policy that maximizes the cumulative rewards received in an unknown environment. Typically, deep RL algorithms learn a policy in an online trial-and-error fashion using millions to billions of data. However, data collection could be costly and risky in some practical applications such as healthcare [56] and autonomous driving [38]. To tackle this challenge, offline RL (also known as batch RL) [21, 22] learns a near-optimal policy based on a dataset collected a priori without further interactions with the environment. Although there has been great progress in offline RL [72, 20, 16, 54, 62, 6], these works implicitly require that the offline dataset is generated by the real-world environment, which may fail in practice. Taking robotics [18, 37] as an example, the experimenter trains agents in a simulated physical environment and then deploy them in real-world environments. Since the experimenter does not have access to the true physical environment, there is a mismatch between the simulated environment used to generate the offline dataset and the real-world environment used to deploy the agents. Such a mismatch is

*Alphabetical order. Email to miaolu@stanford.edu

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

commonly referred to as the *sim-to-real gap* [42, 76]. Since the optimal policy is sensitive to the model [31, 9], the potential sim-to-real gap may lead to the poor performance of RL algorithms.

A promising solution to remedy this issue is robust RL [13, 9, 32] – training a robust policy that performs well in a bad or even adversarial environment. A line of work on deep robust RL [43, 44, 41, 30, 53, 75, 19] demonstrates the superiority of the trained robust policy in real world environments. Furthermore, the recent work of Hu et al. [11] theoretically proves that the ideal robust policy can attain near optimality in dealing with problems with sim-to-real gap, but this work does not suggest how to learn a robust policy from a theoretical perspective. In order to understand robust RL from the theoretical side, robust Markov decision process (RMDP) [13, 9] has been proposed, and many recent works [77, 47, 29] design sample-efficient learning algorithms for robust offline RL. These works mainly focus on the tabular case, which is not capable of tackling large state spaces. Meanwhile, in the non-robust setting, a line of works [16, 54, 62, 73, 45] show that "pessimism" is the general learning principle for designing algorithms that can overcome the distributional shift problem faced by offline RL. In particular, in the context of function approximation, Xie et al. [62] and Uehara and Sun [54] leverage the pessimism principle and propose generic algorithms in the model-free and model-based fashion, respectively. Hence, it is natural to ask the following questions:

**Q1:** What is the general learning principle for robust offline RL?
**Q2:** Based on this learning principle, can we design a generic algorithm for robust offline RL in the context of function approximation?

To answer these two questions, we need to tackle the following two intertwined challenges: (i) distributional shift, that is, the mismatch between offline data distribution and the distribution induced by the optimal robust policy. In robust offline RL, the distributional shift has two sources – behavior policy and perturbed model, where the latter is the unique challenge not presented in non-robust RL; and (ii) function approximation. Existing works mainly focus on the tabular case, and it remains elusive how to add reasonable structure conditions to make RMDPs with large state spaces tractable. Despite these challenges, we answer the aforementioned two questions affirmatively.

**Contributions.** We study robust offline RL in a general framework, which not only includes existing known tractable $\mathcal{S} \times \mathcal{A}$-rectangular tabular RMDPs, but also subsumes several newly proposed models (e.g., $\mathcal{S} \times \mathcal{A}$-rectangular factored RMDPs, $\mathcal{S} \times \mathcal{A}$-rectangular kernel RMDPs, and $\mathcal{S} \times \mathcal{A}$-rectangular neural RMDPs) as special cases. Under this general framework, we propose a generic model-based algorithm, dubbed as Doubly Pessimistic Model-based Policy Optimization ($\texttt{P}^2\texttt{MPO}$), which consists of a model estimation subroutine and a policy optimization step based on *doubly pessimistic* value estimators. We note that the model estimation subroutine can be flexibly chosen according to structural conditions of specific RMDP examples. Meanwhile, the adoption of doubly pessimistic value estimators in the face of model estimation uncertainty and environment uncertainty plays a key role in overcoming the distributional shift problem in robust offline RL.

From the theoretical perspective, we characterize the optimality of $\texttt{P}^2\texttt{MPO}$ with partial coverage. In particular, we show that the suboptimality gap of $\texttt{P}^2\texttt{MPO}$ is upper bounded by the model estimation error (see Condition 3.2) and the robust partial coverage coefficient (see Assumption 3.3). For concrete examples of RMDPs, by customizing specific model estimation mechanisms and plugging them into $\texttt{P}^2\texttt{MPO}$, we show that $\texttt{P}^2\texttt{MPO}$ enjoys a $n^{-1/2}$ convergence rate with robust partial coverage data, where $n$ is the number of trajectories in the offline dataset. In summary, we identify a general learning principle — *double pessimism* — for robust offline RL. Based on this principle, we can perform sample-efficient robust offline RL with robust partial coverage data via general function approximation. See Table 1 for a summary of our results and a comparison with mostly related works.

### 1.1 Related Works

**Robust reinforcement learning in robust Markov decision processes.** Robust RL is usually modeled as a robust MDP (RMDP) [13, 9], and its planning has been well studied [13, 9, 65, 60, 57]. Recently, robust RL in RMDPs has attracted considerable attention, and a growing body of works studies this problem in the generative model [68, 39, 49, 58, 69, 66, 7], online setting [59, 3, 8], and offline setting [77, 40, 47, 29]. Our work focuses on robust offline RL, and we provide a more in-depth comparison with Zhou et al. [77], Shi and Chi [47], Ma et al. [29] as follows. Under the full coverage condition (a uniformly lower bounded data distribution), Zhou et al. [77] provide the first sample-efficient algorithm for $\mathcal{S} \times \mathcal{A}$-rectangular tabular RMDPs. After, Shi and Chi [47]

Table 1: A comparison with closely related works on robust offline RL. ✓ means the work can tackle this model with robust partial coverage data, ✓! means the work requires full coverage data to solve the model, and ✗ means the work cannot tackle the model. Lightblue color denotes the models that are first proposed and proved tractable in this work.

| | Zhou et al. [77] | Shi and Chi [47] | Ma et al. [29] | This Work |
|---|---|---|---|---|
| $\mathcal{S} \times \mathcal{A}$-rectangular tabular RMDP | ✓! | ✓ | ✗ | ✓ |
| $d$-rectangular linear RMDP | ✗ | ✗ | ✓ | ✓ |
| $\mathcal{S} \times \mathcal{A}$-rectangular factored RMDP | ✗ | ✗ | ✗ | ✓ |
| $\mathcal{S} \times \mathcal{A}$-rectangular kernel RMDP | ✗ | ✗ | ✗ | ✓ |
| $\mathcal{S} \times \mathcal{A}$-rectangular neural RMDP | ✗ | ✗ | ✗ | ✓ |

leverage the pessimism principle and design a sample-efficient offline algorithm that only requires robust partial coverage data for $\mathcal{S} \times \mathcal{A}$-rectangular tabular RMDPs. Ma et al. [29] propose a new $d$-rectangular RMDP and develop a pessimistic style algorithm that can find a near-optimal robust policy with partial coverage data. In comparison, we provide a generic algorithm that can not only solve the models in Zhou et al. [77], Shi and Chi [47], Ma et al. [29], but can also tackle various newly proposed RMDP models such as $\mathcal{S} \times \mathcal{A}$-rectangular factored RMDP, $\mathcal{S} \times \mathcal{A}$-rectangular kernel RMDP, and $\mathcal{S} \times \mathcal{A}$-rectangular neural RMDP. See Table 1 for a summary. Moreover, we propose a new pessimistic type learning principle "double pessimism" for robust offline RL. Although Shi et al. [48] and Ma et al. [29] adopt the similar algorithmic idea in tabular or linear settings, neither of them have identified a general learning principle for robust offline RL in the regime of large state space.

**Non-robust offline RL and pessimism principle.** The line of works on offline RL aims to design efficient learning algorithms that find an optimal policy given an offline dataset collected a priori. Prior works [33, 2, 5] typically require a dataset of full coverage, which assumes that the offline data have good coverage of all state-action pairs. In order to avoid such a strong coverage condition on data, the *pessimism* principle – being conservative in policy or value estimation of those state-action pairs that are not sufficiently covered by data – has been proposed. Based on this principle, a long line of works [see e.g., 16, 54, 62, 63, 45, 73, 71, 64, 48, 24, 26, 74, 28, 46] propose algorithms that can learn the optimal policy only with the *partial coverage data*. The partial coverage data only requires to cover the state-action pairs visited by the optimal policy. Among these works, our work is mostly related to the work of Uehara and Sun [54], which proposes a generic model-based algorithm for non-robust offline RL. Our algorithm for robust offline RL is also in a model-based fashion, and our study covers some models such as $\mathcal{S} \times \mathcal{A}$-rectangular kernel and neural RMDPs whose non-robust counterparts are not studied by Uehara and Sun [54]. More importantly, our algorithm is based on a newly proposed *double pessimism* principle, which is tailored for robust offline RL and is in parallel with the pessimism principle used in non-robust offline RL. Also, we show that the performance of our proposed algorithm depends on the notion of *robust partial coverage coefficient*, which is also different from the notions of partial coverage coefficient in previous non-robust offline RL works [16, 62, 54].

