PESSIMISTIC MODEL-BASED OFFLINE REINFORCEMENT LEARNING UNDER PARTIAL COVERAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study model-based offline Reinforcement Learning with general function approximation without a full coverage assumption on the offline data distribution. We present an algorithm named *Constrained Pessimistic Policy Optimization (CPPO)* which leverages a general function class and uses a constraint over the model class to encode pessimism. Under the assumption that the ground truth model belongs to our function class (i.e., realizability in the function class), CPPO has a PAC guarantee with offline data only providing partial coverage, i.e., it can learn a policy that competes against any policy that is covered by the offline data. We then demonstrate that this algorithmic framework can be applied to many specialized Markov Decision Processes where additional structural assumptions can further refine the concept of partial coverage. Two notable examples are: (1) *low-rank MDP with representation learning* where the partial coverage condition is defined using a relative condition number measured by the unknown ground truth feature representation; (2) *factored MDP* where the partial coverage condition is defined using density ratio based concentrability coefficients associated with individual factors.

1 INTRODUCTION

Offline Reinforcement Learning (RL) is one of the important areas of RL where the learner is presented with a static dataset consisting of transition-related information (state, action, reward, and next state) collected by some behavior policy, and needs to learn purely from the offline data without any future online interaction with the environment. Offline RL is used in a number of applications where online random experimentation is costly or dangerous such as health care (Kosorok & Laber, 2019), digital marketing (Chen et al., 2019) and robotics (Levine et al., 2020).

The performance guarantees of offline RL often rely on two quantities: the coverage of the offline data and the property of the function approximation used in the algorithms. For instance, for the classic Fitted-Q-iteration (FQI) algorithm (Ernst et al., 2005; Munos & Szepesvári, 2008), it requires (a) full coverage in the offline data, i.e., $\max_{(s,a)} d^\pi(s,a)/\rho(s,a) < \infty$ for *any* stochastic policies π including history-dependent non-Markovian policies, where $d^\pi(s,a)$ is a state-action occupancy distribution of a policy π and $\rho(s,a)$ is an offline distribution, (b) realizability in a Q function class, i.e., the optimal Q function belongs to the function class, and (c) Bellman completeness, i.e., applying the Bellman operator on any function in the function class results in a new function that also belongs to the function class (see the first row in Table 1). Among these three assumptions, the full coverage and the Bellman completeness are particularly strong. The full coverage means that the behavior policy needs to be exploratory enough, although figuring out an exploratory policy itself is an extremely hard problem for large-scale MDPs. The Bellman completeness assumption does not have a monotonic property, i.e., even starting with a function class that originally permits Bellman completeness, slightly increasing the capacity of the function class could result in a new class that does not have Bellman completeness anymore. Thus, we aim to relax the assumptions on the offline data and the function class. Particularly, we are interested in the following question:

> Given a realizable function class and an offline distribution that only provides partial coverage, can we learn a policy that is able to compete with any policy that is covered by the offline distribution?

We study this question from a model-based learning perspective and provide an affirmative answer to the question. More specifically, different from FQI, we start with a realizable model class, i.e.,

Methods	Type	Coverage	Additional Structures
FQI (Munos & Szepesvári, 2008)	**F**	Full: $\max_{s,a} \frac{d^{\pi}(s,a)}{\rho(s,a)} < \infty, \forall \pi$	Bellman complete
Minimax Way (Uehara et al., 2020)	**F**	Full: $\max_{s,a} \frac{d^{\pi}(s,a)}{\rho(s,a)} < \infty, \forall \pi$	Realizability in density ratio
Duan et al. (2020)	**F**	Full: $\mathbb{E}_{s,a\sim\rho}\phi(s,a)\phi(s,a)^\top$ is PSD	Linear Bellman complete
Xie & Jiang (2020)	**F**	Full: $\max_{s,a,s'} \frac{P^\star(s'\mid s,a)}{\rho(s')} < \infty$	None
Liu et al. (2020)	**F**	Partial† : $\max_{s,a} \frac{d^{\pi^*}(s,a)}{\rho(s,a)} < \infty$	Bellman / Policy class complete
Rashidinejad et al. (2021)	**F**	Partial: $\max_{s,a} \frac{d^{\pi^*}(s,a)}{\rho(s,a)} < \infty$	Tabular MDP
Jin et al. (2020b); Zhang et al. (2021b)	**F**	Partial††: $\max_x \frac{x^\top \mathbb{E}_{s,a\sim d^{\pi^*}}\phi(s,a)(\phi(s,a))^\top x}{x^\top \mathbb{E}_{s,a\sim\rho}\phi(s,a)(\phi(s,a))^\top x} < \infty$	Linear MDP (Jin et al., 2020a)
Xie et al. (2021)	**F**	Partial: $\max_f \frac{\|f-\mathcal{T}f\|_2^2}{\|f-\mathcal{T}f\|_\mu^2} < \infty$	Bellman complete
Zanette et al. (2021)	**F**	Partial : $\max_x \frac{x^\top \mathbb{E}_{s,a\sim d^{\pi^*}}\phi(s,a)(\phi(s,a))^\top x}{x^\top \mathbb{E}_{s,a\sim\rho}\phi(s,a)(\phi(s,a))^\top x} < \infty$	Linear Bellman complete
Batch (Ross & Bagnell, 2012)	**B**	Full: $\max_{s,a} \frac{d^{\pi}(s,a)}{\rho(s,a)} < \infty, \forall \pi$	None
Milo (Chang et al., 2021)	**B**	Partial: $\max_x \frac{x^\top \mathbb{E}_{s,a\sim d^{\pi^*}}\phi(s,a)(\phi(s,a))^\top x}{x^\top \mathbb{E}_{s,a\sim\rho}\phi(s,a)(\phi(s,a))^\top x} < \infty$	KNR / GP
CPPO (Ours)	**B**	Partial†††: $\max_{s,a} \frac{d^{\pi^*}(s,a)}{\rho(s,a)} < \infty$	None
		Partial: $\max_x \frac{x^\top \mathbb{E}_{s,a\sim d^{\pi^*}}\phi(s,a)(\phi(s,a))^\top x}{x^\top \mathbb{E}_{s,a\sim\rho}\phi(s,a)(\phi(s,a))^\top x} < \infty$	Linear MDP /KNR / GP
		Partial: $\max_x \frac{x^\top \mathbb{E}_{s,a\sim d^{\pi^*}}\phi^\star(s,a)(\phi^\star(s,a))^\top x}{x^\top \mathbb{E}_{s,a\sim\rho}\phi^\star(s,a)(\phi^\star(s,a))^\top x} < \infty$	Low-rank MDP (unknown ϕ^*)

Table 1: Comparison among existing works regarding their type, coverage, and additional structural assumptions on the function class or MDPs. Type **F** means model-free and type **B** means model-based. Partial coverage means [2]that the offline distribution ρ covers a state-action distribution induced by a comparator policy π^*. † means it assumes an accurate density estimator for $\rho(s,a)$. †† means although the analysis in Jin et al. (2020a) is done under the full coverage for linear MDPs, based on the argument (Zhang et al., 2021b), we can show the algorithm has the PAC guarantee under partial coverage in terms of the relative condition number for linear MDPs. † † † means that we can refine it to a more adaptive quantity for the model class. Refer to Definition 1. All the methods in the table require realizability in the function class.

the ground truth transition falls into the model class. We further abandon the strong full coverage assumption, and instead, assume partial coverage which means the offline data distribution only covers a state-action distribution of *some high-quality comparator policy* π^* (π^* is not necessarily the optimal policy, and π^* could be non-Markovian), i.e., $\max_{s,a} d^{\pi^*}(s,a)/\rho(s,a) < \infty$, We design an algorithm — Constrained Pessimistic Policy Optimization (CPPO), which can learn a policy that is as good as any comparator policy π^* that is covered by the offline data.

While one could assume density ratio based concentrability coefficient ($\max_{s,a} d^{\pi^*}(s,a)/\rho(s,a)$) to be under control for small size MDPs, in large-scale MDPs (e.g. continuous state space), the density ratio could quickly become an extremely large quantity which makes the performance guarantee vacuous. Under additional structural assumptions of the underlying MDPs, we can refine the density ratio based concentrability coefficient to more natural and tighter quantities. Notably, we consider the offline representation learning setting where the underlying MDPs permit a low-rank structure (unlikely linear MDPs (Jin et al., 2020a; Yang & Wang, 2020), we do not assume the ground truth state-action feature representation ϕ is known, and instead we need to learn ϕ) and we show that we can refine the density ratio to a relative condition number that is defined using the *unknown* true state-action feature representation ϕ. Intuitively this means that as long as there exists a high-quality comparator policy that only visits the subspace (defined using the true representation ϕ) that is covered by the offline data, CPPO can compete against such a policy, *even without knowing the true* ϕ. Such bounded relative condition number assumption is weaker than the bounded density ratio assumption. While the concept of relative condition number was originally introduced in the online RL setting (e.g., Agarwal et al. (2020c;a) with a known linear feature ϕ), and later was introduced in offline RL (Zhang et al. (2021b); Chang et al. (2021)), these prior works *all* rely on the fact that the feature representation ϕ is known to the learner a priori (see Table 1 for the comparison).

Our contributions. Our contributions are two folds, which we summarize below:

1. We show that in the model-based setting, realizability and partial coverage (measured in density ratio based concentrability coefficient) is enough to learn a high-quality comparator policy (Theorem 1). Notably, this result holds for *any MDPs with realizable model classes*. This is in sharp contrast to the state-of-art provable model-free offline RL results: see Table 1 on page 2 for detailed comparisons to prior works.

[2]Strictly speaking, in Jin et al. (2020b); Rashidinejad et al. (2021), a comparator policy is restricted to the optimal policy. In Chang et al. (2021); Zanette et al. (2021); Xie et al. (2021) and CPPO, a comparator policy can be any policy.

2. Under additional structural assumptions (e.g., kernelized nonlinear regulators (KNRs) (Kakade et al., 2020), linear MDPs (Yang & Wang, 2020), linear mixture MDPs (Ayoub et al., 2020), low-rank MDPs, factored MDPs (Kearns & Koller, 1999)), we show that we can further refine the density ratio based concentrability coefficients. Two notable settings: (a) for offline representation learning in low-rank MDPs, the density ratio concentrability coefficient is refined to be a relative condition number under the true (but unknown) representation (Theorem 2); (b) for factored MDPs, the concentrability coefficient is refined using the density ratios associated to individual factors (Theorem 3).

While we focus on the model-based setting, it is worth noting that realizability in the model-based setting is usually considered stronger than the one in the model-free setting. However, it is unknown whether or not partial coverage and realizability in Q class can ensure efficient learning. Indeed, it is still an open question if efficient learning is possible even with a full coverage condition (measured by the density ratio concentrability coefficient) and realizability. Our theoretical results provide a sharp contrast between model-based and model-free approaches in offline RL.

2 Related work

We discuss two families of related works: offline RL and representation learning in RL.

Offline RL. Insufficient coverage of the dataset due to the lack of online exploration is known as the main challenge in offline RL (Wang et al., 2020). To deal with this problem, a number of methods have been recently proposed from both model-free (Wu et al., 2019; Touati et al., 2020; Kumar et al., 2020; Liu et al., 2020; Rezaeifar et al., 2021; Fujimoto et al., 2019; Fakoor et al., 2021; Ghasemipour et al., 2021; Buckman et al., 2020) and model-based perspectives (Yu et al., 2020; Kidambi et al., 2020; Matsushima et al., 2020; Yin et al., 2021). More or less, their methods rely on the idea of pessimism and its variants in the sense that the learned policy can avoid uncertain regions not covered by offline data. As a theoretical side, Munos & Szepesvári (2008); Duan et al. (2020; 2021); Fan et al. (2020) proved FQI has a PAC (probably approximately correct) guarantee under realizability, the global coverage, and Bellman completeness. Other offline model-free RL methods such as minimax offline RL methods also require realizability and the global coverage (Chen & Jiang, 2019; Antos et al., 2008; Uehara et al., 2021; Duan et al., 2021; Zhang et al., 2020; Nachum et al., 2019). Recently, by being inspired by aforementioned the pessimism idea, Jin et al. (2020a); Rajaraman et al. (2020) showed that FQI with an additional pessimistic bonus (penalty) term can weaken the assumption from the global coverage to partial coverage. Comparing to their works, our analysis focuses on a model-based method. The offline model-based method is known to have a PAC guarantee under the realizability and the global coverage (Ross & Bagnell, 2012; Chen & Jiang, 2019). As the most closely related work, Chang et al. (2021) proved a model-based method with an additional penalty term can weaken the assumption from the global coverage to the partial coverage for structured MDPs such as KNRs and Gaussian Processes models (Deisenroth & Rasmussen, 2011). In this work, we consider *arbitrary* MDPs with a realizable model class and aim for PAC bounds under a partial coverage condition.

Representation learning. Representation learning for low-rank MDPs (ground truth feature representation is unknown) in online learning is studied from a model-based perspective (Agarwal et al., 2020b) and model-free perspective (Modi et al., 2021). In the online setting, Zhang et al. (2021a); Papini et al. (2021) also study representation learning under different model assumptions. Comparing with these works, since our setting is offline, the algorithm and analysis are totally different.

In the offline setting, Ni et al. (2021) study dimensionality reduction in a given kernel space, and Hao et al. (2021) study feature selection in sparse linear MDPs. Their focus is different as they do not study PAC guarantees under partial coverage. The detailed comparison is given in Section C.

3 Preliminaries

We consider an episodic finite horizon Markov Decision process (MDP) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, \gamma, r, d_0\}$ where $P : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition, $r : \mathcal{S}\times\mathcal{A} \to [0, 1]$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps from state to distribution over actions. Given a policy π and a transition distribution P, V^π_P denotes the expected cumulative reward of π under P, d_0 and r. Similarly, $Q^\pi_P : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $A^\pi_P : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are a Q-function and advantage-function under P and π. Given a transition P, we denote $\pi(P)$

as the optimal policy associated with model P. We also denote $d_P^\pi \subset \Delta(\mathcal{S} \times \mathcal{A})$ as the average state-action distribution of π under the transition model P, i.e, $d_P^\pi = (1-\gamma)\sum_{t=0}^\infty \gamma^t d_{P,t}^\pi$, where $d_{P,t}^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ is the distribution of (s_t, a_t) under π and P at a time-step t. We denote the true transition distribution as $P^\star$, which we do not know in advance. For simplicity, we suppose r is known. The extension to the unknown case is straightforward.

In the offline RL setting, we have an offline distribution $\rho \in \Delta(\mathcal{S} \times \mathcal{A})$, and an offline dataset $\mathcal{D} = \{s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)}\}_{i=1}^n$ which is sampled in the following way: $s, a \sim \rho, r = r(s,a), s' \sim P^\star(\cdot|s,a)$ where $P^\star$ denotes the ground truth transition model. Our main goal is obtaining $\pi(P^\star) = \arg\max_{\pi\in\Pi} V_{P^\star}^\pi$ from this offline dataset without any further interaction with the environment. We often denote $\mathbb{E}_\mathcal{D}[f(s,a,s')] = 1/n \sum_{(s,a,s')\in\mathcal{D}} f(s,a,s')$. Our goal is to construct an offline RL algorithm Alg, which maps from $\mathcal{D}$ to Π, where $\Pi \subseteq \mathcal{S} \to \Delta(\mathcal{A})$ is a policy class so that the suboptimality gap $V_{P^\star}^{\pi^*} - V_{P^\star}^{\mathrm{Alg}(\mathcal{D})}$ for any comparator policy $\pi^* \in \Pi$ is minimized. In this work, we in default consider Π to be the unrestricted Markovian policy class unless otherwise stated. Hereafter, $c, c_1, c_2, \cdots$ are always universal constants.

Partial coverage. Throughout this work, we do not assume ρ has global coverage. The global coverage in this work means that the density ratio based concentrability coefficient $d_{P^\star}^\pi(s,a)/\rho(s,a)$ is upper-bounded by some constant $C \in \mathbb{R}^+$ for all polices $\pi \in \Pi$, or the feature covariance matrix corresponding to the offline distribution $\mathbb{E}_{s,a\sim\rho}\phi(s,a)\phi(s,a)^\top$ ($\phi \in \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a feature representation) is full rank and has a non-zero minimum eigenvalue, which are commonly used assumptions in offline RL (Munos, 2005; Antos et al., 2008; Chen & Jiang, 2019; Duan et al., 2020). Under the full coverage, they show the output policy can compete with the globally optimal policy $\pi(P^\star)$. However, this assumption may not be true in practice as computing an exploratory policy itself is a challenging task for large-scale RL problems. Instead, we are interested in the partial coverage setting such as $d_{P^\star}^{\pi^*}(s,a)/\rho(s,a) \leq C$, which means the state-action occupancy measure under some comparator policy π^* is covered by the offline dataset. We want to design an algorithm that can compete against *any* policy π^* that is covered by the offline data. This assumption is much weaker than the global coverage. We can further reduce the density-ratio concentrability coefficient to a tighter quantity that is adaptive to the function class as we will show in Section 4.

4 Pessimistic Model-based Offline RL

We first introduce a general model-based algorithm that has a PAC guarantee of the suboptimality gap under partial coverage defined with a newly introduced concentrability coefficient. The algorithm takes a realizable model class as input and outputs a policy that is as good as any comparator policy that is covered by the offline data in the sense of the bounded concentrability coefficient.

Our algorithm, *Constrained Pessimistic Policy Optimization (CPPO)* (Algorithm 1), takes a realizable hypothesis class $\mathcal{M}$ (with $P^\star \in \mathcal{M}$) consisting of $|\mathcal{M}|$ candidate models as input, computes the maximum likelihood estimator (MLE) $\widehat{P}_{\mathrm{MLE}}$ using the given offline data $\mathcal{D} = \{s,a,s'\}$. It then forms a min-max objective subject to a constraint. The min-max objective introduces pessimism via searching for the least favorable model P (in terms of its policy's value V_P^π) that is feasible with respect to the constraint. We can also express the constrained optimization procedure using a version space $\mathcal{M}_\mathcal{D}$ and a policy optimization procedure defined below:

$$\max_{\pi\in\Pi} \min_{P\in\mathcal{M}_\mathcal{D}} V_P^\pi, \quad \text{where } \mathcal{M}_\mathcal{D} = \left\{P \mid P \in \mathcal{M}, \mathbb{E}_\mathcal{D}\left[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot|s,a), P(\cdot|s,a))^2\right] \leq \xi\right\}, \quad (1)$$

where $\mathrm{TV}(P_1, P_2)$ is a total variation (TV) distance between two distributions P_1 and P_2. The version space $\mathcal{M}_\mathcal{D}$ contains models that are not far away from $\widehat{P}_{\mathrm{MLE}}$ in terms of the average TV distance under $\mathcal{D}$. The version space is constructed such that with high probability $P^\star \in \mathcal{M}_\mathcal{D}$.

Below we state the algorithm's performance guarantee. Assuming for now that $P^\star \in \mathcal{M}_\mathcal{D}$ holds with high probability, then, $\hat{V}^\pi := \min_{P\in\mathcal{M}_\mathcal{D}} V_P^\pi$ is a pessimistic policy evaluation estimator, which satisfies $\hat{V}^\pi \leq V_{P^\star}^\pi$ for all $\pi \in \Pi$. Using the idea of pessimism, we have the following observation:

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = V_{P^\star}^{\pi^*} - \hat{V}^{\pi^*} + \hat{V}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - \hat{V}^{\pi^*} + \hat{V}^{\hat{\pi}} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - \hat{V}^{\pi^*},$$

where the first inequality uses $\hat{\pi} = \arg\max_{\pi\in\Pi} \hat{V}^\pi$ and the second inequality uses $\hat{V}^\pi \leq V_{P^\star}^\pi$ for all $\pi \in \Pi$. Thus, the final error only incurs the policy evaluation error for the comparator policy π^*, which leads to the error only depending on the concentrability coefficient for the comparator policy.

Algorithm 1 Constrained Pessimistic Policy Optimization (CPPO)

1: **Require**: Models $\mathcal{M}$, dataset $\mathcal{D}$, parameter ξ, policy class Π (note Π could be unrestricted)
2: Obtain the estimator $\hat{P}_{\mathrm{MLE}}$ by MLE: $\widehat{P}_{\mathrm{MLE}} = \arg\max_{P\in\mathcal{M}} \mathbb{E}_{\mathcal{D}}[\ln P(s' \mid s,a)]$.
3: Constrained policy optimization:

$$\hat{\pi} = \arg\max_{\pi\in\Pi} \min_{P\in\mathcal{M}} V_P^{\pi},\ \text{s.t.}, \mathbb{E}_{\mathcal{D}}\left[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot|s,a), P(\cdot|s,a))^2\right] \le \xi.$$

4: **Return** $\hat{\pi}$

We define the following new concentrability coefficient that uses the model class $\mathcal{M}$:

Definition 1 (Model-based Concentrability Coefficient)**.** *For a comparator policy π^*, we define the concentrability coefficient $C^{\dagger}_{\pi^*}$ as follows:*

$$C^{\dagger}_{\pi^*} = \sup_{P'\in\mathcal{M}} \frac{\mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[\mathrm{TV}(P'(\cdot|s,a),P(\cdot|s,a))^2]}{\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P'(\cdot|s,a),P(\cdot|s,a))^2]}.$$

The following theorem shows CPPO learns a policy that competes against π^* when $C^{\dagger}_{\pi^*} < \infty$.

Theorem 1 (PAC Bound for CPPO with general function class)**.** *Assume $P^\star \in \mathcal{M}$. We set $\xi = c_1 \frac{\ln(c_2|\mathcal{M}|/\delta)}{n}$. Then, with probability $1-\delta$, for any comparator policy $\pi^* \in \Pi$ (Π can be the unrestricted policy class containing non-Markovian policies),*

$$V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} \le c_3(1-\gamma)^{-2}\sqrt{\frac{C^{\dagger}_{\pi^*}\ln(c_2|\mathcal{M}|/\delta)}{n}}.$$

To the best of our knowledge, this is the *first* algorithm that achieves a PAC guarantee for *any MDPs* under the partial coverage assumption $C^{\dagger}_{\pi^*} < \infty$ with only a realizable hypothesis class. We emphasize that the inequality in the above *uniformly* holds for *all* policies with probability $1-\delta$ including *history-dependent non-Markovian policies* (see Remark 4). Note that the ability to compete against non-Markovian policies in offline RL is meaningful when the offline data does not cover the optimal policy $\pi^\star$ (i.e., there could be a high-quality history-dependent policy that is covered by the offline data against which we want to compete). In model-free approaches, this type of result generally cannot be obtained. Indeed, the model-free approach from Xie et al. (2021) requires Π to be a restricted Markovian policy class, since their bound contains $\mathrm{poly}(\ln(|\Pi|))$ dependence. For the detailed discussion, refer to Remark 2.

The quantity $C^{\dagger}_{\pi^*}$ adaptively captures the discrepancy between the offline data and the state-action occupancy measure under a comparator policy π^* depending on the model class $\mathcal{M}$. For example, $C^{\dagger}_{\pi^*}$ can be reduced to a relative condition number in KNRs. Besides, it is always upper bounded by the density ratio based concentrability coefficient:

$$C_{\pi^*,\infty} := \sup_{(s,a)} \frac{d^{\pi^*}_{P^\star}(s,a)}{\rho(s,a)}.$$

Theorem 1 consider the case where the hypothesis class $\mathcal{M}$ is finite. When the hypothesis class is infinite, we can still obtain the PAC guarantee by utilizing the generalized result in Section A for any realizable model class with valid statistical complexity (e.g., localized Rademacher complexity).

Prior works that achieve PAC guarantees with only realizable model classes rely on much stronger global coverage $\sup_\pi C_{\pi,\infty} < \infty$(Chen & Jiang, 2019). Even when the comparator policy is the optimal policy $\pi(P^\star)$, the partial coverage condition $C_{\pi(P^\star),\infty} < \infty$ is weaker. Existing pessimistic model-based algorithms and their theoretical results (Chang et al., 2021) often assume that a *point-wise* model uncertainty measure is given as a by-product of model fitting, which limits the applicability to special linear models such as KNRs/GPs. CPPO can work for any MDPs with the realizable function class having a valid statistical complexity such that the MLE properly works.

Remark 1. *With a slight modification of the proof, we can also show Theorem 1 by substituting $C^{\dagger}_{\pi^*}$ with* L2 *density-ratio based concentrability coefficient $C_{\pi^*,2} = \mathbb{E}_{(s,a)\sim\rho}[(d^{\pi^*}_{P^\star}(s,a)/\rho(s,a))^2]$. We have $C_{\pi^*,2} \le C_{\pi^*,\infty}$ while $C_{\pi^*,2}$ and $C^{\dagger}_{\pi^*}$ are generally not comparable. The detail is in Section D.*

Remark 2 (Comparison to the model-free approach from Xie et al. (2021); Zanette et al. (2021))**.** *Xie et al. (2021) study the model-free setting where the function class $\mathcal{Q}$ models Q functions*

assumed to be Bellman complete for any Markovian policy in Π. *While directly comparing model-based approaches to model-free approaches is hard as they use different inductive biases in function classes, we can leverage the approach from Chen & Jiang (2019, Corollary 6) to convert a model class* $\mathcal{M}$ *to a pair of* $\mathcal{Q}$ *and* Π *class. Specifically, we can convert a model class* $\mathcal{M}$ *to a pair of* $\mathcal{Q}$ *class and* Π *class such that* $\mathcal{Q}$ *will be realizable and also Bellman complete with respect to all* $\pi \in \Pi$. *After such conversion from the model-based setting to the model-free setting, running the algorithm from Xie et al. (2021) using* $\mathcal{Q}$ *and* Π *achieves* $V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = \sqrt{C^\diamond \ln(|\mathcal{M}||\Pi|/n)}, \forall \pi^* \in \Pi$, *where* $C^\diamond$ *is some concentrability coefficient. For the detailed derivation, we refer readers to Appendix E. Since the suboptimality gap from such conversion incurs* $\log|\Pi|$, *a policy class* Π *cannot be too large. Especially, unlike our results, it cannot take the unrestricted policy class as* Π. *This restriction* cannot *be fixed even if we use natural policy gradient (NPG) algorithms unless models have special structures (Xie et al., 2021; Zanette et al., 2021). The details are given in Section E.*

Thus, our theorem indicates two advantages of model-based approaches: (1) realizability in function class is enough to ensure a PAC guarantee under a partial coverage condition, (2) it can compete against a larger pool of candidate policies including history-dependent non-Markovian policies.