### 1.2 Notations

For any set $A$, we use $2^A$ to denote the collection of all the subsets of $A$. For any measurable space $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the collection of probability measures over $\mathcal{X}$. For any integer $n$, we use $[n]$ to denote the set $\{1, \cdots, n\}$. Throughout the paper, we use $D(\cdot\|\cdot)$ to denote a (pseudo-)distance between two probability measures (or densities). In specific, we define the KL-divergence $D_{\mathrm{KL}}(p\|q)$ between two probability densities $p$ and $q$ over $\mathcal{X}$ as

$$D_{\mathrm{KL}}(p\|q) = \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x,$$

and we define the TV-distance $D_{\mathrm{TV}}(p\|q)$ between two probability densities $p$ and $q$ over $\mathcal{X}$ as

$$D_{\mathrm{TV}}(p\|q) = \frac{1}{2} \int_{\mathcal{X}} |q(x) - p(x)| \, \mathrm{d}x.$$

Given a function class $\mathcal{F}$ equipped with some norm $\|\cdot\|_{\mathcal{F}}$, we denote by $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{\mathcal{F}})$ the $\epsilon$-bracket number of $\mathcal{F}$, and $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{\mathcal{F}})$ the $\epsilon$-covering number of $\mathcal{F}$.

## 2 Preliminaries

### 2.1 A Unified Framework of Robust Markov Decision Processes

We introduce a unified framework for studying episodic robust Markov decision processes (RMDP), which we denote as a tuple $(\mathcal{S}, \mathcal{A}, H, P^\star, R, \mathcal{P}_{\mathrm{M}}, \mathbf{\Phi})$. The set $\mathcal{S}$ is the state space with possibly infinite cardinality, $\mathcal{A}$ is the action space with finite cardinality. The integer $H$ is the length of each episode. The set $P^\star = \{P_h^\star\}_{h=1}^H$ is the collection of transition kernels where each $P_h^\star : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$, and $R = \{R_h\}_{h=1}^H$ is the collection of reward functions where each $R_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$. We use $\Delta(\mathcal{S})$ to note the probability simplex on $\mathcal{S}$ (i.e. the space of probability measures with support on $\mathcal{S}$).

We consider a model-based perspective of reinforcement learning, and we denote $\mathcal{P} = \{P(\cdot|\cdot, \cdot) : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})\}$ as the space of all transition kernels. The space $\mathcal{P}_{\mathrm{M}} \subseteq \mathcal{P}$ of the RMDP is a realizable model space which contains the transition kernel $P^\star$, i.e., $P_h^\star \in \mathcal{P}_{\mathrm{M}}$ for any step $h \in [H]$. Finally, the RMDP is equipped with a mapping $\mathbf{\Phi} : \mathcal{P}_{\mathrm{M}} \mapsto 2^{\mathcal{P}}$ that characterizes the *robust set* of any transition kernel in $\mathcal{P}_{\mathrm{M}}$. Formally, for any transition kernel $P \in \mathcal{P}_{\mathrm{M}}$, we call $\mathbf{\Phi}(P)$ the *robust set* of $P$. One can interpret the transition kernel $P_h^\star \in \mathcal{P}_{\mathrm{M}}$ as the transition kernel of the training environment, while $\mathbf{\Phi}(P_h^\star)$ contains all the possible transition kernels of the test environment.

**Remark 2.1.** *The mapping $\mathbf{\Phi}$ is defined on the realizable model space $\mathcal{P}_{\mathrm{M}}$, while for generality we allow the image of $\mathbf{\Phi}$ to be outside of $\mathcal{P}_{\mathrm{M}}$. That is, a $\widetilde{P} \in \mathbf{\Phi}(P)$ for some $P \in \mathcal{P}_{\mathrm{M}}$ might be in $\mathcal{P}_{\mathrm{M}}^c$.*

**Policy and robust value function.** Given an RMDP $(\mathcal{S}, \mathcal{A}, H, P^\star, R, \mathcal{P}_{\mathrm{M}}, \mathbf{\Phi})$, we consider using a Markovian policy to make decision. A Markovian policy $\pi$ is defined as $\pi = \{\pi_h\}_{h=1}^H$ with $\pi_h : \mathcal{A} \mapsto \Delta(\mathcal{S})$ for each step $h \in [H]$. For simplicity, we use *policy* to refer to a Markovian policy.

Given any policy $\pi$, we define the *robust value function* of $\pi$ with respect to any set of transition kernels $P = \{P_h\}_{h=1}^H \subseteq \mathcal{P}_{\mathrm{M}}$ as the following, for each step $h \in [H]$,

$$V_{h,P,\mathbf{\Phi}}^\pi(s) = \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h), 1 \le h \le H} V_h^\pi(s; \{\widetilde{P}_h\}_{h=1}^H), \quad \forall s \in \mathcal{S}, \tag{2.1}$$

$$Q_{h,P,\mathbf{\Phi}}^\pi(s, a) = \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h), 1 \le h \le H} Q_h^\pi(s, a; \{\widetilde{P}_h\}_{h=1}^H), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{2.2}$$

Here $V_h^\pi(\cdot; \{\widetilde{P}_h\}_{h=1}^H)$ and $Q_h^\pi(\cdot; \{\widetilde{P}_h\}_{h=1}^H)$ are the *state-value function* and the *action-value function* [52] of policy $\pi$ in the standard episodic MDP $(\mathcal{S}, \mathcal{A}, H, \{\widetilde{P}_h\}_{h=1}^H, R)$,

$$V_h^\pi(s; \{\widetilde{P}_h\}_{h=1}^H) = \mathbb{E}_{\{\widetilde{P}_h\}_{h=1}^H, \pi}\left[\sum_{i=h}^H R_i(s_i, a_i) \middle| s_h = s\right], \quad \forall s \in \mathcal{S}, \tag{2.3}$$

$$Q_h^\pi(s, a; \{\widetilde{P}_h\}_{h=1}^H) = \mathbb{E}_{\{\widetilde{P}_h\}_{h=1}^H, \pi}\left[\sum_{i=h}^H R_i(s_i, a_i) \middle| s_h = s, a_h = a\right], \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, \tag{2.4}$$

where the expectation $\mathbb{E}_{\{\widetilde{P}_h\}_{h=1}^H, \pi}[\cdot]$ is taken with respect to the trajectories induced by the transition kernel $\{\widetilde{P}_h\}_{h=1}^H$ and the policy $\pi$. Intuitively, the robust value function of a policy $\pi$ given transition kernel $P$ is defined as the least expected cumulative reward achieved by $\pi$ when the transition kernel varies in the robust set of $P$. This is how an RMDP takes the perturbed models into consideration.

**$\mathcal{S} \times \mathcal{A}$-rectangular robust set and robust Bellman equation.** Ideally, we would like to consider robust value function that has recursive expressions, just like the Bellman equation satisfied by (2.3) in a standard MDP [52]. To this end, we focus on a generally adopted kind of robust set in our unified framework, which is called the *$\mathcal{S} \times \mathcal{A}$-rectangular robust set* [13].

**Assumption 2.2** ($\mathcal{S} \times \mathcal{A}$-rectangular robust set)**.** *We assume that the mapping $\mathbf{\Phi}$ induces $\mathcal{S} \times \mathcal{A}$-rectangular robust sets. Specifically, the mapping $\mathbf{\Phi}$ satisfies, for $\forall P \in \mathcal{P}_{\mathrm{M}}$,*

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

*for some (pseudo-) distance $D(\cdot\|\cdot)$ on $\Delta(\mathcal{S})$ and some $\rho \in \mathbb{R}_+$. Intuitively, $\mathcal{S} \times \mathcal{A}$-rectangular requires that $\mathbf{\Phi}(P)$ gives decoupled robust sets for $P(\cdot|s, a)$ across different $(s, a)$-pairs. The (pseudo-)distance $D(\cdot\|\cdot)$ can be chosen as a $\phi$-divergence [68, 10] or a $p$-Wasserstein-distance [35].*

Thanks to the $\mathcal{S} \times \mathcal{A}$-rectangular assumption on the mapping $\mathbf{\Phi}$, the robust value functions (2.1) of any policy $\pi$ then satisfy a recursive expression, which is called robust Bellman equation [13, 36].

**Proposition 2.3** (Robust Bellman equation)**.** *Under Assumption 2.2, for any $P = \{P_h\}_{h=1}^H$ where $P_h \in \mathcal{P}_{\mathrm{M}}$ and any $\pi = \{\pi_h\}_{h=1}^H$ with $\pi_h : \mathcal{S} \mapsto \Delta(\mathcal{A})$, the following robust Bellman equation holds,*

$$V_{h,P,\mathbf{\Phi}}^{\pi}(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}[Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a)], \quad \forall s \in \mathcal{S}, \tag{2.5}$$

$$Q_{h,P,\mathbf{\Phi}}^{\pi}(s,a) = R_h(s,a) + \inf_{\widetilde{P}_h \in \mathbf{\Phi}(P_h)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s,a)}[V_{h+1,P,\mathbf{\Phi}}^{\pi}(s')], \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{2.6}$$

To be self-contained, in Appendix B we provide a detailed proof of the robust Bellman equation in our framework under Assumption 2.2. Equation (2.5) actually says that the infimum over all the transition kernels (recall the definition of $V_{h,P,\mathbf{\Phi}}^{\pi}$ in (2.1)) can be decomposed into a "one-step" infimum over the current transition kernel, i.e., $\inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}$, and an infimum over the future transition kernels, i.e., $V_{h+1,P,\mathbf{\Phi}}^{\pi}$. Such a property is crucial to the algorithmic design and theoretical analysis for RMDPs.

### 2.2 Examples of Robust Markov Decision Processes

In this section, we give concrete examples for the general RMDP framework proposed in Section 2.1. Most existing works on RMDPs hinge on the finiteness assumption on the state space, which fails to deal with prohibitively large or even infinite state space. In our framework, RMDPs can be considered in the paradigm of infinite state space, for which we adopt various powerful function approximation tools including kernel and neural functions. Also, we introduce a new type of RMDP named robust factored MDP, which is a robust extension of standard factored MDPs [17].

**Remark 2.4.** *Besides $\mathcal{S} \times \mathcal{A}$-rectangular-type robust sets (Assumption 2.2), our unified framework of RMDP can also cover other types of robust sets considered in some previous works as special cases, including $\mathcal{S}$-rectangular robust set [61] and $d$-rectangular robust set for linear MDPs [29]. See Section A for a discussion about these two types of robust sets.*

In the sequel, we introduce concrete examples of our framework of RMDP.

**Example 2.5** ($\mathcal{S} \times \mathcal{A}$-rectangular robust tabular MDP)**.** *When the state space $\mathcal{S}$ is a finite set, we call the corresponding model an $\mathcal{S} \times \mathcal{A}$-rectangular robust tabular MDP. Recently, there is a line of works on the $\mathcal{S} \times \mathcal{A}$-rectangular robust tabular MDP [77, 68, 39, 27, 47, 40, 8, 10, 35, 58, 69, 66, 7]. For $\mathcal{S} \times \mathcal{A}$-rectangular robust tabular MDPs, we choose $\mathcal{P}_{\mathrm{M}} = \mathcal{P}$ containing all possible transitions.*

**Remark 2.6.** *We point out that our framework of RMDP under $\mathcal{S} \times \mathcal{A}$-rectangular assumption covers substantially more model than $\mathcal{S} \times \mathcal{A}$-rectangular robust tabular MDP since our state space $\mathcal{S}$ can be infinite. The model space $\mathcal{P}_{\mathrm{M}}$ can be adapted to function approximation methods to handle the infinite state space. Thus any efficient algorithm developed for our framework of RMDPs **can not** be covered by algorithms for $\mathcal{S} \times \mathcal{A}$-rectangular robust tabular MDPs. Example 2.7 and 2.8 are infinite state space $\mathcal{S} \times \mathcal{A}$-rectangular robust MDPs with function approximations.*

**Example 2.7** ($\mathcal{S} \times \mathcal{A}$-rectangular robust MDP with kernel function approximations)**.** *We consider an infinite state space $\mathcal{S} \times \mathcal{A}$-rectangular robust MDP whose realizable model space $\mathcal{P}_{\mathrm{M}}$ is in a reproduced kernel Hilbert space (RKHS). Let $\mathcal{H}$ be a RKHS associated with a positive definite kernel $\mathcal{K} : (\mathcal{S}\times\mathcal{A}\times\mathcal{S}) \times (\mathcal{S}\times\mathcal{A}\times\mathcal{S}) \mapsto \mathbb{R}_+$ (See Appendix D.3.1 for a review of the basics of RKHS). We denote the feature mapping of $\mathcal{H}$ by $\boldsymbol{\psi} : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \mapsto \mathcal{H}$. With $\mathcal{H}$, an $\mathcal{S} \times \mathcal{A}$-rectangular robust MDP with kernel function approximation is defined as an $\mathcal{S} \times \mathcal{A}$-rectangular robust MDP with*

$$\mathcal{P}_{\mathrm{M}} = \big\{P(s'|s,a) = \langle \boldsymbol{\psi}(s,a,s'), \boldsymbol{f}\rangle_{\mathcal{H}} : \boldsymbol{f} \in \mathcal{H}, \|f\|_{\mathcal{H}} \le B_{\mathrm{K}}\big\}, \tag{2.7}$$

*for some $B_{\mathrm{K}} > 0$. Here we implicitly identify $P(\cdot|\cdot,\cdot)$ as the density of the corresponding distribution with respect to a proper base measure on $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$.*

**Example 2.8** ($\mathcal{S} \times \mathcal{A}$-rectangular robust MDP with neural function approximations)**.** *We consider an infinite state space $\mathcal{S}\times\mathcal{A}$-rectangular robust MDP whose realizable model space $\mathcal{P}_{\mathrm{M}}$ is parameterized by an overparameterized neural network. We first define a two-layer fully-connected neural network on some $\mathcal{X} \subseteq \mathbb{R}^{d_{\mathcal{X}}}$ as*

$$\mathrm{NN}(\mathbf{x}; \mathbf{W}, \mathbf{a}) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(\mathbf{x}^\top \mathbf{w}_j), \quad \forall \mathbf{x} \in \mathcal{X}, \tag{2.8}$$

*where $m \in \mathbb{N}_+$ is the number of hidden units of* NN, *$(\mathbf{W}, \mathbf{a})$ is the parameters given by $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_m) \in \mathbb{R}^{d \times m}$, $\mathbf{a} = (a_1, \cdots, a_m)^\top \in \mathbb{R}^m$, and $\sigma(\cdot)$ is the activation function. Now we assume that the state space $\mathcal{S} \subseteq \mathbb{R}^{d_\mathcal{S}}$ for some $d_\mathcal{S} \in \mathbb{N}_+$. Also, we identify actions via one-hot vectors in $\mathbb{R}^{|\mathcal{A}|}$, i.e., we represent $a \in \mathcal{A}$ by $(0, \cdots, 0, 1, 0, \cdots, 0)$ with 1 in the $a$-th coordinate. Let $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ with $d_\mathcal{X} = 2d_\mathcal{S} + |\mathcal{A}|$. Then an $\mathcal{S} \times \mathcal{A}$-rectangular robust MDP with neural function approximation is defined as an $\mathcal{S} \times \mathcal{A}$-rectangular robust MDP with $\mathcal{P}_\mathrm{M}$ given by*

$$\mathcal{P}_\mathrm{M} = \left\{ P(s'|s,a) = \mathrm{NN}((s,a,s'); \mathbf{W}, \mathbf{a}^0) : \|\mathbf{W} - \mathbf{W}^0\|_2 \le B_\mathrm{N} \right\}, \tag{2.9}$$

*for some $B_\mathrm{N} > 0$ and some fixed $(\mathbf{W}^0, \mathbf{a}^0)$ which can be interpreted as the initialization. We refer to Appendix D.4.1 for more details about neural function approximations and analysis techniques.*

**Example 2.9** ($\mathcal{S} \times \mathcal{A}$-rectangular robust factored MDP)**.** *We consider a factored MDP equipped with $\mathcal{S} \times \mathcal{A}$-rectangular factored robust set. A standard factored MDP [17] is defined as follows. Let $d$ be an integer and $\mathcal{O}$ be a finite set. The state space $\mathcal{S}$ is factored as $\mathcal{S} = \mathcal{O}^d$. For each $i \in [d]$, $s[i]$ is the $i$-coordinate of $s$ and it is only influenced by $s[\mathrm{pa}_i]$, where $\mathrm{pa}_i \subseteq [d]$. That is, the transition of a factored MDP can be factorized as*

$$P_h^\star(s'|s,a) = \prod_{i=1}^d P_{h,i}^\star(s'[i]|s[\mathrm{pa}_i], a).$$

*Here we let the realizable model space $\mathcal{P}_\mathrm{M}$ consist of all the factored transition kernels, i.e.,*

$$\mathcal{P}_\mathrm{M} = \left\{ P(s'|s,a) = \prod_{i=1}^d P_i(s'[i]|s[\mathrm{pa}_i], a) \ : \ P_i : \mathcal{S}[\mathrm{pa}_i] \times \mathcal{A} \mapsto \Delta(\mathcal{O}), \forall i \in [d] \right\}.$$

*For an $\mathcal{S} \times \mathcal{A}$-rectangular robust factored MDP, we define $\mathbf{\Phi}$ as, for any transition kernel $P(s'|s,a) = \prod_{i=1}^d P_i(s'[i]|s[\mathrm{pa}_i], a) \in \mathcal{P}_\mathrm{M}$, $\mathbf{\Phi}(P) = \bigotimes_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{P}_{\mathrm{Fac},\rho}(s,a;P)$, with*

$$\mathcal{P}_{\mathrm{Fac},\rho}(s,a;P) = \left\{ \prod_{i=1}^d \widetilde{P}_i(\cdot) : \widetilde{P}_i(\cdot) \in \Delta(\mathcal{O}), D(\widetilde{P}_i(\cdot) \| P_i(\cdot|s[\mathrm{pa}_i], a)) \le \rho_i, \forall i \in [d] \right\}.$$

*for some (pseudo-)distance $D$ on $\Delta(\mathcal{O})$ and some positive real numbers $\{\rho_i\}_{i=1}^d$.*

### 2.3 Offline Reinforcement Learning in Robust Markov Decision Processes

In this section, we define the offline RL protocol in a RMDP $(\mathcal{S}, \mathcal{A}, H, P^\star, R, \mathcal{P}_\mathrm{M}, \mathbf{\Phi})$. The learner is given the realizable model space $\mathcal{P}_\mathrm{M}$ and the robust mapping $\mathbf{\Phi}$, but the learner doesn't know the transition kernel $P^\star$. For simplicity, we assume that the learner knows the reward function $R$[2].