5 Examples with Refined Concentrability Coefficients

In the previous section, our results apply to any MDP as long as its true transition belongs to a function class $\mathcal{M}$. In this section, we consider several concrete MDPs with additional structural conditions. We show that by leveraging the additional structural conditions, we can refine the model-based concentrability coefficient to more natural quantities. The examples that we discuss here are: linear mixture MDPs which generalize linear MDPs from Yang & Wang (2020) and tabular MDPs, KNRs which generalize LQRs, low-rank MDPs, and factored MDPs.

Before proceeding, we clarify CPPO cannot capture linear MDPs in Jin et al. (2020a) that is different from the one (Yang & Wang, 2020) we use, and linear Bellman-complete MDPs (Duan et al., 2020) without any modification since MLE is no longer applicable to them. However, other objective functions for learning models could be applied to these models (e.g., see the nonparametric model-based learning approach from Lykouris et al. (2021); Neu & Pike-Burke (2020) in the online setting).

5.1 Tabular MDPs and Linear Mixture MDPs

Tabular MDPs Tabular MDPs are MDPs, where the state and action spaces are finite. Although the corresponding hypothesis class for tabular MDPs is infinite, we can still run MLE, that is, estimating $P^\star$ by the empirical distribution. Then, Algorithm 1 has the following guarantee.

Corollary 1 (PAC bound for tabular MDP)**.** *We set* $\xi = c_1 \frac{|\mathcal{S}|^2|\mathcal{A}|\ln(n|\mathcal{S}|\mathcal{A}|c_2/\delta)}{n}$. *Then with probability* $1-\delta$, *for all* $\pi^* \in \Pi$,

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq c_3(1-\gamma)^{-2}\left\{\sqrt{\frac{C_{\pi^*,\infty}|\mathcal{S}|^2|\mathcal{A}|\ln(n|\mathcal{S}||\mathcal{A}|c_4/\delta)}{n}}\right\}.$$

Here, for tabular MDPs with $\mathcal{M} = \{P : P(\cdot|s,a) \in \Delta(\mathcal{S}), \forall s,a\}$, the model-based concentrability coefficient in Definition 1 is equal to the density ratio based concentrability coefficient $C_{\pi^*,\infty}$.

Linear mixture MDPs The definition is as follows (Ayoub et al., 2020; Modi et al., 2020).

Definition 2 (Linear mixture MDPs)**.** *Given a feature vector* $\psi : (\mathcal{S}, \mathcal{A}, \mathcal{S}) \to \mathbb{R}^d$, *a linear mixture MDP is an MDP where the ground truth transition is* $P^\star(s'|s,a) := \theta^{\star\top}\psi(s,a,s'), \theta^\star \in \mathbb{R}^d$.

By setting, $\psi(s,a,s') = \mu(s') \bigotimes \phi(s,a)$ ($\otimes$ denotes the Kronecker product), linear mixture MDPs include the following linear MDPs (Yang & Wang, 2020):

Definition 3 (Linear MDPs)**.** *Linear MDP has* $P^\star(s'|s,a) := \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} M_{ij}^\star \mu_i(s')\phi_j(s,a)$ *with* $\mu : \mathcal{S} \to \mathbb{R}^{d_1}$ *and* $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{d_2}$ *are known features, and* $M^\star \subset \mathbb{R}^{d_1\times d_2}$.

We use CPPO to learn on linear mixture MDPs. The corresponding $\mathcal{M}$ is

$$\mathcal{M}_{\text{Mix}} = \left\{\theta^\top\psi(s,a,s') \mid \theta \in \Theta \subset \mathbb{R}^d, \int \theta^\top\psi(s,a,s')\mathrm{d}(s') = 1 \quad \forall(s,a)\right\}.$$

Given a function $V : \mathcal{S} \to \mathbb{R}$, define the state-action feature indexed by V as $\psi_V(s,a) := \int \psi(s,a,s')V(s')\mathrm{d}(s')$, we have the following PAC guarantee.

Corollary 2 (PAC bound for linear mixture MDPs)**.** *Suppose* $\inf_{s,a,s'} P^\star(s' \mid s,a) \geq c_3 > 0$, $\Theta = \{\theta : \|\theta\|_2 \leq R\}$, $\|\psi_V(s,a)\|_2 \leq 1, \forall V \in \mathcal{S} \to [0,1]$ *and* $P^\star \in \mathcal{M}_{Mix}$. *We set* $\xi = c_1 d \ln^2(c_2 nR/\delta)/n$. *Then, with probability* $1-\delta$, *for any* π^* *in* Π *(again* Π *can be the unrestricted policy class), CPPO outputs a policy* $\hat{\pi}$ *such that:*

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq c_4(1-\gamma)^{-2}\sqrt{\min(dC_{\pi^*}^\dagger, d^2\bar{C}_{\pi^*,\mathrm{mix}})\frac{\ln^2(c_5 nR/\delta)}{n}}, \tag{2}$$

where the concentrability coefficient $\bar{C}_{\pi^*,\mathrm{mix}}$ *is defined as:*

$$\bar{C}_{\pi^*,\mathrm{mix}} := \sup_{P\in\mathcal{Z}_{P^\star}} \sup_{x\in\mathbb{R}^d} \left(\frac{x^\top \Sigma_{\pi^*,\psi_{V_P^{\pi^*}}} x}{x^\top \Sigma_{\rho,\psi_{V_P^{\pi^*}}} x}\right)$$

with the localized class $\mathcal{Z}_{P^\star} := \{P : \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \leq \xi\}$, $\Sigma_{\rho,\psi_{V_P^{\pi^*}}} = \mathbb{E}_{(s,a)\sim\rho}[\psi_{V_P^{\pi^*}}(s,a)\psi_{V_P^{\pi^*}}(s,a)^\top]$, *and* $\Sigma_{\pi^*,\psi_{V_P^{\pi^*}}} = \mathbb{E}_{s,a\sim d_{P^\star}^{\pi^*}}[\psi_{V_P^{\pi^*}}(s,a)\psi_{V_P^{\pi^*}}(s,a)^\top]$.

When specializing to linear MDPs, the above bound still holds with $\bar{C}_{\pi^*,\mathrm{mix}}$ *being replaced by the relative condition number* $\bar{C}_{\pi^*}$*:*

$$\bar{C}_{\pi^*} := \sup_{x\in\mathbb{R}^d} \frac{x^T\Sigma_{\pi^*}x}{x^\top\Sigma_\rho, x}, \text{ where } \Sigma_\rho = \mathbb{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top],\ \Sigma_{\pi^*} = \mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}[\phi(s,a)\phi(s,a)^\top].$$

This is the first PAC-guarantee result in the offline setting under partial coverage $\bar{C}_{\pi^*,\mathrm{mix}} < \infty$ for linear mixture MDPs. $\bar{C}_{\pi^*,\mathrm{mix}}$ is a newly-introduced concentrabiliy coefficient for linear mixture MDPs. This coefficient is measured on the integrated feature vectors $\phi_V(s,a)$ for $V : S \to [0,1]$. The class of V is localized, i.e., we consider state-value functions $V_P^{\pi^*}(s)$ for all P centered around $P^\star$ (i.e., $P \in \mathcal{Z}_{P^\star}$). Such localization property ensures that $\bar{C}_{\pi^*,\mathrm{mix}} \leq C_{\pi^\star}^\dagger$ (see Lemma 10 in Section G).

Note that these relative condition number based quantifiers are always tighter than the density ratio based concentrability coefficients (i.e., $\max\{\bar{C}_{\pi^*}, \bar{C}_{\pi^*,\mathrm{mix}}\} \leq C_{\pi^*,\infty}$). For the special case where $\phi(s,a)$ is a one-hot encoding vector, then they are reduced to the density ratio based concentrability coefficient. In a non-tabular setting, even if when the density ratio is infinite, the relative condition number can be still finite. Intuitively, the bounded relative condition number implies that the offline data covers the subspace that the comparator policy π^* visits.

We remark $P^\star(s' \mid s,a) \geq c_3 > 0$ in Corollary 2 is a technical condition to calculate the entropy integral of the hypothesis class easily. It can be potentially discarded by a more careful argument following (van de Geer, 2000, Chapter 7). The norm assumption $\|\psi_V(s,a)\|_2 < 1$ is also commonly assumed in the online setting (Zhou et al., 2021).

5.2 Kernelized Nolinear Regulators

We consider the example of KNRs in this section. A kernelized Nonlinear Regulator (KNR) (Kakade et al., 2020) is a model where the ground truth transition $P^\star(s'|s,a)$ is defined as $s' = W^\star\phi(s,a)+\epsilon$, $\epsilon \sim \mathcal{N}(0,\zeta^2\mathbf{I})$, with $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ being a possibly nonlinear feature mapping. We denote the corresponding model on W by $P(W)$. We can apply Algorithm 1 and obtain its guarantee. Especially, since $\mathrm{TV}(P(W)(\cdot \mid s,a), P(W^\star)(\cdot \mid s,a))^2 = \Theta(\|(W-W^\star)\phi(s,a)\|_2^2)$ (Devroye et al., 2018), $C_{\pi^*}^\dagger$ is upper-bounded by the relative condition number $\bar{C}_{\pi^*}$.

Then, we can also recover the result of Chang et al. (2021) which proposes a reward penalty-based pessimistic offline RL algorithm. The detail is given in Section B. In summary, we can show

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq c_1(1-\gamma)^{-2}\min(d^{1/2}, \bar{R})\sqrt{\bar{R}}\sqrt{\tfrac{d_S\bar{C}_{\pi^*}\ln(1+n)}{n}},.$$

where $\bar{R} := \mathrm{rank}[\Sigma_\rho]\{\mathrm{rank}[\Sigma_\rho] + \ln(c_2/\delta)\}$.

This implies CPPO can learn a policy that can compete against π^* with partial coverage $\bar{C}_{\pi^*} < \infty$. Note that the condition $\bar{C}_{\pi^*} < \infty$ does not require Σ_ρ to be full-rank. Also the bound uses $\mathrm{rank}[\Sigma_\rho]$ instead of d, which means that our bound is distribution dependent and is still valid even when $d = \infty$ as long as the offline data only concentrate on a low-dimensional subspace. Regarding the formal argument with $d = \infty$ in RKHS models, refer to Appendix B.1.

5.3 LOW-RANK MDPS WITH REPRESENTATION LEARNING

We consider the representation learning in offline RL. Following FLAMBE (Agarwal et al., 2020b), we study low-rank MDPs but in the offline setting. Note that low-rank MDPs here are a more generalized model of the aforementioned linear MDPs (Yang & Wang, 2020) since the true feature representation $\phi^\star$ in a low-rank MDP is unknown.

Definition 4 (Low rank MDPs)**.** *The ground-truth model $P^\star$ admits a low rank decomposition with a dimension d if there exists two embedding functions $\mu^* : \mathcal{S} \to \mathbb{R}^d, \phi^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ s.t. $P^\star(s' \mid s, a) = \mu^*(s')^\top \phi^*(s, a)$. Neither μ^* nor ϕ^* is known to the learner.*

One interesting special case of a low-rank MDP is the following latent variable model (see Agarwal et al. (2020b) for more details).

Definition 5 (Latent variable models)**.** *There exists a latent space $\mathcal{Z}$ along with functions $\mu^* : \mathcal{Z} \to \Delta(\mathcal{S})$ and $\phi^* : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{Z})$ s.t. $P^\star(\cdot \mid s, a) = \sum_{z \in \mathcal{Z}} \mu^*(\cdot \mid z)\phi^*(z \mid s, a)$.*

To tackle representation learning under partial coverage on low-rank MDPs, we setup function classes as follows: given two function classes $\Psi \subset \mathcal{S} \to \mathbb{R}^d, \Phi \subset \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ (both are realizable in the sense that $\mu^* \in \Psi$ and $\phi^* \in \Phi$), we consider a hypothesis class $\{\mu(s')^\top \phi(s, a); \mu \in \Psi, \phi \in \Phi\}$. Then, CPPO (Algorithm 1) and Theorem 1 still work under this setting.

Here we show that by leveraging the low-rankness, we can refine the concentrability coefficient to a relative condition number defined by the unknown true representation ϕ^*. We emphasize that this does not depend on the other features. Particularly, given a comparator policy π^*, we define $\bar{C}_{\pi^*,\phi^\star}$:

$$\bar{C}_{\pi^*,\phi^\star} = \sup_{x \in \mathbb{R}^d} \frac{x^\top \Sigma_{\pi^*} x}{x^\top \Sigma_\rho x}, \quad \Sigma_{\pi^*} := \mathbb{E}_{s,a \sim d^{\pi^*}_{P^\star}} \phi^*(s, a)\phi^*(s, a)^\top, \quad \Sigma_\rho := \mathbb{E}_{s,a \sim \rho} \phi^*(s, a)\phi^*(s, a)^\top.$$

We can show CPPO learns a policy that can compete against π^* such that $\bar{C}_{\pi^*,\phi^\star} < \infty$.

Theorem 2 (PAC bound for low-rank MDP)**.** *We set $\xi = c_1 \frac{\ln(|\Phi||\Psi|c_2/\delta)}{n}$. Suppose (a): $\|\phi(s, a)\|_2 \leq 1, \forall (s, a) \in \mathcal{S} \times \mathcal{A}, \forall \phi \in \Phi$, $\int \mu(s')^\top \phi(s, a) \mathrm{d}(s') = 1$ and $\int \|\mu(s)\|_2 \mathrm{d}s \leq \sqrt{d}, \forall \mu \in \Psi, \phi \in \Phi$, (b) $\rho(s, a) = d^{\pi_b}_{P^\star}(s, a)$, (c) $P^\star(s'|s, a) = \mu^*(s')^\top \phi^*(s, a)$ for some $\mu^* \in \Psi, \phi^* \in \Phi$. With probability at least $1 - \delta$, for all $\pi^* \in \Pi$ (again Π can be an unrestricted policy class), CPPO (Algorithm 1) finds $\hat{\pi}$ such that:*

$$V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} \leq c_3 \sqrt{\bar{C}_{\pi^*,\phi^\star} \omega_{\pi^*} \mathrm{rank}(\Sigma_\rho) \frac{\ln(|\Psi||\Phi|c_4/\delta)}{(1-\gamma)^4 n}}, \; \omega_{\pi^*} = \left(\max_{(s,a)} \frac{\pi^*(a|s)}{\pi_b(a|s)} \right) \tag{3}$$

To the best of our knowledge, this is the first established PAC result under the partial coverage condition $\bar{C}_{\pi^*,\phi^\star} < \infty, \omega_{\pi^*} < \infty$ for low-rank MDPs in the offline setting, where the covariance matrices Σ_ρ and Σ_{π^*} are defined with respect to the true $\phi^\star$, and the offline covariance matrix Σ_ρ is not necessarily full rank. We also emphasize that our bound in Theorem 2 is distribution dependent, i.e., it depends on $\mathrm{rank}(\Sigma_\rho)$ rather than the exact rank d. Note that $\mathrm{rank}(\Sigma_\rho) \leq d$, and $\mathrm{rank}(\Sigma_\rho)$ could be much smaller than d when the offline distribution only concentrates on a low-dimensional subspace (defined using ϕ^*).

Remark 3 (More discussion of assumptions in Theorem 2)**.** *(a) is a regularity condition as assumed in FLAMBE. For (b), ρ can be the stationary distribution regarding π_b and $P^\star$. While we still need the density ratio of the comparator policy and the behavior policy ω_{π^*} is bounded to obtain the PAC guarantee, this is much weaker than $\max_{s,a} d_{\pi^*}(s, a)/\rho(s, a)$ is bounded,*

5.4 FACTORED MDPS

The last example we include is the factored MDP (Kearns & Koller, 1999) defined as follows:

Definition 6 (Factored MDPs)**.** *Let $d \in \mathbb{N}^+$ and $\mathcal{O}$ being a small finite set. The state space $\mathcal{S} = \mathcal{O}^d$, and for each state s, we denote $s[i] \in \mathcal{O}$ as the i-th variable of the state s. For each $i \in [1, \cdots, d]$, the parents of i, $\mathrm{pa}_i \subset [1, \cdots, d]$, is the subset of state variables that directly influences i, i.e., the transition is defined as follows:*

$$\forall s, a, s' : P^\star(s'|s, a) = \textstyle\prod_{i=1}^d P^\star_i(s'[i]|s[\mathrm{pa}_i], a).$$

We will denote $\mathcal{S}_i = \mathcal{O}^{|\mathrm{pa}_i|}$, and given $s \in \mathcal{S}$, we will have $s[\mathrm{pa}_i] \in \mathcal{S}_i$

Due to the factorization, the transition operator $P^\star$ can be described with $L := \sum_{i=1}^d |\mathcal{A}||\mathcal{O}|^{1+|\mathrm{pa}_i|}$ many parameters. In contrast, the non-factored transition will need $O(|\mathcal{O}|^d)$ parameters. When $|\mathrm{pa}_i| \ll d \, \forall i$, it is expected that we can learn this model with lower sample complexity by leveraging the factorization which has been demonstrated in the online setting (Kearns & Koller, 1999). We remark a factored MDP is an example where model-based approaches are necessary as neither the optimal policy nor the Q functions of the factored MDP are factored (Koller & Parr, 2000).

We will slightly modify Algorithm 1 to take the factorization into consideration. First, we perform MLE for model learning: each factor $P_i^\star$ is independently learned via MLE:

$$\forall i \in [d], \widehat{P}_{\mathrm{MLE},i} = \arg\max_P \mathbb{E}_{\mathcal{D}}[\ln P(s'[i]|s[\mathrm{pa}_i], a)], \quad \widehat{P} = \textstyle\prod_i \widehat{P}_{\mathrm{MLE},i}.$$

Next, the constrained policy optimization procedure is defined as

$$\hat{\pi} = \arg\max_{\pi} \min_{P := \prod_i P_i} V_P^\pi, \text{ s.t.}, \mathbb{E}_{\mathcal{D}}[\mathrm{TV}(P_i(\cdot \mid s,a), \widehat{P}_{\mathrm{MLE},i}(\cdot \mid s,a))^2] \leq \xi_i \, (\forall i \in [1, \cdots, d]).$$

Note that in the above objective, there is no restriction on the policy, i.e., the $\arg\max$ operator searches over all possible Markovian policies.

To analyze the performance of the above modified CPPO, we introduce the specialized concentration coefficient for factored MDPs that utilizes the factored structure. We focus on density ratio based concentrability coefficients since in a factored MDP with the function class $\mathcal{M} := \{P = \prod_i P_i : P_i \in \mathcal{S}_i \to \Delta(\mathcal{O})\}$, the concentrability coefficient associated with $\mathcal{M}$ in Definition 1 will be reduced to the density ratio. For any π^*, we define two types of concentrability coefficients for the factored MDP as follows:

$$\ddot{C}_{\pi^*,\infty} := \max_{j \in [1,\cdots,d]} \max_{s_j \in \mathcal{S}_j, a \in \mathcal{A}} \frac{d_{P^\star}^{\pi^*}(s_j,a)}{\rho(s_j,a)}, \quad \ddot{C}_{\pi^*,2} := \max_{j \in [1,\cdots,d]} \mathbb{E}_{(s_j,a)\sim\rho}\left[\left(\frac{d_{P^\star}^{\pi^*}(s_j,a)}{\rho(s_j,a)}\right)^2\right],$$

where for $s_j \in \mathcal{S}_j$, we denote $\nu(s_j,a) := \sum_{s \in \mathcal{S}: s[\mathrm{pa}_j] = s_j} \nu(s,a)$ for any distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$. While we can show that $\ddot{C}_{\pi^*,\infty} \leq \ddot{C}_{\pi^*,2}$, we include $\ddot{C}_{\pi^*,\infty}$ here since this quantity is more interpretable. Comparing to the classic density ratio based concentrability coefficient defined on the original state space $\mathcal{S}$, here $\ddot{C}_{\pi^*,\infty}$ and $\ddot{C}_{\pi^*,2}$ are defined over each state space $\mathcal{S}_j$ associated with each factor j. Note that when $|\mathrm{pa}_j| = \Theta(1)$, $|\mathcal{S}_j|$ is exponentially smaller than $|\mathcal{S}|$. One can verify that $\ddot{C}_{\pi^*,\infty} \leq C_{\pi^*,\infty}, \ddot{C}_{\pi^*,2} \leq C_{\pi^*,2}$ (see Appendix F.8), where $C_{\pi^*,\infty}$ and $C_{\pi^*,2}$ both ignore the factored structure and treat $\mathcal{S}$ as a whole single space. This formally demonstrates the benefit of the factored structure in terms of the coverage condition in offline RL.

With the new definition of the concentrability coefficients, now we are ready to state the PAC bound of CPPO for factored MDPs. Recall $L := \sum_{i=1}^d L_i, L_i = |\mathcal{A}||\mathcal{O}|^{1+|\mathrm{pa}_i|}$.

Theorem 3 (PAC bound for factored MDP)**.** *We set $\xi_i = c_1 \frac{L_i \ln(L_i c_2 d/\delta)}{n}$. Then with probability $1-\delta$, CPPO finds a policy $\hat{\pi}$ such that for all comparator policy $\pi^* \in \Pi$ (Π can be unrestricted),*

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq c_3 (1-\gamma)^{-2} \sqrt{\frac{d \ddot{C}_{\pi^*,2} L \cdot \ln(n L c_4 d/\delta)}{n}}.$$

Comparing with Corollary 1 for tabular MDPs, we see that our sub-optimality gap scales polynomially with respect to L, i.e., the complexity of the factored MDP, rather than $|\mathcal{S}|$ which can be $\Omega(\exp(d))$.

6 Conclusion

We study model-based offline RL with function approximation under partial coverage. We show that for the model-based setting, realizability in function class and partial coverage together are enough to learn a policy that is comparable to *any* policies (including history-dependent policies) covered by the offline distribution. Our result demonstrates a sharp contrast to model-free offline RL approaches which often require additional structural conditions in the function class (e.g., Bellman completion) and have restrictions on the pool of candidate policies that they can compete against.

REFERENCES

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13399–13412, 2020a.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20095–20107, 2020b.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 64–66, 2020c.

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*, 71:89–129, 2008.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of machine learning research*, 18(21):1–38, 2017.

Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Jonathan D Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. Mitigating covariate shift in imitation learning via offline data without great coverage. 2021.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1042–1051, 2019.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on web search and data mining*, WSDM '19, pp. 456–464, 2019.

Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pp. 465–472. Citeseer, 2011.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2701–2709, 2020.

Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. *arXiv preprint arXiv:2103.13883*, 2021.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

Rasool Fakoor, Jonas Mueller, Pratik Chaudhari, and Alexander J Smola. Continuous doubly constrained batch reinforcement learning. *arXiv preprint arXiv:2102.09225*, 2021.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 486–489, 2020.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, and Shixiang Shane Gu. Emaq: Expected-max q-learning operator for simple yet effective offline and online rl. In *International Conference on Machine Learning*, pp. 3682–3691. PMLR, 2021.

Botao Hao, Yaqi Duan, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning*, pp. 4063–4073. PMLR, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143, 2020a.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020b.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15312–15325, 2020.

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

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823. Curran Associates, Inc., 2020.

Daphne Koller and Ronald Parr. Policy iteration for factored mdps. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pp. 326–334, 2000.

Michael R. Kosorok and Eric B. Laber. Precision medicine. 6:263–286, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

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.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1264–1274, 2020.

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pp. 3242–3245. PMLR, 2021.

Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Gu. Deployment-efficient reinforcement learning via model-based offline optimization. *ICLR*, 2020.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020. PMLR, 2020.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv: 2102.07035*, 2021.

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

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *arXiv preprint arXiv:2007.01891*, 2020.

Chengzhuo Ni, Anru Zhang, Yaqi Duan, and Mengdi Wang. Learning good state and action representations via tensor decomposition. *arXiv preprint arXiv:2105.01136*, 2021.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. *arXiv preprint arXiv:2104.03781*, 2021.

Nived Rajaraman, Lin F Yang, Jiantao Jiao, and Kannan Ramachandran. Toward the fundamental limits of imitation learning. *arXiv preprint arXiv:2009.05990*, 2020.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. 2005.

Shideh Rezaeifar, Robert Dadashi, Nino Vieillard, Léonard Hussenot, Olivier Bachem, Olivier Pietquin, and Matthieu Geist. Offline reinforcement learning as anti-exploration. *arXiv preprint arXiv:2106.06431*, 2021.

Stéphane Ross and J Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1905–1912, 2012.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 1015–1022, 2010.

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 *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2898–2933, 2019.

Ahmed Touati, Amy Zhang, Joelle Pineau, and Pascal Vincent. Stable policy optimization via off-policy divergence regularization. *arXiv preprint arXiv:2003.04108*, 2020.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9659–9668, 2020.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pp. 654–663, Arlington, Virginia, USA, 2013. AUAI Press.

S van de Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

Martin J Wainwright. *High-Dimensional Statistics : A Non-Asymptotic Viewpoint*. Cambridge University Press, New York, 2019.

Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline rl with linear function approximation?. *arXiv preprint arXiv:2010.11895*, 2020.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10746–10756, 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*, 2021.

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

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*, 2021.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural computation*, 17(9):2077–2098, 2005.

Weitong Zhang, Jiafan He, Dongruo Zhou, Amy Zhang, and Quanquan Gu. Provably efficient representation learning in low-rank markov decision processes. *arXiv preprint arXiv:2106.11935*, 2021a.

Xuezhou Zhang, Yiding Chen, Jerry Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. *arXiv preprint arXiv:2106.06630*, 2021b.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021.

A GENERALIZATION OF THEOREM 1

We present the generalized version of Theorem 1 when the hypothesis class is infinite. We define the modified function class of $\mathcal{M}$:

$$\mathcal{H} = \left\{ \sqrt{\frac{P + P^*}{2}} \mid P \in \mathcal{M} \right\}.$$

Given a function class $\mathcal{F}$, let $\mathcal{N}_{[]}(\delta, \mathcal{F}, d)$ be the bracketing number of $\mathcal{F}$ w.r.t the metric $d(a, b)$ given by

$$d(a, b) = \mathbb{E}_{(s,a)\sim\rho} \left[\int (a(s' \mid s, a) - b(s' \mid s, a))^2 \mathrm{d}(s') \right]^{1/2}.$$

Then, the entropy integral of $\mathcal{F}$ is given by

$$J_B(\delta, \mathcal{F}, d) = \max \left(\int_{\delta^2/2}^{\delta} (\log \mathcal{N}_{[]}(u, \mathcal{F}, d))^{1/2} \mathrm{d}u, \delta \right). \tag{4}$$

We also define the localized class of $\mathcal{H}$:

$$\mathcal{H}(\delta) = \{h \in \mathcal{H} : \mathbb{E}_{(s,a)\sim\rho}[h^2(P(\cdot \mid s, a) \| P^\star(\cdot \mid s, a))] \leq \delta^2\},$$

where $h(P(\cdot \mid s, a) \| P^\star(\cdot \mid s, a))$ denotes Hellinger distance defined by

$$\left(0.5 \int \{\sqrt{P(s' \mid s, a)} - \sqrt{P^\star(s' \mid s, a)}\}^2 \mathrm{d}(s') \right)^{1/2}.$$

Based on Theorem 7.4 (van de Geer, 2000), the MLE has the following guarantee.