**Offline dataset generation.** We assume that the learner is given an offline dataset $\mathbb{D}$ that consists of $n$ i.i.d. trajectories generated from the standard MDP $(\mathcal{S}, \mathcal{A}, H, P^\star, R)$ using some behavior policy $\pi^\mathrm{b}$. For each $\tau \in [n]$, the trajectory has the form of $\{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{h=1}^H$, satisfying that $a_h^\tau \sim \pi_h^\mathrm{b}(\cdot|s_h^\tau)$, $r_h^\tau = R_h(s_h^\tau, a_h^\tau)$, and $s_{h+1}^\tau \sim P_h^\star(\cdot|s_h^\tau, a_h^\tau)$ for each step $h \in [H]$.

Given transition kernels $P = \{P_h\}_{h=1}^H$ and a policy $\pi$, we use $d_{P,h}^\pi(\cdot,\cdot)$ to denote the state-action visitation distribution at step $h$ when following policy $\pi$ and transition kernel $P$. With this notation, the distribution of $(s_h^\tau, a_h^\tau)$ can be written as $d_{P^\star,h}^{\pi^\mathrm{b}}$ or simply $d_{P^\star,h}^\mathrm{b}$, for each $\tau \in [n]$ and $h \in [H]$. We also use $d_{P^\star,h}^{\pi^\mathrm{b}}(\cdot)$ to denote the marginal distribution of state at step $h$ when there is no confusion.

**Learning objective.** In offline robust RL, the goal is to learn the policy $\pi^\star$ from the offline dataset $\mathbb{D}$ which maximizes the robust value function $V_{1,P^\star,\mathbf{\Phi}}^\pi$, that is,

$$\pi^\star = \operatorname*{argsup}_{\pi \in \Pi} V_{1,P^\star,\mathbf{\Phi}}^\pi(s_1), \quad s_1 \in \mathcal{S}, \tag{2.10}$$

where $\Pi = \{\pi = \{\pi_h\}_{h=1}^H \,|\, \pi_h : \mathcal{S} \mapsto \Delta(\mathcal{A})\}$ denotes the collection of all Markovian policies. In view of (2.10), we call $\pi^\star$ the *optimal robust policy*. Equivalently, we want to learn a policy $\widehat{\pi} \in \Pi$ which minimizes the suboptimality gap between $\widehat{\pi}$ and $\pi^\star$, defined as[3]

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) := V_{1,P^\star,\mathbf{\Phi}}^{\pi^\star}(s_1) - V_{1,P^\star,\mathbf{\Phi}}^{\widehat{\pi}}(s_1), \quad \forall s_1 \in \mathcal{S}. \tag{2.11}$$

[2]This is reasonable since learning the reward function is easier than learning the transition kernel.

[3]Without loss of generality, we assume that the initial state is fixed to some $s_1 \in \mathcal{S}$. Our algorithm and theory can be directly extended to the case when $s_1 \sim \rho \in \Delta(\mathcal{S})$.

**Algorithm 1** Doubly Pessimistic Model-based Policy Optimization ($\texttt{P}^2\texttt{MPO}$)

1: **Input**: model space $\mathcal{P}_{\mathrm{M}}$, mapping $\mathbf{\Phi}$, dataset $\mathbb{D}$, policy class $\Pi$, algorithm $\texttt{ModelEst}$.
2: Model estimation step:
3: Obtain a confidence region $\widehat{\mathcal{P}} = \texttt{ModelEst}(\mathbb{D}, \mathcal{P}_{\mathrm{M}})$.
4: Doubly pessimistic policy optimization step:
5: Set policy $\widehat{\pi}$ as $\mathrm{argsup}_{\pi\in\Pi} J_{\texttt{Pess}^2}(\pi)$, where $J_{\texttt{Pess}^2}(\pi)$ is defined in (3.1).
6: **Output**: $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$.

# 3 Algorithm: Generic Framework and Unified Theory

In this section, we propose Doubly Pessimistic Model-based Policy Optimization ($\texttt{P}^2\texttt{MPO}$) algorithm to solve offline RL in the RMDP framework we introduce in Section 2.1, and we establish a unified theoretical guarantee for $\texttt{P}^2\texttt{MPO}$. Our proposed algorithm and theory show that *double pessimism* is a general principle for designing efficient algorithms for offline robust RL. The algorithm features three key points: i) learning the optimal robust policy $\pi^\star$ approximately; ii) requiring only a partial coverage property of the offline dataset $\mathbb{D}$; iii) able to handle infinite state space via function approximations.

We first introduce our proposed algorithm framework $\texttt{P}^2\texttt{MPO}$ in Section 3.1. Then we establish a unified analysis for $\texttt{P}^2\texttt{MPO}$ in Section 3.2. Our algorithm framework can be specified to solve all the concrete examples of RMDP we introduce in Section 2.2, which we show in Section 4.

## 3.1 Algorithm Framework: P$^2$MPO

The $\texttt{P}^2\texttt{MPO}$ algorithm framework (Algorithm 1) consists of a *model estimation step* and a *doubly pessimistic policy optimization step*, which we introduce in the following respectively.

**Model estimation step (Line 3).** The $\texttt{P}^2\texttt{MPO}$ algorithm framework first constructs an estimation of the transition kernels $P^\star = \{P_h^\star\}_{h=1}^H$, i.e., it estimates the dynamic of the training environment. It implements a sub-algorithm $\texttt{ModelEst}(\mathbb{D}, \mathcal{P}_{\mathrm{M}})$ that returns a confidence region $\widehat{\mathcal{P}}$ for $P^\star = \{P_h^\star\}_{h=1}^H$. Specifically, $\widehat{\mathcal{P}} = \{\widehat{\mathcal{P}}_h\}_{h=1}^H$ with $\widehat{\mathcal{P}}_h \subseteq \mathcal{P}_{\mathrm{M}}$ for each step $h \in [H]$.

The sub-algorithm $\texttt{ModelEst}$ can be tailored to specific RMDPs. We refer to Section 4 for detailed implementations of $\texttt{ModelEst}$ for different examples of RMDPs introduced in Section 2.2. Ideally, to ensure sample-efficient learning, we need $\widehat{\mathcal{P}} = \texttt{ModelEst}(\mathbb{D}, \mathcal{P}_{\mathrm{M}})$ to satisfy: i) the transition kernels $P^\star = \{P_h^\star\}_{h=1}^H$ are contained in $\widehat{\mathcal{P}} = \{\widehat{\mathcal{P}}_h\}_{h=1}^H$; ii) each transition kernel $P_h \in \widehat{\mathcal{P}}_h$ enjoys a small "robust estimation error" which is highly related to the robust Bellman equation in (2.5). We quantify these two conditions of $\widehat{\mathcal{P}}$ for sample-efficient learning in Section 3.2.

**Doubly pessimistic policy optimization step (Line 5).** After **model estimation step**, $\texttt{P}^2\texttt{MPO}$ performs policy optimization to find the optimal robust policy. To learn the optimal robust policy in the face of uncertainty, $\texttt{P}^2\texttt{MPO}$ adopts a *double pessimism* principle. To explain, this general principle has two sources of pessimism: i) pessimism in the face of data uncertainty; ii) pessimism to find a robust policy. Specifically, for any policy, we first estimate its robust value function via two infimums, where one is an infimum over the confidence set constructed in the model estimation step, and one is an infimum over the robust sets. Formally, for any $\pi \in \Pi$, we define the doubly pessimistic estimator

$$J_{\texttt{Pess}^2}(\pi) = \inf_{P_h\in\widehat{\mathcal{P}}_h, 1\le h\le H} \inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h), 1\le h\le H} V_1^\pi(s_1; \{\widetilde{P}_h\}_{h=1}^H), \tag{3.1}$$

where $V_1^\pi$ is the standard value function of policy $\pi$ defined in (2.3). Then $\texttt{P}^2\texttt{MPO}$ outputs the policy $\widehat{\pi}$ that maximizes the doubly pessimistic estimator $J_{\texttt{Pess}^2}(\pi)$ defined in (3.1), i.e.,

$$\widehat{\pi} = \mathop{\mathrm{argsup}}_{\pi\in\Pi} J_{\texttt{Pess}^2}(\pi). \tag{3.2}$$

The novelty of the doubly pessimistic policy optimization step is performing pessimism from the two sources (data uncertainty and robust optimization) simultaneously. Compared with the previous works on standard offline RL in MDPs [62, 54] and offline RL in RMDPs without pessimism [68, 77, 40], they only contain one source of pessimism in algorithm design, contrasting with our algorithm.