Theorem 4 (MLE guarantee with general function approximation)**.** *We take a function $G(\epsilon) : [0, 1] \to \mathbb{R}$ s.t. $G(\epsilon) \geq J_B[\epsilon, \mathcal{H}(\epsilon), d]$ and $G(\epsilon)/\epsilon^2$ is a non-increasing function w.r.t ϵ. Then, letting ξ_n be a solution to $\sqrt{n}\epsilon^2 \geq cG(\epsilon)$ w.r.t ϵ. With probability $1 - \delta$, we have*

$$\mathbb{E}_{(s,a)\sim\rho}[\|\hat{P}_{\mathrm{MLE}}(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1^2] \leq c_1 \left\{ \xi_n + \sqrt{\log(c_2/\delta)/n} \right\}^2.$$

We remark that the original guarantee in (van de Geer, 2000) is given for the estimation of unconditional distributions. The adaption to the conditional case is straightforward. For more details, refer to Section A.1. Besides, when we assume the convexity of the function class, the entropy integral with bracketing (4) can be replaced with the entropy integral with covering number (Wainwright, 2019, Chapter 14).

By using the above notation and the MLE guarantee, we can generalize Theorem 1.

Theorem 5 (Finite sample error bound of CPPO with an infinite hypothesis class)**.** *Assume $P^\star \in \mathcal{M}$. Let $f(P)(s, a) = \mathrm{TV}(P(\cdot \mid s, a), P^\star(\cdot \mid s, a))^2$. Define*

$$\mathcal{M}_1 = \left\{ P : \mathbb{E}_{(s,a)\sim\rho} [f(P)(s, a)] \leq c \left(\xi_n^2 + \frac{\ln(c/\delta)}{n} \right) \right\},$$
$$\mathcal{M}_2 = \left\{ P : \mathbb{E}_{(s,a)\sim\mathcal{D}}[f(P)(s, a)] \leq c \left(G(\mathcal{M}_1) + \xi_n^2 + \frac{\ln(c/\delta)}{n} \right) \right\},$$
$$G(\mathcal{M}_1) = \mathbb{E}[\sup_{P\in\mathcal{M}_1} |(\mathbb{E}_\mathcal{D} - \mathbb{E}_\rho)[f(P)]|], \; G(\mathcal{M}_2) = \mathbb{E}[\sup_{P\in\mathcal{M}_\mathcal{D}} |(\mathbb{E}_\mathcal{D} - \mathbb{E}_\rho)[f(P)]|].$$

Here, in $G(\mathcal{M}_1)$ and $G(\mathcal{M}_2)$, the expectation is taken over the data. We set $\xi = cG(\mathcal{M}_1) + c\xi_n^2 + c\left(\frac{\ln(c/\delta)}{n}\right)$. Then, for all $\pi^* \in \Pi$, we have

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq (1 - \gamma)^{-2} c_1 \sqrt{C_{\pi^*}^\dagger} \sqrt{G(\mathcal{M}_2) + G(\mathcal{M}_1) + \xi_n^2 + \frac{\ln(c/\delta)}{n}}.$$

This theorem shows once we can calculate $\mathcal{G}_{\mathcal{M}_1}, \mathcal{G}_{\mathcal{M}_2}$ and ξ_n, we can obtain the tight rate. Importantly, $\mathcal{G}_{\mathcal{M}_1}$ and $\mathcal{G}_{\mathcal{M}_2}$ are upper-bounded by the localized versions of Rademacher complexities based on symmetrization argument. Hence, their rates are faster than the ones of the nonlocalized versions.

For example, when $|\mathcal{M}|$ is finite, we first have $\xi_n = \sqrt{\ln(|\mathcal{M}|c/\delta)/n}$. Then, from Bernstein's inequality (Wainwright, 2019, Exercise 2.8) and union bound, $G(\mathcal{M}_1)$ is upper-bounded by

$$G(\mathcal{M}_1) \lesssim \underbrace{\xi_n}_{\text{Variance term}} \times \underbrace{\sqrt{\ln(|\mathcal{M}|c/\delta)/n}}_{\text{Union bound}} \lesssim \xi_n^2.$$

Similarly, from empirical Bernstein's inequality,

$$G(\mathcal{M}_2) \lesssim \underbrace{\xi_n}_{\text{Variance term}} \times \underbrace{\sqrt{\ln(|\mathcal{M}|c/\delta)/n}}_{\text{Union bound}} \lesssim \xi_n^2.$$

Then, we can obtain the result of Theorem 1:

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq (1-\gamma)^{-2} c_1 \sqrt{C_{\pi^*}^{\dagger}} \sqrt{\frac{\ln(|\mathcal{M}|c/\delta)}{n}}, \quad \forall \pi^* \in \Pi.$$

We stress if we use Hoeffeding's inequality above, we would immediately get the slower rate $O(n^{-1/4})$. To calculate $G(\mathcal{M}_1)$ and $G(\mathcal{M}_2)$ in a tight manner, we need to leverage the knowledge that the variance of each element in $\mathcal{M}_1$ and $\mathcal{M}_2$ is controlled from the restriction $P \in \mathcal{M}_1$ or $P \in \mathcal{M}_2$.

A.1 Rate of Convergence of Maximum Likelihood Estimation with Infinite Hypothesis Class

We aim for obtaining a PAC guarantee of MLE following (van de Geer, 2000). We explain how we should modify the proof of van de Geer (2000) for unconditional density estimation to conditional density estimation. For simplicity, we assume $P^\star > 0$.

We first introduce the notation:

$$\bar{P} = (P + P^\star)/2, g_P = 0.5 \log \frac{\bar{P}}{P^\star}, (\bar{\mathcal{M}})^{1/2} = \left\{ \sqrt{\bar{P}} \mid P \in \mathcal{M} \right\}.$$

Recall

$$h^2(P_1(\cdot \mid s,a), P_2(\cdot \mid s,a)) = \left(0.5 \int P_1^{1/2}(s' \mid s,a) - P^{1/2}(s' \mid s,a) \mathrm{d}(s') \right)^{0.5}.$$

Here, from Lemma 4.2 (van de Geer, 2000), the following holds:

Lemma 1 (Some property of Hellinger distance)**.**

$$\mathbb{E}_{(s,a)\sim\rho}[h^2(\bar{P}_1(\cdot \mid s,a), \bar{P}_2(\cdot \mid s,a))] \leq 0.5 \mathbb{E}_{(s,a)\sim\rho}[h^2(P_1(\cdot \mid s,a), P_2(\cdot \mid s,a))],$$
$$\mathbb{E}_{(s,a)\sim\rho}[h^2(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))] \leq \mathbb{E}_{(s,a)\sim\rho}[16 h^2(\bar{P}(\cdot \mid s,a), P^\star(\cdot \mid s,a))].$$

We also recall Hellinger distance is stronger than TV distance:

Lemma 2 (Relation of Hellinger distance and TV distance)**.**

$$\mathrm{TV}(P_1(\cdot \mid s,a), P_2(\cdot \mid s,a)) \leq \sqrt{2} h(P_1(\cdot \mid s,a), P_2(\cdot \mid s,a)).$$

The following lemma is useful to connect the log-loss and the Hellinger distance.

Lemma 3 (Basic Inequality for MLE)**.**

$$\mathbb{E}_{(s,a)\sim\rho}[h^2(\hat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))] \leq (\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho, s'\sim P^\star(\cdot|s,a)})[g_P(s,a,s')].$$

This is proved by Lemma 4.1 (van de Geer, 2000). To simplify the notation, we define

$$H^2(P_1, P_2) = \mathbb{E}_{(s,a)\sim\rho}[h^2(P_1(\cdot \mid s,a), P_2(\cdot \mid s,a))].$$

Here, our goal is showing with probability $1-\delta$,

$$H^2(\hat{P}_{\mathrm{MLE}}, P^\star) \leq \{\xi_n + \sqrt{\log(c_2/\delta)/n}\}^2.$$

This is proved by showing for $x \geq \xi_n$,

$$\mathrm{P}(H^2(\hat{P}_{\mathrm{MLE}}, P^\star) \geq x^2) \leq c\exp(-nx^2/c^2).$$

This corresponds to the statement in Theorem 7.4 (van de Geer, 2000). To prove the above, we first use

$$\mathrm{P}(H^2(\hat{P}_{\mathrm{MLE}}, P^\star) \geq x^2) \leq \mathrm{P}(16H^2(\bar{\hat{P}}_{\mathrm{MLE}}, P^\star) \geq x^2),$$

from Lemma 1. Then, from Lemma 3, this is upper-bounded by

$$\mathrm{P}(\sup_{P\in\mathcal{M}, H^2(\bar{P},P^\star)\geq x^2/16} \nu_n(g_P) - \sqrt{n}H^2(P,P^\star) \geq 0) \tag{5}$$

where $\nu_n = \sqrt{n}(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho, s'\sim P^\star(\cdot|s,a)})$. To prove the term 5 is less than $c\exp(-n\delta^2/c^2)$, we use Theorem 5.11 (van de Geer, 2000), that is, some uniform inequality based on entropy with bracketing. The rest of the proof is the same as Theorem 7.4 (van de Geer, 2000). In summary, the only difference is we use the distance $H(a,b)$ tailored to the conditional density estimation instead of unconditional density estimation.

B More details for KNRs

We explain the algorithm and present the PAC guaranteed for KNRs. Here, we denote the dimension of $\mathcal{S}$ by $d_{\mathcal{S}}$.

We tailor Algorithm 1 to KNRs as follows to obtain a tighter guarantee. First, MLE procedure is replaced with $\hat{W}_{\mathrm{MLE}}$ by regularized MLE:

$$\hat{W}_{\mathrm{MLE}} = \operatorname*{arg\,min}_{W\in\mathbb{R}^{d_{\mathcal{S}}\times d}} \mathbb{E}_{\mathcal{D}}[\|W\phi(s,a) - s'\|_2^2] + \lambda\|W\|_F^2,$$

where $\|\cdot\|_F$ is a Frobenius norm. Then, the final policy optimization procedure is

$$\hat{\pi} = \arg\max_{\pi\in\Pi} \min_{W\in\mathcal{W}_{\mathcal{D}}} V^{\pi}_{P(W)}, \text{ s.t.}, \mathcal{W}_{\mathcal{D}} = \{W \in \mathbb{R}^{d_{\mathcal{S}}\times d} : \|(\hat{W}_{\mathrm{MLE}} - W)(\Sigma_n)^{1/2}\|_2 \leq \xi\}$$

where $\Sigma_n = \sum_{i=1}^n \phi(s_i,a_i)\phi^\top(s_i,a_i)$. We state the theoretical guarantee for KNRs below.

Corollary 3 (PAC bound for KNRs)**.** *Assume $\|\phi(s,a)\|_2 \leq 1, \forall(s,a) \in \mathcal{S}\times\mathcal{A}$. We set*

$$\xi = \sqrt{2\lambda\|W^\star\|_2^2 + 8\zeta^2\left(d_{\mathcal{S}}\ln(5) + \ln(1/\delta) + \bar{\mathcal{I}}_n\right)}, \quad \bar{\mathcal{I}}_n = \ln\left(\det(\Sigma_n)/\det(\lambda\mathbf{I})\right).$$

Suppose the KNR model is well-specified. By letting $\|W^\star\|_2^2 = O(1), \zeta^2 = O(1), \lambda = O(1)$, with probability $1-\delta$, for all π^*, we have

$$V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} \leq c_1 H^2 \min(d^{1/2}, \bar{R})\sqrt{\bar{R}}\sqrt{\frac{d_{\mathcal{S}}\bar{C}_{\pi^*,P^\star}\ln(1+n)}{n}}, \quad \textit{where } \bar{R} := \mathrm{rank}[\Sigma_\rho]\{\mathrm{rank}[\Sigma_\rho] + \ln(c_2/\delta)\}.$$

The proof is deferred to Section F.6.

B.1 KNRs in RKHS

We consider the ground truth model $P^\star(s'|s,a)$ defined as $s' = g^\star(s,a) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \zeta^2\mathbf{I})$ where each component in $g^\star$ belongs to an RKHS $\mathcal{H}_k$ with a kernel $k(\cdot,\cdot)$. We assume $s \in \mathbb{R}^{d_{\mathcal{S}}}$.

We denote a $d_{\mathcal{S}}$ dimensional RKHS as $\bigoplus \mathcal{H}_k$. We also denote the corresponding model to g by $P(g)$. We can learn $g^\star$ by regularized MLE (kernel ridge regression) using offline dataset $\mathcal{D}$. We tailor Algorithm 1 as follows.