We note that a recent work [47] also studied robust offline RL in $\mathcal{S}\times\mathcal{A}$-rectangular robust tabular MDPs (Example 2.5) using pessimism techniques. Compared with our double pessimism principle,

their algorithm performs pessimism in face of data uncertainty i) depending on the tabular structure of the model since a point-wise pessimism penalty is needed and ii) depending on the specific form of the robust set $\mathbf{\Phi}(P)$, which makes it difficult to adapt to the infinite state space case with general function approximations and general types of robust set $\mathbf{\Phi}(P)$.

### 3.2 Unified Theoretical Analysis

In this section, we establish a unified theoretical analysis for the P$^2$MPO algorithm framework proposed in Section 3.1. We first specify the two conditions that the **model estimation step** of P$^2$MPO should satisfy in order for sample-efficient learning. Then we establish an upper bound of suboptimality of the policy obtained by P$^2$MPO given that these two conditions are satisfied. In Section 4, we show that the specific implementations of the sub-algorithm ModelEst for the RMDPs examples in Section 2.2 satisfy these two conditions, which results in tailored suboptimality bounds for these examples.

**Conditions.** The two conditions on the **model estimation step** are given by the following.

**Condition 3.1** ($\delta$-accuracy). With probability at least $1-\delta$, it holds that $P_h^\star \in \widehat{\mathcal{P}}_h$ for any $h \in [H]$.

**Condition 3.2** ($\delta$-model estimation error). With probability at least $1-\delta$, it holds that

$$\mathbb{E}_{(s,a)\sim d^{\pi_{\mathrm{b}}}_{P^\star,h}}\left[\Big(\inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}\widetilde{\mathbb{P}}_h(V^{\pi^\star}_{h+1,P,\mathbf{\Phi}})(s,a)-\inf_{\widetilde{P}_h\in\mathbf{\Phi}(P^\star_h)}\widetilde{\mathbb{P}}_h(V^{\pi^\star}_{h+1,P,\mathbf{\Phi}})(s,a)\Big)^2\right]\le \mathrm{Err}^{\mathbf{\Phi}}_h(n,\delta).$$

for any $P=\{P_h\}_{h=1}^H$ with $P_h\in\widehat{\mathcal{P}}_h$. Here $\widetilde{\mathbb{P}}_h(V^{\pi^\star}_{h+1,P,\mathbf{\Phi}})(s,a)=\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s,a)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')]$

Condition 3.1 requires that the confidence region $\widehat{\mathcal{P}}_h$ contains the transition kernel of the training environment $P_h^\star$ with high probability. Condition 3.2 requires that each transition kernel $P_h\in\widehat{\mathcal{P}}_h$ induces an error from $P_h^\star$ less than certain quantity $\mathrm{Err}^{\mathbf{\Phi}}_h(n,\delta)$, where the error is adapted from the robust Bellman equation (2.5) and involves an infimum over the robust set of $P_h$ and $P_h^\star$. In specific implementations of ModelEst for RMDP examples in Section 4, we show that the quantity $\mathrm{Err}^{\mathbf{\Phi}}_h(n,\delta)$ generally scales with $\widetilde{\mathcal{O}}(n^{-1})$, where $n$ is the number of trajectories in the offline dataset.

**Suboptimality analysis.** Now we establish a unified suboptimality bound for the P$^2$MPO algorithm framework. Thanks to the double pessimism principle of P$^2$MPO, we can prove a suboptimality bound while only making a mild *robust partial coverage assumption* on the dataset.

**Assumption 3.3** (Robust partial coverage)**.** *We assume that*

$$C^\star_{P^\star,\mathbf{\Phi}}:=\sup_{1\le h\le H}\sup_{P=\{P_h\}_{h=1}^H,P_h\in\mathbf{\Phi}(P^\star_h)}\mathbb{E}_{(s,a)\sim d^{\pi_{\mathrm{b}}}_{P^\star,h}}\left[\left(\frac{d^{\pi^\star}_{P,h}(s,a)}{d^{\pi_{\mathrm{b}}}_{P^\star,h}(s,a)}\right)^2\right]<+\infty,$$

*and we call $C^\star_{P^\star,\mathbf{\Phi}}$ the robust partial coverage coefficient.*

To interpret, Assumption 3.3 only requires that the dataset covers the visitation distribution of the optimal policy $\pi^\star$, but in a robust fashion since $C^\star_{P^\star,\mathbf{\Phi}}$ considers all possible transition kernels in the robust set $\mathbf{\Phi}(P^\star)$. The robust consideration in $C^\star_{P^\star,\mathbf{\Phi}}$ is because in RMDPs the policies are all evaluated in a robust way. This partial-coverage-style assumption is much weaker than full-coverage-style assumptions [68, 77, 40] which require either a uniformly lower bounded dataset distribution or covering the visitation distribution of any $\pi\in\Pi$. For $\mathcal{S}\times\mathcal{A}$-rectangular robust tabular MDPs (Example 2.5), the robust partial coverage coefficient $C^\star_{P^\star,\mathbf{\Phi}}$ is similar with the partial coverage coefficient proposed by [47] who studied tabular RMDPs under partial coverage. We highlight that beyond $\mathcal{S}\times\mathcal{A}$-rectangular robust tabular MDPs, our robust partial coverage assumption can handle other examples of RMDPs (Section 2.2) under our unified theory.

Our main result is the following theorem. See Appendix C for a detailed proof.

**Theorem 3.4** (Suboptimality of P$^2$MPO)**.** *Under Assumptions 2.2 and 3.3, suppose that Algorithm 1 implements a sub-algorithm that satisfies Conditions 3.1 and 3.2, then with probability at least $1-2\delta$,*

$$\mathrm{SubOpt}(\widehat{\pi};s_1)\le\sqrt{C^\star_{P^\star,\mathbf{\Phi}}}\cdot\sum_{h=1}^H\sqrt{\mathrm{Err}^{\mathbf{\Phi}}_h(n,\delta)}.$$

When $\mathrm{Err}^{\mathbf{\Phi}}_h(n,\delta)$ achieves a rate of $\widetilde{\mathcal{O}}(n^{-1})$, then P$^2$MPO enjoys a $\widetilde{\mathcal{O}}(n^{-1/2})$-suboptimality. In the following Section 4, we give specific implementations of the model estimation step of P$^2$MPO for each example of RMDP in Section 2. The implementations will make Conditions 3.1 and 3.2 satisfied and thus specify the unified result Theorem 3.4.

## 4 Implementations of P$^2$MPO for Examples of RMDPs

In this section, we provide concrete implementations of the `ModelEst` sub-algorithm in P$^2$MPO (Algorithm 1). In Section 4.1, we implement `ModelEst` for all the RMDPs that satisfy Assumption 2.2, and we specify the suboptimality bounds in Theorem 3.4 to Examples 2.5, 2.7, 2.8 in Section 2.2. In Section 4.2, we implement `ModelEst` for $\mathcal{S} \times \mathcal{A}$-rectangular robust factored MDPs (Example 2.9) and specify Theorem 3.4 to this example.

### 4.1 Model Estimation for General RMDPs with $\mathcal{S} \times \mathcal{A}$-rectangular Robust Sets

Using the offline data $\mathbb{D}$, we first construct the *maximum likelihood estimator* (MLE) of the transition kernel $P^\star$. Specifically, for each step $h \in [H]$, we define

$$\widehat{P}_h = \mathop{\arg\max}_{P \in \mathcal{P}_{\mathrm{M}}} \frac{1}{n} \sum_{\tau=1}^{n} \log P(s_{h+1}^\tau | s_h^\tau, a_h^\tau). \tag{4.1}$$

After, we construct a confidence region for the MLE estimator, denoted by $\widehat{\mathcal{P}}$. Specifically, $\widehat{\mathcal{P}}$ contains all transitions which have a small total variance distance from $\widehat{P}$. For each step $h \in [H]$, we define

$$\widehat{\mathcal{P}}_h = \bigg\{ P \in \mathcal{P}_{\mathrm{M}} : \frac{1}{n} \sum_{\tau=1}^{n} \|\widehat{P}_h(\cdot | s_h^\tau, a_h^\tau) - P(\cdot | s_h^\tau, a_h^\tau)\|_1^2 \le \xi \bigg\}. \tag{4.2}$$

Here $\xi > 0$ is a tuning parameter that controls the size of the confidence region $\widehat{\mathcal{P}}_h$. Finally, we set $\texttt{ModelEst}(\mathbb{D}, \mathcal{P}_{\mathrm{M}}) = \widehat{\mathcal{P}} = \{\widehat{\mathcal{P}}_h\}_{h=1}^H$ with $\widehat{\mathcal{P}}_h$ given in (4.2). In the sequel, we mainly consider the distance $D(\cdot\|\cdot)$ in Assumption 2.2 to be KL-divergence and TV-distance. The following corollary specifies Theorem 3.4 to model estimation step given by (4.2). See Appendix D for a detailed proof.