First, by letting $x := (s,a), x_i = (s_i,a_i)$, MLE procedure is replaced with $\hat{g}_{\mathrm{MLE}}$ by regularized MLE:

$$\hat{g}_{\mathrm{MLE}}(\cdot) = S(\mathbf{K}_n + \zeta^2\mathbf{I})^{-1}\bar{k}_n(\cdot), \quad S = [s'_1, \cdots, s'_n] \in \mathbb{R}^{d_{\mathcal{S}}\times n}, \quad \bar{k}_n(x) = [k(x_1,x), \cdots, k(x_n,x)]^\top,$$

$$\{\mathbf{K}_n\}_{i,j} = k(x_i,x_j)\,(1\leq i\leq n, 1\leq j\leq n),\ k_n(x,x') = k(x,x') - \bar{k}_n(x)^\top(\mathbf{K}_n + \zeta^2\mathbf{I})^{-1}\bar{k}_n(x'),$$

where the notation $\|\cdot\|_{k_n}$ is a norm associated with an RKHS with a kernel $k_n(\cdot,\cdot)$. The final optimization procedure is replaced with

$$\hat{\pi} = \arg\max_{\pi\in\Pi} \min_{g\in\mathcal{G}_{\mathcal{D}}} V^{\pi}_{P(g)}, \text{s.t.}, \mathcal{G}_{\mathcal{D}} = \{g \in \bigoplus \mathcal{H}_k : \sum_{i=1}^{d_{\mathcal{S}}} \|\hat{g}_{i,\text{MLE}} - g_i\|^2_{k_n} \leq \xi^2\}.$$

We state the theoretical guarantee for KNRs with RKHS below. Before proceeding to the result, we prepare several notations and definitions. For simplicity, following (Srinivas et al., 2010), we suppose as follows:

Assumption 1. *$k(x,x) \leq 1, \forall x \in \mathcal{S}\times\mathcal{A}$ and there exists a set of pairs of eigenvalues and eigenfunctions $\{\mu_i, \psi_i\}_{i=1}^{\infty}$, where $\int \rho(x)\psi_i(x)\psi_i(x)dx = 1$ for all i and $\int \rho(x)\psi_i(x)\psi_j(x)dx = 0$ for $i \neq j$.*

The above is ensured by Mercer's theorem (Rasmussen & Williams, 2005). Eigenfunctions and eigenvalues essentially defines an infinite-dimensional feature mapping $\phi(x) := [\sqrt{\mu_1}\psi_1(x), \ldots, \sqrt{\mu_\infty}\psi_\infty(x)]^\top$. By setting eigenvalues $\{\mu_1, \ldots, \mu_\infty\}$ is in non-increasing order, we define the effective dimension below:

Definition 7 (Effective dimension)**.** $d^* = \min\{j \in \mathbb{N} : j \geq B(j+1)n/\zeta^2\}$, $B(j) = \sum_{k=j}^{\infty} \mu_k$.

The effective dimension d^* is commonly used and calculated for many kernels (Zhang, 2005; Bach, 2017; Valko et al., 2013). In finite-dimensional linear kernels $\{x \mapsto a^\top\phi(x); a \in \mathbb{R}^d\}$ ($k(x,x) = \phi^\top(x)\phi(x)$), we have $d^* \leq \text{rank}[\Sigma_\rho]$. Thus, d^* is regarded a natural extension of $\text{rank}[\Sigma_\rho]$ to infinite-dimensional models. Note that d^* itself is offline distribution dependent, i.e., the eigenvalues and eigenfunctions are defined using the offline distribution ρ.

With the above preparations in mind, we present the theoretical result below.

Corollary 4 (PAC bound for RKHS models)**.** *Let $\Sigma_{\pi^*} = \mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[\phi(s,a)\phi(s,a)^\top], \Sigma_\rho = \mathbb{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top]$. We set ξ:*

$$\xi = \sqrt{d_{\mathcal{S}}\{2 + 150\ln^3(d_{\mathcal{S}}n/\delta)\mathcal{I}_n\}}, \quad \mathcal{I}_n = \ln(\det(\mathbf{I} + \zeta^{-2}\mathbf{K}_n))$$

and $\zeta^2 = O(1)$. With probability at least $1-\delta$, for all comparator policy π^* with $\bar{C}_{\pi^*,P^\star} = \sup_{\|x\|_2\leq 1}(x^\top\Sigma_{\pi^*}x/x^\top\Sigma_\rho x) < \infty$, we have:

$$V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} \leq c_1(1-\gamma)^{-2}\{d^* + \ln(c_2/\delta)\}d^*\sqrt{\frac{d_{\mathcal{S}}\bar{C}_{\pi^*,P^\star}\ln^3(c_3 d_{\mathcal{S}}n/\delta)\ln(1+n)}{n}},$$

This implies the algorithm has a valid PAC guarantee under the partial coverage of KNRs with RKHS. The proof is deferred in Section F.6.

C More Related Works

Ni et al. (2021) assumes the transition operator can be properly embedded into predefined Reproducing Kernel Hilbert Spaces and learns low-dimensional state-action representations via kernelized embedding and low-rank tensor decomposition. However, they did not study the errors for policy optimization after using these learned features. Regarding offline distribution coverage, Ni et al. (2021) assumes that the feature covariance matrix (feature associated with the pre-defined kernel) of the offline distribution is full rank.

Hao et al. (2021) studies an OPE problem on sparse linear Bellman complete MDPs in the offline learning setting where they assume all covariance matrices (covariance matrices that correspond to all possible subsets of features) under the offline distribution are full rank as well. We study policy optimization in low-rank MDPs (with unknown feature representation), and we do not assume full coverage, i.e., we do not assume the feature covariance matrix is full rank, and indeed our result is distribution-dependent since it scales with respect to the rank of the covariance matrix that is defined using the ground truth feature representation.

D RESULTS WITH DIFFERENT CONCENTRABILITY COEFFICIENTS

In this section, we show we can obtain the different versions of Theorem 1 using different concentrability coefficients. We introduce the following concentrability coefficients:

$$\acute{C}_{\pi^*} = \left(\sup_{P \in \mathcal{M}} \frac{\mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))]}{\mathbb{E}_{(s,a)\sim \rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))]} \right)^2, \; C_{\pi^*,2} = \mathbb{E}_{(s,a)\sim\rho}\left[\left(\frac{d^{\pi^*}_{P^\star}(s,a)}{\rho(s,a)} \right)^2 \right].$$

The quantity $\acute{C}_{\pi^*}$ is similar to $C^\dagger_{\pi^*}$.

By slightly modifying the proof of Theorem 1, the following theorem is yielded.

Theorem 6 (PAC bound for CPPO with different concentrability coefficients). *Assume $P^\star \in \mathcal{M}$. We set $\xi = c_1 \frac{\ln(c_2|\mathcal{M}|/\delta)}{n}$. Then, with probability $1-\delta$, for any comparator policy π^*,*

$$V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} \le c_3 (1-\gamma)^{-2} \sqrt{\acute{C}_{\pi^*} \frac{\ln(c_4|\mathcal{M}|/\delta)}{n}},$$
$$V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} \le c_3 (1-\gamma)^{-2} \sqrt{C_{\pi^*,2} \frac{\ln(c_4|\mathcal{M}|/\delta)}{n}}.$$

The proof is deferred to Section F.1.

Though $\acute{C}_{\pi^*}$ is not generally comparable to $C^\dagger_{\pi^*}$, we still believe $C^\dagger_{\pi^*}$ is more preferable quantity. This is because this $\acute{C}_{\pi^*}$ does not result in the relative condition number in KNRs.

The density ratio based concentrability coefficient $C_{\pi^*,2}$ is always smaller than or equal to $C_{\pi^*,\infty}$. This is seen as follows:

$$C_{\pi^*,2} = \mathbb{E}_{(s,a)\sim\rho}\left[\left(\frac{d^{\pi^*}_{P^\star}(s,a)}{\rho(s,a)} \right)^2 \right] = \mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}\left[\left(\frac{d^{\pi^*}_{P^\star}(s,a)}{\rho(s,a)} \right) \right] \le C_{\pi^*,\infty}.$$

We remark $C_{\pi^*,2}$ is generally not comparable to $C^\dagger_{\pi^*}$.

E COMPARISON TO XIE ET AL. (2021)

We compare a result in (Xie et al., 2021) to our result in detail. Let $\mathcal{F}$ be a function class for Q-functions. Here, we consider a more general version of their algorithm by replacing the original $\mathcal{E}(f,\pi;\mathcal{D})$ in their algorithm with

$$\mathcal{E}(f,\pi;\mathcal{D}) := \mathcal{L}(f,f;\pi,\mathcal{D}) - \min_{g\in\mathcal{G}} \mathcal{L}(g,f;\pi,\mathcal{D}).$$

In their original algorithm, they set $\mathcal{G} = \mathcal{F}$. Here, we consider the version such that a discriminator class $\mathcal{G}$ can be different from $\mathcal{F}$.

They show the PAC result under partial coverage as follows. Here, $\mathcal{T}^\pi_{P^\star}$ is a Bellman operator under π and $P^\star$:

$$\mathcal{T}^\pi_{P^\star} : \{\mathcal{S}\times\mathcal{A}\to\mathbb{R}\} \ni f \mapsto r(s,a) + \mathbb{E}_{P^*(s'|s,a)}[f(s',\pi)] \in \{\mathcal{S}\times\mathcal{A}\to\mathbb{R}\}.$$

Theorem 7 (Extension of Result in (Xie et al., 2021)). *Suppose realizaibility $Q^\pi_{P^\star} \in \mathcal{F}, \forall \pi \in \Pi$ and closeness $\max_{f\in\mathcal{F}} \min_{g\in\mathcal{G}} \mathbb{E}_{s,a\sim\rho}[(g - \mathcal{T}^\pi_{P^\star} f)^2(s,a)] = 0, \forall \pi \in \Pi$. Then, with $1-\delta$, for any $\pi^* \in \Pi$, we have*

$$V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} = O(\sqrt{C^\diamond \ln(|\Pi||\mathcal{F}||\mathcal{G}|/\delta)/n}), \quad C^\diamond = \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[(f-\mathcal{T}f)^2(s,a)]}{\mathbb{E}_{(s,a)\sim\rho}[(f-\mathcal{T}f)^2(s,a)]}.$$

By combining this result with the conversion from model-free results to model-based results in (Chen & Jiang, 2019, Corollary 6), we can obtain the following result under partial coverage.

Theorem 8 (PAC guarantee from the direct application of (Xie et al., 2021) to mode-based RL). *Assume $P^\star \in \mathcal{M}$. Then, there exists an algorithm s.t. with $1-\delta$, for any policy $\pi^\star \in \Pi$,*

$$V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} = O(\sqrt{C^\diamond \ln(|\Pi||\mathcal{M}|/\delta)/n}).$$

As we mentioned, this is worse than our result since it includes $|\Pi|$. Besides, the algorithm can only compete against policies restricted in Π, while our algorithm works for the unrestricted policy class Π which could even include history dependent policies. For completeness, we give the proof as follows.

We remark that their results (Theorem 4.1) with NPG that can possibly compete with any stochastic policies, are not applicable here. This is because they need an assumption that the comparator policy π^* needs to satisfy $Q^{\pi^*}_{P^\star} \in \mathcal{F}$ and $\max_{f\in\mathcal{F}} \min_{g\in\mathcal{G}} \mathbb{E}_{s,a\sim\rho}[(g - \mathcal{T}^{\pi^*}_{P^\star} f)^2(s,a)] = 0$, which does not hold for the corresponding Q-function class $\mathcal{F}$ after the conversion. As a notable exception, when the model is a linear Bellman-complete MDP (Zanette et al., 2021), any stochastic policies satisfy the Bellman completeness for the linear Q-function class; then, their algorithms can learn policies that can compete with any stochastic policies satisfying partial coverage.

Proof of Theorem 8. Given a model class $\mathcal{M}$, consider the following reduction. We define a Q-function class:

$$\mathcal{F} = \{q^\pi_P \mid \pi \in \Pi, P \in \mathcal{M}\}.$$

Then, we define a discriminator class $\mathcal{G}$:

$$\mathcal{G} = \{\mathcal{T}^{\pi'}_{P'} q^\pi_P \mid \pi \in \Pi, \pi' \in \Pi, P \in \mathcal{M}, P' \in \mathcal{M}\}.$$

The above satisfies the realizability $Q^\pi_{P^\star} \in \mathcal{F}, \forall \pi \in \Pi$ and the closedness $\mathcal{T}^\pi_{P^\star}\mathcal{F} \subset \mathcal{G}, \forall \pi \in \Pi$. Thus, the assumptions in Theorem 7 are satisfied. Then, we have

$$\begin{aligned} V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} &= O(\sqrt{C^\diamond \ln(|\Pi||\mathcal{F}||\mathcal{G}|/\delta)/n}) \\ &= O(\sqrt{C^\diamond \ln(|\Pi||\mathcal{M}|/\delta)/n}), \end{aligned}$$

noting $|\mathcal{F}| = |\Pi||\mathcal{M}|$ and $|\mathcal{G}| = |\Pi|^2|\mathcal{M}|^2$.

□

F Missing Proofs

Below we use $c, c_1, c_2, \cdots$ to denote universal constants. For a d-dimensional vector a and a matrix $A \in \mathbb{R}^{d\times d}$, we denote $\|a\|^2_A = a^\top A a$. Here, $a \lesssim B$ means $a \le cB$ for some universal constant. c

F.1 Proofs for General Function Approximation

Proof of Theorem 1. From Lemma 6, the MLE guarantee gives us the following generalization bound: with probability $1-\delta$,

$$\mathbb{E}_{s,a\sim\rho}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \lesssim \frac{\ln(|\mathcal{M}|/\delta)}{n}. \tag{6}$$

Letting

$$A(P) \coloneqq |\mathbb{E}_{s,a\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] - \mathbb{E}_{\mathcal{D}}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]|.$$

with probability $1-\delta$, from union bound and Bernstein's inequality, we also have

$$A(P) \le \sqrt{\frac{c_1 \mathrm{var}_{(s,a)\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]\ln(|\mathcal{M}|/\delta)}{n}} + \frac{c_2 \ln(|\mathcal{M}|/\delta)}{n}, \forall P \in \mathcal{M}. \tag{7}$$

Hereafter, we condition on the above two events. Recall that we construct the version space using $\mathcal{D}$ and $\widehat{P}_{\mathrm{MLE}}$ as follows:

$$\mathcal{M}_{\mathcal{D}} \coloneqq \left\{P \in \mathcal{M} : \mathbb{E}_{\mathcal{D}}[\mathrm{TV}(P(\cdot \mid s,a), \hat{P}_{\mathrm{MLE}}(\cdot \mid s,a))^2] \le \xi\right\}.$$

First Step: Show $P^\star \in \mathcal{M}_\mathcal{D}$ in high-probability. We set $\xi = c\frac{\ln(|\mathcal{M}|/\delta)}{n}$. Conditioning on the above two events equations (6) and (7), we prove $P^\star \in \mathcal{M}_\mathcal{D}$. This is proved by

$$
\begin{aligned}
&\mathbb{E}_\mathcal{D}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \\
&= \mathbb{E}_\mathcal{D}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] - \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \\
&+ \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \\
&= \mathbb{E}_\mathcal{D}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] - \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] + \frac{c_1 \ln(|\mathcal{M}|/\delta)}{n} \\
&\lesssim \sqrt{\frac{\mathrm{var}_{(s,a)\sim\rho}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \ln(|\mathcal{M}|/\delta)}{n}} + \frac{\ln(|\mathcal{M}|/\delta)}{n} && \text{(From (7))} \\
&\lesssim \sqrt{\frac{\mathrm{E}_{(s,a)\sim\rho}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \ln(|\mathcal{M}|/\delta)}{n}} + \frac{\ln(|\mathcal{M}|/\delta)}{n} \\
& && (\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2 \le 4) \\
&\lesssim \frac{1}{n}\ln(|\mathcal{M}|/\delta). && \text{(Plug in MLE guarantee)}
\end{aligned}
$$

Second Step: Show $\mathbb{E}_{s,a\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \le c\xi, \quad \forall P \in \mathcal{M}_\mathcal{D}$ in high probability. We show for any $P \in \mathcal{M}_\mathcal{D}$, the distance between $P^\star$ is sufficiently controlled in terms of TV distance. More concretely (conditioning on the above two events (6) and (7)), we show

$$
\mathbb{E}_{s,a\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \lesssim \xi, \quad \forall P \in \mathcal{M}_\mathcal{D}.
$$

In order to observe this, for any $P \in \mathcal{M}_\mathcal{D}$, we have

$$
\begin{aligned}
&\mathbb{E}_\mathcal{D}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \\
&\le 2\mathbb{E}_\mathcal{D}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P(\cdot \mid s,a))^2] + 2\mathbb{E}_\mathcal{D}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \le 4\xi \\
& && \text{(From } (a+b)^2 \le 2a^2 + 2b^2.)
\end{aligned}
$$

Thus, we have:

$$
\begin{aligned}
&\mathbb{E}_{s,a\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \\
&= \mathbb{E}_{s,a\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] - \mathbb{E}_\mathcal{D}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] + \mathbb{E}_\mathcal{D}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \\
&\le A(P) + c\xi. && (8)
\end{aligned}
$$

Here, from (7), we have

$$
A(P) \le \sqrt{\frac{c_1 \mathrm{var}_{(s,a)\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]] \ln(|\mathcal{M}|/\delta)}{n}} + \frac{c_2 \ln(|\mathcal{M}|/\delta)}{n}, \forall P \in \mathcal{M}_\mathcal{D}.
$$

Then, for any $P \in \mathcal{M}_\mathcal{D}$, we have

$$
\begin{aligned}
A(P) &\le \sqrt{\frac{c_1 \mathrm{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^4] \ln(|\mathcal{M}|/\delta)}{n}} + \frac{c_2 \ln(|\mathcal{M}|/\delta)}{n} \\
&\le \sqrt{\frac{4c_1 \mathrm{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \ln(|\mathcal{M}|/\delta)}{n}} + \frac{c_2 \ln(|\mathcal{M}|/\delta)}{n} \\
& && ([\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \le 4.) \\
&\le \sqrt{\frac{4c_1 (A(P) + c\xi) \ln(|\mathcal{M}|/\delta)}{n}} + \frac{c_2 \ln(|\mathcal{M}|/\delta)}{n}.
\end{aligned}
$$

From $(a+b)^2 \le 2a^2 + 2b^2$,

$$
\begin{aligned}
A^2(P) &\lesssim \left(\sqrt{\frac{c(A(P)+\xi)\ln(|\mathcal{M}|/\delta)}{n}} + \frac{c\ln(|\mathcal{M}|/\delta)}{n}\right)^2 \lesssim \frac{(A(P)+\xi)\ln(|\mathcal{M}|/\delta)}{n} + \left\{\frac{c\ln(|\mathcal{M}|/\delta)}{n}\right\}^2 \\
&\lesssim \frac{(A(P)+\xi)\ln(|\mathcal{M}|/\delta)}{n} && (\xi \text{ includes } \ln(|\mathcal{M}|/\delta)) \\
&\lesssim \frac{(A(P) + 1/n\ln(|\mathcal{M}|/\delta))\ln(|\mathcal{M}|/\delta)}{n}.
\end{aligned}
$$

Then, we have

$$A^2(P) - B_1 A(P) - B_2 \leq 0, \quad B_1 = c\ln(|\mathcal{M}|/\delta)/n, \quad B_2 = c(1/n)^2 \ln(|\mathcal{M}|/\delta)^2.$$

This concludes

$$0 \leq A(P) \leq \frac{B_1 + \sqrt{B_1^2 + 4B_2}}{2} \leq c(B_1 + \sqrt{B_2}) \leq c\frac{\ln(|\mathcal{M}|/\delta)}{n} \lesssim \xi.$$

Thus, by using the above $A(P) \lesssim \xi(P \in \mathcal{M}_{\mathcal{D}})$ and (8), with probability $1-\delta$, we have:

$$\mathbb{E}_{s,a\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \leq A(P) + c\xi \lesssim \xi, \quad P \in \mathcal{M}_{\mathcal{D}}.$$

Third Step: Calculate the final error bound taking the distribution shift into account For any $P \in \mathcal{M}_{\mathcal{D}}$, we prove

$$V_{P^\star}^{\pi^*} - V_P^{\pi^*} \leq (1-\gamma)^{-2} c\sqrt{C_{\pi^*}^{\dagger}}\sqrt{\frac{\ln(|\mathcal{M}|/\delta)}{n}}. \tag{9}$$

For any $P \in \mathcal{M}_{\mathcal{D}}$, this is proved as follows:

$$\begin{aligned}
V_{P^\star}^{\pi^*} - V_P^{\pi^*} &\leq (1-\gamma)^{-2}\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))] \\
&\qquad\qquad \text{(Simulation lemma, Lemma 5)} \\
&\leq (1-\gamma)^{-2}\sqrt{\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]} \\
&\leq (1-\gamma)^{-2}\sqrt{C_{\pi^*}^{\dagger}\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]} \\
&\leq c(1-\gamma)^{-2}\sqrt{C_{\pi^*}^{\dagger}}\sqrt{\frac{\ln(|\mathcal{M}|/\delta)}{n}}. \quad \text{(Based on the consequence of the second step)}
\end{aligned}$$

Combining all things together, with probability $1-\delta$, for any $\pi^* \in \Pi$, we have

$$\begin{aligned}
V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} &\leq V_{P^\star}^{\pi^*} - \min_{P\in\mathcal{M}_{\mathcal{D}}} V_P^{\pi^*} + \min_{P\in\mathcal{M}_{\mathcal{D}}} V_P^{\pi^*} - V_{P^\star}^{\hat{\pi}} \\
&\leq V_{P^\star}^{\pi^*} - \min_{P\in\mathcal{M}_{\mathcal{D}}} V_P^{\pi^*} + \min_{P\in\mathcal{M}_{\mathcal{D}}} V_P^{\hat{\pi}} - V_{P^\star}^{\hat{\pi}} && \text{(definition of } \hat{\pi}\text{)} \\
&\leq V_{P^\star}^{\pi^*} - \min_{P\in\mathcal{M}_{\mathcal{D}}} V_P^{\pi^*} && \text{(Fist step, } P^\star \in \mathcal{M}_{\mathcal{D}}\text{)} \\
&\lesssim (1-\gamma)^{-2} c_1\sqrt{C_{\pi^*}^{\dagger}}\sqrt{\frac{\ln(|\mathcal{M}|c_2/\delta)}{n}}. && \text{(From (9))}
\end{aligned}$$

□

Remark 4 (To compete with all history-dependent polices)**.** *Consider the case where* Π *is all Markovian polices. We want to show we can compete with all history-dependent non-Markovian polices:*

$$\bar{\Pi} = \left\{\prod_{i=1}^{\infty}\pi_i \mid \pi_i \in \left[\left(\prod_{k=1}^{i-1}\mathcal{S}\times\mathcal{A}\right) \to \Delta(\mathcal{A})\right]\right\}.$$

We take an element π^* *from* $\bar{\Pi}$*. Then,* $V_{P^\star}^{\pi^*}$ *and* $d_{P^\star}^{\pi^*}$ *are still well-defined. Then, every step in the proof still holds. The only step we need to check carefully is this line:*

$$\begin{aligned}
V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} &\leq V_{P^\star}^{\pi^*} - \min_{P\in\mathcal{M}_{\mathcal{D}}} V_P^{\pi^*} + \min_{P\in\mathcal{M}_{\mathcal{D}}} V_P^{\pi^*} - V_{P^\star}^{\hat{\pi}} \\
&\leq V_{P^\star}^{\pi^*} - \min_{P\in\mathcal{M}_{\mathcal{D}}} V_P^{\pi^*} + \min_{P\in\mathcal{M}_{\mathcal{D}}} V_P^{\hat{\pi}} - V_{P^\star}^{\hat{\pi}}.
\end{aligned}$$

This is proved by $\max_{\pi\in\bar{\Pi}} V_P^{\pi} = \max_{\pi\in\Pi} V_P^{\pi}$ *for any* P*.*

F.2 Proofs for General function approximation with infinite hypothesis class

Proof of Theorem 5. From Theorem 4, the MLE guarantee gives us the following generalization bound: with probability $1-\delta$,

$$\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \lesssim \left(\xi_n^2 + \frac{\ln(c/\delta)}{n}\right). \tag{10}$$

We define

$$\mathcal{M}_1 = \left\{P \in \mathcal{M} : \mathbb{E}_{(s,a)\sim\rho}\left[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2\right] \le c\left(\xi_n^2 + \sqrt{\frac{\ln(c/\delta)}{n}}\right)\right\}.$$

for some large c. From a functional Bernstein's inequality (Lemma 12), by defining

$$f(P)(s,a) = \mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2,\ G(\mathcal{M}_1) = \mathbb{E}[\sup_{P\in\mathcal{M}_1} |(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_\rho)[f(P)]|].$$

with probability $1-\delta$, we have

$$\begin{aligned}
\sup_{P\in\mathcal{M}_1} |(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_\rho)[f(P)]| &\lesssim G(\mathcal{M}_1) + \{G(\mathcal{M}_1) + \sup_{P\in\mathcal{M}_1} \mathbb{E}_\rho[f(P)^2]\}^{1/2}\sqrt{\frac{\log(c_1/\delta)}{n}} + \frac{\log(c_1/\delta)}{n} \\
&\lesssim G(\mathcal{M}_1) + \{G(\mathcal{M}_1) + \sup_{P\in\mathcal{M}_1} \mathbb{E}_\rho[f(P)]\}^{1/2}\sqrt{\frac{\log(c_1/\delta)}{n}} + \frac{\log(c_1/\delta)}{n} \\
&\lesssim G(\mathcal{M}_1) + \xi_n^2 + \left(\frac{\ln(c/\delta)}{n}\right).
\end{aligned} \tag{11}$$

Similarly, by defining

$$\begin{aligned}
\mathcal{M}_2 &= \left\{P : \mathbb{E}_{(s,a)\sim\mathcal{D}}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \le cG(\mathcal{M}_1) + c\xi_n^2 + c\left(\frac{\ln(c/\delta)}{n}\right)\right\}, \\
G(\mathcal{M}_2) &= \mathbb{E}[\sup_{P\in\mathcal{M}_{\mathcal{D}}} |(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_\rho)[f(P)]|], \\
Z_n &= \sup_{P\in\mathcal{M}_2} \mathbb{E}_{(s,a)\sim\rho}[f(P)(s,a)].
\end{aligned}$$

from a functional Bernstein's inequality, with probability $1-\delta$, we have

$$\begin{aligned}
\sup_{P\in\mathcal{M}_2} |(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_\rho)[f(P)]| &\lesssim G(\mathcal{M}_2) + \{G(\mathcal{M}_2) + Z_n\}^{1/2}\sqrt{\frac{\log(c_1/\delta)}{n}} + \frac{\log(c_1/\delta)}{n} \\
&\lesssim G(\mathcal{M}_2) + \sqrt{Z_n\frac{\ln(c/\delta)}{n}} + \frac{\ln(c/\delta)}{n}
\end{aligned} \tag{12}$$

Hereafter, we condition on the above three events:(10), (11) and (12).