**Corollary 4.1** (Suboptimality of P$^2$MPO: $\mathcal{S} \times \mathcal{A}$-rectangular robust MDP)**.** *Under Assumption 2.2, 3.3, setting the tuning parameter $\xi$ 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 $C_1, C_2 > 0$, P$^2$MPO with model estimation step given by (4.2) satisfies that*

- ♠ *when $D(\cdot\|\cdot)$ is KL-divergence and Assumption D.3 holds with parameter $\underline{\lambda}$, then with probability at least $1 - 2\delta$,*

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \le \frac{\sqrt{C^\star_{P^\star, \mathbf{\Phi}}} H^2 \exp(H/\underline{\lambda})}{\rho} \cdot \sqrt{\frac{C_1' \log(C_2' H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}}.$$

- ♠ *when $D(\cdot\|\cdot)$ is TV-divergence, then with probability at least $1 - 2\delta$,*

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \le \sqrt{C^\star_{P^\star, \mathbf{\Phi}}} H^2 \cdot \sqrt{\frac{C_1' \log(C_2' H \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}}.$$

**$\mathcal{S} \times \mathcal{A}$-rectangular robust tabular MDP (Example 2.5).** When $\mathcal{S}$ is finite as in Example 2.5, the MLE estimator (4.1) coincides the empirical estimator

$$\widehat{P}_h(s'|s,a) = \frac{\sum_{\tau=1}^n \mathbf{1}\{s_h^\tau = s, a_h^\tau = a, s_{h+1}^\tau = s'\}}{1 \vee \sum_{\tau=1}^n \mathbf{1}\{s_h^\tau = s, a_h^\tau = a\}}, \tag{4.3}$$

which is adopted by [77, 68, 39, 47, 40]. Furthermore, in Example 2.5, the realizable model space $\mathcal{P}_{\mathrm{M}} = \{P : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})\}$. When $\mathcal{S}$ and $\mathcal{A}$ are finite, we can bound the bracket number of $\mathcal{P}_{\mathrm{M}}$ as

$$\log \mathcal{N}_{[]}(1/n^2, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty}) \le 2|\mathcal{S}|^2|\mathcal{A}| \log(n). \tag{4.4}$$

Combining (4.4) and Corollary 4.1, we can conclude that: i) under TV-distance the suboptimality of P$^2$MPO for $\mathcal{S} \times \mathcal{A}$-rectangular robust tabular RMDP is given by $\mathcal{O}(H^2 \sqrt{C^\star_{P^\star, \mathbf{\Phi}} |\mathcal{S}|^2 |\mathcal{A}| \log(nH/\delta)/n})$, ii) under KL-divergence the suboptimality of P$^2$MPO for $\mathcal{S} \times \mathcal{A}$-rectangular robust tabular MDP is given by $\mathcal{O}(H^2 \exp(H/\underline{\lambda})/\rho \cdot \sqrt{C^\star_{P^\star, \mathbf{\Phi}} |\mathcal{S}|^2 |\mathcal{A}| \log(nH/\delta)/n})$. We prove (4.4) in Appendix D.2.

**Remark 4.2.** *We note that for KL-divergence robust sets, the dependence on* $\exp(H)$ *is due to the usage of general function approximations, which also appears in a recent work [29] for RMDPs with linear function approximations. For the special case of robust tabular MDPs, under KL-divergence, existing work [47] derived sample complexities without* $\exp(H)$*, but with an additional dependence on* $1/d^{\mathrm{b}}_{\min}$ *and* $1/P^{\star}_{\min}$ *Here* $d^{\mathrm{b}}_{\min} = \min_{(s,a,h):d^{\pi^{\mathrm{b}}}_{P^\star,h}(s,a)>0} d^{\pi^{\mathrm{b}}}_{P^\star,h}(s,a)$ *and* $P^{\star}_{\min} = \min_{(s,s',h):P_h(s'|s,\pi^\star_h(s))>0} P^\star_h(s'|s,\pi^\star_h(s))$*. We remark that our analysis for* `P`$^2$`MPO` *algorithm can be tailored to the tabular case and become* $\exp(H)$*-free using their techniques, with the cost of an additional dependence on* $1/d^{\mathrm{b}}_{\min}$ *and* $1/P^{\star}_{\min}$*. But we note that in the infinite state space case, both the* $1/d^{\mathrm{b}}_{\min}$*-dependence and the* $1/P_{\min}$*-dependence becomes problematic. So, it serves as an interesting future work to answer whether one can derive both* $\exp(H)$*-free and* $(1/d^{\mathrm{b}}_{\min}, 1/P^{\star}_{\min})$*-free results for (general) function approximations under KL-divergence.*

**$\mathcal{S}\times\mathcal{A}$-rectangular robust MDP with kernel and neural function approximations (Examples 2.7 and 2.8).** By specifying the bracket numbers in Corollary 4.1, we can provide the detailed suboptimality guarantees for $\mathcal{S}\times\mathcal{A}$-rectangular robust MDP with kernel and neural function approximations. Due to space limitations, we defer the detailed results to Appendices D.3 and D.4.

### 4.2 Model Estimation for $\mathcal{S}\times\mathcal{A}$-rectangular Robust Factored MDPs (Example 2.9)

We first construct MLE estimator for each factor $P^\star_{h,i}$ of the transition $P^\star_h = \prod_{i=1}^d P^\star_{h,i}$, that is,

$$\widehat{P}_{h,i} = \mathop{\arg\max}_{P_i:\mathcal{S}[\mathrm{pa}_i]\times\mathcal{A}\mapsto\Delta(\mathcal{O})} \frac{1}{n}\sum_{k=1}^n \log P(s^\tau_{h+1}[i]|s^\tau_h[\mathrm{pa}_i], a^\tau_h). \tag{4.5}$$

Then given $\{\widehat{P}_{h,i}\}_{i=1}^d$ we construct a confidence region that is factored across $i\in[d]$. Specifically,

$$\widehat{\mathcal{P}}_h = \left\{ P(s'|s,a) = \prod_{i=1}^d P_i(s'[i]|s[\mathrm{pa}_i],a) : \frac{1}{n}\sum_{i=1}^n \|(P_i - \widehat{P}_{h,i})(\cdot|s^\tau_h[\mathrm{pa}_i], a^\tau_h)\|_1^2 \le \xi_i, \forall i \right\}. \tag{4.6}$$

Finally, we set `ModelEst`$(\mathbb{D}, \mathcal{P}_{\mathrm{M}}) = \widehat{\mathcal{P}} = \{\widehat{\mathcal{P}}\}_{h=1}^H$ with $\widehat{\mathcal{P}}_h$ given in (4.6). The following corollary specifies Theorem 3.4 to model estimation step given by (4.6). See Appendix E for a detailed proof.

**Corollary 4.3** (Suboptimality of `P`$^2$`MPO`: $\mathcal{S}\times\mathcal{A}$-rectangular robust factored MDP)**.** *Supposing the RMDP is an* $\mathcal{S}\times\mathcal{A}$*-rectangular robust factored MDP, under the same Assumptions and parameter choice in Theorem 3.4 and Proposition E.1,* `P`$^2$`MPO` *with model estimation step given by (4.6) satisfies*

- ♣ *when* $D(\cdot\|\cdot)$ *is KL-divergence and Assumption D.3 holds with parameter* $\underline{\lambda}$*, then with probability at least* $1-2\delta$*, (defining* $\rho_{\min} = \min_{i\in[d]}\rho_i$*)*

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \le \frac{\sqrt{C^\star_{P^\star,\mathbf{\Phi}}} H^2 \exp(H/\underline{\lambda})}{\rho_{\min}} \cdot \sqrt{\frac{dC_1' \sum_{i=1}^d |\mathcal{O}|^{1+|\mathrm{pa}_i|}|\mathcal{A}| \log(C_2' nd/\delta)}{n}}.$$

- ♣ *when* $D(\cdot\|\cdot)$ *is TV-divergence, then with probability at least* $1-2\delta$*,*

$$\mathrm{SubOpt}(\widehat{\pi}; s_1) \le \sqrt{C^\star_{P^\star,\mathbf{\Phi}}} H^2 \sqrt{\frac{dC_1' \sum_{i=1}^d |\mathcal{O}|^{1+|\mathrm{pa}_i|}|\mathcal{A}| \log(C_2' nd/\delta)}{n}}.$$

Compared with the suboptimality bounds for $\mathcal{S}\times\mathcal{A}$-rectangular robust MDPs in Section 4.1, the suboptimality of $\mathcal{S}\times\mathcal{A}$-rectangular robust factored MDPs with `ModelEst` given in (4.6) only scales with $\sum_{i=1}^d |\mathcal{O}|^{1+|\mathrm{pa}_i|}$ instead of scaling with $|\mathcal{S}| = \prod_{i=1}^d |\mathcal{O}|$ which is of order $\exp(d)$. This justifies the benefit of considering $\mathcal{S}\times\mathcal{A}$-rectangular robust factored MDPs when the transition kernels of training and testing environments enjoy factored structures.

## 5 Conclusion and Discussions

This paper proposes a general learning principle — double pessimism — for robust offline RL. Based on this learning principle, we propose a generic algorithm that only requires robust partial coverage data to solve $\mathcal{S}\times\mathcal{A}$-rectangular RMDPs with general function approximation. Our results are ready to be extended to $d$-rectangular linear RMDPs [29]. See Appendix A for details. In Appendix A, we also provide some challenges to perform sample efficient RL in $\mathcal{S}$-rectangular RMDPs.

## Acknowledgments and Disclosure of Funding

The material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF 1915967, 2118199, 2229012, 2312204.

## References

[1] Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pages 10–4, 2019.

[2] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

[3] Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR, 2021.

[4] Qi Cai, Zhuoran Yang, Csaba Szepesvari, and Zhaoran Wang. Optimistic policy optimization with general function approximations. 2020.

[5] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

[6] Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. *arXiv preprint arXiv:2202.02446*, 2022.

[7] Pierre Clavier, Erwan Le Pennec, and Matthieu Geist. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

[8] Jing Dong, Jingwei Li, Baoxiang Wang, and Jingzhao Zhang. Online policy optimization for robust mdp. *arXiv preprint arXiv:2209.13841*, 2022.

[9] Laurent El Ghaoui and Arnab Nilim. Robust solutions to markov decision problems with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[10] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Robust $\phi$-divergence mdps. 2022.

[11] Jiachen Hu, Han Zhong, Chi Jin, and Liwei Wang. Provable sim-to-real transfer in continuous domain with partial observations. *arXiv preprint arXiv:2210.15598*, 2022.

[12] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pages 1695–1724, 2013.

[13] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

[14] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[15] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[16] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[17] Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pages 740–747, 1999.

[18] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[19] Yufei Kuang, Miao Lu, Jie Wang, Qi Zhou, Bin Li, and Houqiang Li. Learning robust policy against disturbance in transition dynamics via state-conservative policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7247–7254, 2022.

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

[21] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

[22] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[23] Chris Junchi Li, Dongruo Zhou, Quanquan Gu, and Michael I Jordan. Learning two-player mixture markov games: Kernel function approximation and correlated equilibrium. *arXiv preprint arXiv:2208.05363*, 2022.

[24] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

[25] Zhihan Liu, Miao Lu, Zhaoran Wang, Michael Jordan, and Zhuoran Yang. Welfare maximization in competitive equilibrium: Reinforcement learning for markov exchange economy. In *International Conference on Machine Learning*, pages 13870–13911. PMLR, 2022.

[26] Zhihan Liu, Yufeng Zhang, Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *International Conference on Machine Learning*, pages 14094–14138. PMLR, 2022.

[27] Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust $q$-learning. In *International Conference on Machine Learning*, pages 13623–13643. PMLR, 2022.

[28] Miao Lu, Yifei Min, Zhaoran Wang, and Zhuoran Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. *arXiv preprint arXiv:2205.13589*, 2022.

[29] Xiaoteng Ma, Zhipeng Liang, Li Xia, Jiheng Zhang, Jose Blanchet, Mingwen Liu, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

[30] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Li Fei-Fei, and Silvio Savarese. Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3932–3939. IEEE, 2017.

[31] Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance in value function estimation. In *Proceedings of the twenty-first international conference on Machine learning*, page 72, 2004.

[32] Jun Morimoto and Kenji Doya. Robust reinforcement learning. *Neural computation*, 17(2): 335–359, 2005.

[33] Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[34] Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

[35] Ariel Neufeld and Julian Sester. Robust $q$-learning algorithm for markov decision processes under wasserstein uncertainty. *arXiv preprint arXiv:2210.00898*, 2022.

[36] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[37] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafał Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *CoRR*, 2018. URL http://arxiv.org/abs/1808.00177.

[38] Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos Theodorou, and Byron Boots. Agile autonomous driving using end-to-end deep imitation learning. *arXiv preprint arXiv:1709.07174*, 2017.

[39] Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR, 2022.

[40] Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*, 2022.

[41] Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. *arXiv preprint arXiv:1712.03632*, 2017.

[42] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

[43] Lerrel Pinto, James Davidson, and Abhinav Gupta. Supervision via competition: Robot adversaries for learning tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1601–1608. IEEE, 2017.

[44] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017.

[45] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

[46] Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*, 2022.

[47] Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

[48] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. *arXiv preprint arXiv:2202.13890*, 2022.

[49] Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust batch contextual bandits. *Management Science*, 2023.

[50] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[51] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

[52] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[53] Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pages 6215–6224. PMLR, 2019.

[54] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

[55] Sara A Van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[56] Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2447–2456, 2018.

[57] Qiuhao Wang, Chin Pang Ho, and Marek Petrik. On the convergence of policy gradient in robust mdps. *arXiv preprint arXiv:2212.10439*, 2022.

[58] Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3370–3398. PMLR, 2023.

[59] Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

[60] Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *International Conference on Machine Learning*, pages 23484–23526. PMLR, 2022.

[61] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[62] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

[63] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

[64] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

[65] Huan Xu and Shie Mannor. Distributionally robust markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

[66] Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9728–9754. PMLR, 2023.

[67] Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

[68] Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

[69] Wenhao Yang, Han Wang, Tadashi Kozuno, Scott M Jordan, and Zhihua Zhang. Avoiding model estimation in robust markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*, 2023.

[70] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems*, 33:13903–13916, 2020.

[71] Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078, 2021.

[72] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

[73] Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

[74] Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

[75] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020.

[76] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744. IEEE, 2020.

[77] Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR, 2021.

# 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*

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

*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

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

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 $c^\dagger > 0$,*

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

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

**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 $\mathbf{\Phi}$ is defined as, for $\forall P \in \mathcal{P}_{\mathrm{M}}$,*

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

*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.,*

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

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

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

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* $(s,a) \in \mathcal{S} \times \mathcal{A}$,

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

On the other hand, using condition (B.1) and the definition of (robust) value function $V_{h,P,\mathbf{\Phi}}^{\pi}$ and $V_h^{\pi}$ 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 $V^{\pi}_{h+1}$ in (2.3), the first equality follows from condition (B.1), and the second equality follows from the definition of $V^{\pi}_{h+1,P,\mathbf{\Phi}}$ in (2.1). Note that the right hand side of (B.5) equals to $Q^{\pi}_{h,P,\mathbf{\Phi}}(s,a)$. Therefore, all the inequalities are actually equalities. On the one hand, from (B.4), we can know that,

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

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

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

This proves the existence of $\{P_i^{\pi,\dagger}\}_{i=h}^{H}$ in the conclusion 2. for step $h$ and $Q^{\pi}_{h,P,\mathbf{\Phi}}$. The remaining of the proof is to prove the $V^{\pi}_{h,P,\mathbf{\Phi}}$ 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 $V^{\pi}_{h,P,\mathbf{\Phi}}$ in (2.1), we have that,

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

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

$$V^{\pi}_{h,P,\mathbf{\Phi}}(s) \le \sum_{a\in\mathcal{A}} \pi_h(a|s) Q^{\pi}_{h,P,\mathbf{\Phi}}(s,a) \tag{B.8}$$

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

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

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

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

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

This proves the $V^{\pi}_{h,P,\mathbf{\Phi}}$ 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

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

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

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

This finishes the proof of Proposition 2.3 under Assumption 2.2. □

# 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

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

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

$$\Delta_{h,P,\mathbf{\Phi}}(s_h,a_h) = Q_{h,P^\star,\mathbf{\Phi}}^{\pi^\star}(s_h,a_h) - Q_{h,P,\mathbf{\Phi}}^{\pi^\star}(s_h,a_h). \qquad \text{(C.4)}$$

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

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

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

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

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 $(s_h,a_h)\in\mathcal{S}$,

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

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

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

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 $\Delta^{(\mathrm{ii})}_{h,P,\mathbf{\Phi}}(s_h,a_h)$. Combining this with (C.6), we can derive that,

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

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

$$\begin{aligned}\mathbb{E}_{a_1\sim\pi^\star_1(\cdot|s_1)}[\Delta_{1,P,\mathbf{\Phi}}(s_1,a_1)] &\le \sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{P^{\pi^\star,\dagger},h}}[\Delta^{(\mathrm{ii})}_{h,P,\mathbf{\Phi}}(s_h,a_h)] \\ &= \sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{P^{\pi^\star,\dagger},h}}\Bigg[\inf_{\widetilde{P}_h\in\mathbf{\Phi}(P^\star_h)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')] \\ &\quad - \inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)} \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')]\Bigg], \end{aligned} \tag{C.8}$$

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 $h\in[H]$,

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

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

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

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

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

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 $\xi$ 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*

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

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

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

*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,*

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

*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*

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

*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

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

By Assumption D.3 and Lemma H.7, we know that the optimal value of $\lambda$ 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
&\left|\inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h^\star)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')] - \inf_{\widetilde{P}_h\in\mathbf{\Phi}(P_h)}\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s_h,a_h)}[V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')]\right| \\
&\quad= \left|\sup_{\lambda\in\mathbb{R}}\left\{-\mathbb{E}_{s'\sim P_h^\star(\cdot|s_h,a_h)}\left[\left(\lambda - V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')\right)_+\right] - \frac{\rho}{2}\left(\lambda - \inf_{s''\in\mathcal{S}} V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s'')\right) + \lambda\right\}\right. \\
&\qquad\left. - \sup_{\lambda\in\mathbb{R}}\left\{-\mathbb{E}_{s'\sim P_h(\cdot|s_h,a_h)}\left[\left(\lambda - V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')\right)_+\right] - \frac{\rho}{2}\left(\lambda - \inf_{s''\in\mathcal{S}} V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s'')\right) + \lambda\right\}\right| \qquad \text{(D.9)} \\
&\quad\le \left|\sup_{\lambda\in\mathbb{R}}\left\{\left(\mathbb{E}_{s'\sim P_h^\star(\cdot|s_h,a_h)} - \mathbb{E}_{s'\sim P_h(\cdot|s_h,a_h)}\right)\left[\left(\lambda - V^{\pi^\star}_{h+1,P,\mathbf{\Phi}}(s')\right)_+\right]\right\}\right| \qquad \text{(D.10)}
\end{aligned}$$

As is shown in Lemma H.8, the optimal value of $\lambda$ 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

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

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

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

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

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

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 $\mathcal{F} = \{f(s,a,s') : \|f\|_\infty \le 1\}$ 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,

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

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),

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

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

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

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

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

### 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 $\epsilon > 0$,*

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

*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 $B_{\mathrm{N}}$ 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 $\mathtt{P}^2\mathtt{MPO}$ for $\mathcal{S} \times \mathcal{A}$-rectangular robust MDPs with neural function approximations is,

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

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

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

#### 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 $\epsilon > 0$,*

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

*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

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

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 $m \geq d_{\mathcal{X}} B_{\mathrm{N}}^4/\epsilon^2$,

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

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 $\xi_i$ defined in (4.6) as*

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

*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),

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

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

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

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

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

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

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

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

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

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

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

which satisfies $0 \le v^j_h \le H$. For each $1 \le j \le d$, we can then upper bound

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

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

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

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 $1-\delta$,*

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

*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 $h\in[H]$,

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

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

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

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

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

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 $\boldsymbol{\Phi}(\cdot)$ in Example A.1, we can derive that

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

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

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

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 $\lambda$, we can derive that

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

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

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

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,

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

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

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

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

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

with probability at least $1 - \delta$, where the first inequality holds since $\exp\{-V^{\pi^\star}_{h+1,P,\boldsymbol{\Phi}}(s')/\lambda_i\} \in \mathcal{V}$, 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

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

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 $1-\delta$,

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

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

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

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

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

For notational simplicity, in the sequel, we denote by

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

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,

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

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

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

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

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

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

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

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 $1-\delta$,*

$$\sup_{v \in \mathcal{V}} \left| \int_{\mathcal{S}} \big( P_h^\star(\mathrm{d}s'|s,a) - \widehat{P}_h(\mathrm{d}s'|s,a) \big) v(s') \right|^2$$
$$\le C_1 \cdot \|\boldsymbol{\phi}(s,a)\|^2_{\mathbf{\Lambda}^{-1}_{h,\alpha}} \cdot \frac{dH^2 \log(C_2 ndH/\delta)}{n},$$

*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 $h \in [H]$,

$$\sup_{v \in \mathcal{V}} \frac{1}{n} \sum_{\tau=1}^{n} \left| \int_{\mathcal{S}} \big( P_h^\star(\mathrm{d}s'|s_h^\tau, a_h^\tau) - \widehat{P}_h(\mathrm{d}s'|s_h^\tau, a_h^\tau) \big) v(s') \right|^2 \le \frac{1}{n} \sum_{\tau=1}^{n} \|\boldsymbol{\phi}(s_h^\tau, a_h^\tau)\|^2_{\mathbf{\Lambda}^{-1}_{h,\alpha}} \cdot \frac{C_1 dH^2 \log(C_2 ndH/\delta)}{n}.$$

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

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

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

$$\sup_{v \in \mathcal{V}} \frac{1}{n} \sum_{\tau=1}^{n} \left| \int_{\mathcal{S}} \big( P_h^\star(\mathrm{d}s'|s_h^\tau, a_h^\tau) - \widehat{P}_h(\mathrm{d}s'|s_h^\tau, a_h^\tau) \big) v(s') \right|^2 \le \frac{C_1 d^2 H^2 \log(C_2 ndH/\delta)}{n} = \xi.$$

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,

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

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

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

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

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

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

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

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

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

for any step $h\in[H]$, $(s_h,a_h)\in\mathcal{S}\times\mathcal{A}$, and $P_h\in\widehat{\mathcal{P}}_h$. 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 $1-\delta$,

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

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,

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

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

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

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

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

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 $\epsilon$-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 $v \in \mathcal{C}_{\mathcal{V},\epsilon}$.

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

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 $\|v - \widehat{v}\|_\infty \le \epsilon$. Then we have that

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

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

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

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,

$$
\begin{aligned}
&\frac{2}{n^2}\cdot\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2\cdot\left\|\sum_{\tau=1}^n\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\left(\int_{\mathcal{S}}P_h^\star(\mathrm{d}s'|s_h^\tau,a_h^\tau)v(s')-v(s_{h+1}^\tau)\right)\right\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2\\
&\quad\le\frac{36}{n}\cdot\big(\log(H/\delta)+d\log(1+4\sqrt{n}Hd/(\underline{\lambda}))+\log(1+4\sqrt{n}Hd/(\underline{\lambda}^2\rho))+d\log(1+n)\big)\cdot\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2\\
&\qquad+\frac{16}{n}\cdot\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2,\\
&\quad\le\frac{C_1d\big(\log(1+C_2nH/\delta)+\log(1+C_3ndH/(\rho\underline{\lambda}^2))\big)}{n}\cdot\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,\alpha}^{-1}}^2,
\end{aligned}
\tag{F.30}
$$

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 $h \in [H]$,

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

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 $\epsilon$-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,

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

This finishes the proof of Lemma F.3. □

### F.2 Other Lemmas

**Lemma F.4** (Covering number of $\mathcal{V}$: KL-divergence case)**.** *The $\epsilon$-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,

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

Thus we conclude that to form an $\epsilon$-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 $\epsilon$-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 $\epsilon$-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 $\xi$ 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*

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

*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* $\pi^{\mathrm{b}}$ *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* $\xi_i$ *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*

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

*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* $\pi^{\mathrm{b}}$ *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

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

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

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

for some absolute constants $c_1, c_2 > 0$. Here $\zeta_h$ is a solution to the inequality $\sqrt{n}\epsilon^2 \ge c_0 G_h(\epsilon)$ w.r.t $\epsilon$, 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 $1-\delta$,

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

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 $1-\delta$,

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

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

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

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

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

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 $1-\delta/2$,

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

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 $1-\delta/2$,

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

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,

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

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

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

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

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

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

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

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. $G_h(\epsilon) \geq J_B(\epsilon, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}})$ and $G_h(\epsilon)/\epsilon^2$ non-increasing w.r.t $\epsilon$. Then, letting $\zeta_h$ be a solution to $\sqrt{n}\epsilon^2 \geq c_0 G_h(\epsilon)$ w.r.t $\epsilon$, where $c_0$ is an absolute constant. With probability at least $1-\delta$, we have that*

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

*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 $\zeta_h$ 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 $\sqrt{n}\epsilon^2 \geq c_0 G_h(\epsilon)$ w.r.t $\epsilon$.*

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

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

where the first inequality follows from Lemma H.6, the second inequality follows from the fact that $\mathcal{N}_{[]}(u_1, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}}) \geq \mathcal{N}_{[]}(u_2, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}})$ for $u_1 \leq u_2$. In the second inequality we assume without loss of generality that $\log(\mathcal{N}_{[]}(\epsilon^2/2, \overline{\mathcal{P}}_h(\epsilon), \|\cdot\|_{2,d^{\mathrm{b}}_{P^\star,h}})) \geq 4$. 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 $\epsilon$ 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 $\sqrt{n}\epsilon^2 \geq c_0 G_h(\epsilon)$ reduces to

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

which equivalents to

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

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 $\zeta_h$ 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 $\mathcal{M}(\epsilon) = \mathcal{N}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty)$)

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

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

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

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

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

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 $\delta$, 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 $\mathcal{M}'(\epsilon) = \mathcal{N}(\epsilon, \mathcal{G}_h, \|\cdot\|_\infty)$)

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

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,

$$\begin{aligned}
\Delta_h(P_h) &\le \sqrt{\frac{8(\Delta_h(P_h) + 4\xi) \log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}} \\
&\quad + 8\sqrt{\frac{\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n^3}} + \frac{8\log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{3n} + \frac{2}{n^2} \\
&\le \sqrt{\frac{8(\Delta_h(P_h) + 4\xi) \log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}} + \frac{c'_1 \log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n} \\
&\le \sqrt{\frac{8\Delta_h(P_h) \log(\mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n}} + \frac{c''_1 \log(c''_2 \mathcal{N}_{[]}(\epsilon, \mathcal{P}_{\mathrm{M}}, \|\cdot\|_{1,\infty})/\delta)}{n},
\end{aligned} \tag{H.6}$$

for some absolute constants $c'_1, c''_1, c''_2 > 0$, where in the last inequality we have applied the definition of $\xi$. 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 $\delta$, 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 $\epsilon$-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 $|f(X)| \le R$, then with probability at least $1-\delta$ the following inequality holds for all $f \in \mathcal{F}$,*

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

*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 $\epsilon \ge 0$ that*

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

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

*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 $\|f\|_\infty \le H$, we have that

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

Thus we can conclude that $\lambda^\star \le H/\sigma$. □

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

*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 $g(\lambda) \le 0$ for $\lambda \le 0$. Thus we have shown that $\lambda^\star \ge 0$. Also, for $\lambda \ge H$, due to $\|f\|_\infty \le H$, 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 $\lambda$. Thus we prove that $\lambda^\star \le H$. □