First step: Show $P^* \in \mathcal{M}_{\mathcal{D}}$ in high probability. From (10), we have

$$\mathbb{E}_{(s,a)\sim\rho}[f(\hat{P}_{\mathrm{MLE}})(s,a)] \le \xi_n^2 + \frac{\ln(c/\delta)}{n}.$$

Thus, $P^\star \in \mathcal{M}_1$. Then, from (11) and (10),

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\mathcal{D}}[f(\hat{P}_{\mathrm{MLE}})(s,a)] \\
&= |\mathbb{E}_{(s,a)\sim\mathcal{D}}[f(\hat{P}_{\mathrm{MLE}})(s,a)] - \mathbb{E}_{(s,a)\sim\rho}[f(\hat{P}_{\mathrm{MLE}})(s,a)]| + \mathbb{E}_{(s,a)\sim\rho}[f(\hat{P}_{\mathrm{MLE}})(s,a)] \\
&\lesssim G(\mathcal{M}_1) + \xi_n^2 + \frac{\ln(c/\delta)}{n}.
\end{aligned}$$

Thus, $P^\star \in \mathcal{M}_{\mathcal{D}}$ recalling the definition of $\mathcal{M}_{\mathcal{D}}$ (we set $\xi = G(\mathcal{M}_1) + \xi_n^2 + \frac{\ln(c/\delta)}{n}$).

Second step: Show the upper bound of $\mathbb{E}_{(s,a)\sim\rho}[f(P)(s,a)]$ for any $P \in \mathcal{M}_{\mathcal{D}}$. For any $P \in \mathcal{M}_{\mathcal{D}}$, as the proof of Theorem 1, we can prove $P \in \mathcal{M}_2$. Thus,

$$\sup_{P\in\mathcal{M}_{\mathcal{D}}} \mathbb{E}_{(s,a)\sim\rho}[f(P)(s,a)] \leq Z_n.$$

Thus, we will hereafter analyze Z_n. Here from (12), for any $P \in \mathcal{M}_2$, we have

$$\begin{aligned}\mathbb{E}_{(s,a)\sim\rho}[f(P)(s,a)] &\leq \sup_{P\in\mathcal{M}_2} \left(|(\mathbb{E}_{(s,a)\sim\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho})[f(P)(s,a)]|\right) + \sup_{P\in\mathcal{M}_2} \mathbb{E}_{(s,a)\sim\mathcal{D}}[f(P)(s,a)] \\ &\lesssim G(\mathcal{M}_2) + \sqrt{Z_n \frac{\ln(c/\delta)}{n}} + G(\mathcal{M}_1) + \xi_n^2 + \frac{\ln(c/\delta)}{n}.\end{aligned}$$

Hence,

$$Z_n \leq G(\mathcal{M}_2) + \sqrt{Z_n \frac{\ln(c/\delta)}{n}} + G(\mathcal{M}_1) + \xi_n^2 + \frac{\ln(c/\delta)}{n}.$$

This shows

$$Z_n \leq G(\mathcal{M}_2) + G(\mathcal{M}_1) + \xi_n^2 + \frac{\ln(c/\delta)}{n}.$$

Third step: Calculate the final error bound. Following the proof of Theorem 1, we can prove

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq (1-\gamma)^{-2} c_1 \sqrt{C_{\pi^*}^{\dagger}} \sqrt{G(\mathcal{M}_2) + G(\mathcal{M}_1) + \xi_n^2 + \frac{\ln(c/\delta)}{n}}.$$

□

F.3 Proofs for general function approximation with different concentrability coefficients

Proof of Theorem 6. We follow the proof of Theorem 1. We only change the third step. For any $P \in \mathcal{M}_{\mathcal{D}}$, we have

$$\begin{aligned}V_{P^\star}^{\pi^*} - V_{P}^{\pi^*} &\leq (1-\gamma)^{-2} \mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))] && \text{(Simulation lemma, Lemma 5)} \\ &\leq (1-\gamma)^{-2} \sqrt{\acute{C}_{\pi^\star}} \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))] \\ &\leq (1-\gamma)^{-2} \sqrt{\acute{C}_{\pi^\star} \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]} && \text{(Jensen's inequality)} \\ &\leq (1-\gamma)^{-2} \sqrt{\acute{C}_{\pi^\star} \ln(|\mathcal{M}|/\delta)/n}. && \text{(Consequence of the second step)}\end{aligned}$$

Then, the statement using $\acute{C}_{\pi^\star}$ is concluded.

Similarly, for any $P \in \mathcal{M}_{\mathcal{D}}$, we have

$$\begin{aligned}V_{P^\star}^{\pi^*} - V_{P}^{\pi^*} &\leq (1-\gamma)^{-2} \mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))] && \text{(Simulation lemma, Lemma 5)} \\ &\leq (1-\gamma)^{-2} \sqrt{C_{\pi^\star,2} \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]} && \text{(CS inequality)} \\ &\leq (1-\gamma)^{-2} \sqrt{C_{\pi^\star,2} \ln(|\mathcal{M}|/\delta)/n}.\end{aligned}$$

Then, the statement using $C_{\pi^\star,2}$ is concluded.

□

F.4 Proofs for Tabular MDPs

Proof of Corollary 1. We prove in a similar way as Theorem 1.

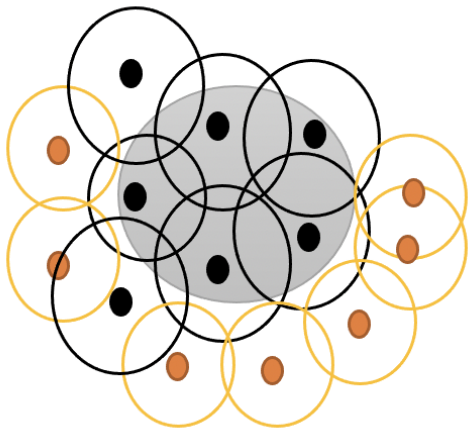

Figure 1: $\mathcal{M}_{\mathcal{D}}$ is colored in gray. $\mathcal{M}'$ corresponds to the set of black dots. Orange dots correspond to $\bar{\mathcal{M}}$, which do not belong to $\mathcal{M}'$.

First step We set $\xi = c\frac{|\mathcal{S}|^2|\mathcal{A}|\ln(n|\mathcal{S}|\mathcal{A}|c_2/\delta)\}}{n}$. Then, from Lemma 7, with probability $1-\delta$, we can show $P^\star \in \mathcal{M}_{\mathcal{D}}$ since

$$\mathbb{E}_{s,a\sim\mathcal{D}}\left[\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2\right] \leq \xi.$$

Hereafter, we condition on the above event.

Second step. Following the second step in the proof of Theorem 1 based on (8), for any $P \in \mathcal{M}_{\mathcal{D}}$, we have

$$\mathbb{E}_{s,a\sim\rho}\left[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2\right] \leq c\xi + A(P) \tag{13}$$

where

$$A(P) \coloneqq |\mathbb{E}_{s,a\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] - \mathbb{E}_{\mathcal{D}}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]|.$$

Our goal here is showing with probability $1-\delta$,

$$A(P) \lesssim \xi, \forall P \in \mathcal{M}_{\mathcal{D}}. \tag{14}$$

To prove (14), consider an ϵ-net $\{P_1(s,a), \cdots, P_K(s,a)\}$ covering a simplex in terms of $\|\cdot\|_1$ [3] for each fixed pair $(s,a) \in \mathcal{S}\times\mathcal{A}$. We take $\epsilon = 1/n$. Since the covering number K is upper-bounded by $(c/\epsilon)^{|\mathcal{S}|}$ (Wainwright, 2019, Lemma 5.7), we can obtain $\bar{M} = \{P_1, \cdots, P_{K^{|\mathcal{S}|\times|\mathcal{A}|}}\}$ s.t. for any possible $P \subset \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$, there exists P_i s.t.

$$\mathrm{TV}(P_i(\cdot \mid s,a), P(\cdot \mid s,a)) \leq \epsilon, \forall (s,a).$$

This implies for any $P \subset \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$, there exists $P_i(\cdot \mid s,a)$ s.t. $\forall(s,a)$,

$$\begin{aligned}
&|\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2 - \mathrm{TV}(P_i(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2| \\
&\leq 4|\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a)) - \mathrm{TV}(P_i(\cdot \mid s,a), P^\star(\cdot \mid s,a))| \quad (a^2-b^2=(a-b)(a+b)) \\
&\leq 4\mathrm{TV}(P\cdot \mid s,a), P_i(\cdot \mid s,a)) \quad (|\|a\|-\|b\|| \leq \|a-b\|) \\
&\leq 4\epsilon. \qquad (15)
\end{aligned}$$

We often use this property (15) hereafter.

Next, we define $\mathcal{M}' \subset \bar{M}$ so that it covers $\mathcal{M}_{\mathcal{D}}$. Concretely, we define $\mathcal{M}'$:

$$\mathcal{M}' = \{P \in \bar{M} : \exists P'' \in \mathcal{M}_{\mathcal{D}}, \mathrm{TV}(P(\cdot \mid s,a), P''(\cdot \mid s,a)) \leq \epsilon \quad \forall(s,a)\}. \tag{16}$$

[3] In the tabular setting, since the state space is countable, it is equivalent to L1 distance.

The construction is illustrated in Figure 1. Here, from the definition, for any $P \in \mathcal{M}_{\mathcal{D}}$, we can also find $P' \in \mathcal{M}'$ s.t.

$$\mathrm{TV}(P(\cdot \mid s,a), P'(\cdot \mid s,a)) \leq \epsilon, \forall (s,a).$$

This is because from the definition of $\bar{M}$, we can always find $P \in \bar{\mathcal{M}}$ satisfying the above. Such P belongs to $\mathcal{M}'$ from the definition of $\mathcal{M}'$. We use this fact later.

Then, from (16) and recalling (13), we have

$$\mathbb{E}_{s,a\sim\rho}\left[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2\right] \lesssim \xi + A(P), \quad \forall P \in \mathcal{M}'. \tag{17}$$

because

$$\begin{aligned}
&\mathbb{E}_{s,a\sim\rho}\left[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2\right] \\
&\leq \mathbb{E}_{s,a\sim\rho}\left[\mathrm{TV}(P(\cdot \mid s,a), P''(\cdot \mid s,a))^2\right] + \mathbb{E}_{s,a\sim\rho}\left[\mathrm{TV}(P''(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2\right] \\
&\leq \mathbb{E}_{s,a\sim\rho}\left[\mathrm{TV}(P''(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2\right] + \epsilon^2 && \text{(Take some } P'' \in \mathcal{M}_{\mathcal{D}} \text{ noting (16))} \\
&\leq c\xi + A(P). && \text{(From (13))}
\end{aligned}$$

Then, with probability $1-\delta$, from Bernstein's inequality, we have

$$A(P) \leq \sqrt{c\frac{\mathrm{var}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]\ln(K^{|\mathcal{S}|\times|\mathcal{A}|}/\delta)}{n}} + \frac{c\ln(K^{|\mathcal{S}|\times|\mathcal{A}|}/\delta)}{n}, \forall P \in \mathcal{M}.$$

Hereafter, we condition on the above event. Based on (17), we can state

$$\mathrm{var}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \lesssim \mathrm{E}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \lesssim \xi + A(P), \quad \forall P \in \mathcal{M}',$$

with probability $1-\delta$. Following the argument of Theorem 1, for $P \in \mathcal{M}'$, we have

$$A^2(P) - A(P)B_1 - B_2 \leq 0, \quad B_1 = \frac{\ln(K^{|\mathcal{S}|\times|\mathcal{A}|}/\delta)}{n}, \quad B_2 = \xi\frac{\ln(K^{|\mathcal{S}|\times|\mathcal{A}|}/\delta)}{n} + \left(\frac{\ln(K^{|\mathcal{S}|\times|\mathcal{A}|}/\delta)}{n}\right)^2.$$

Then, with probability $1-\delta$, we have

$$A(P) \leq \frac{\ln(K^{|\mathcal{S}|\times|\mathcal{A}|}/\delta)}{n} + \sqrt{\frac{\ln(K^{|\mathcal{S}|\times|\mathcal{A}|}/\delta)}{n}}\xi^{1/2} \lesssim \xi, \quad \forall P \in \mathcal{M}'. \tag{18}$$

This shows for any $P \in \mathcal{M}_{\mathcal{D}}$, we have

$$\begin{aligned}
&|\{\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho}\}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]| \\
&\leq |\{\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho}\}[\mathrm{TV}(P'(\cdot \mid s,a), P(\cdot \mid s,a))^2 + \mathrm{TV}(P'(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2]| \\
& && \text{(We take } P' \in \mathcal{M}' \text{ such that (16))} \\
&\leq |\{\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho}\}[\mathrm{TV}(P'(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] + 8\epsilon && \text{(From the definition of } \mathcal{M}') \\
&\lesssim \xi. && \text{(From (18) and } P' \in \mathcal{M}')
\end{aligned}$$

Thus, (14) is proved.

Third step. We follow the third step of Theorem 1:

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \lesssim (1-\gamma)^{-2}\sqrt{C_{\pi^*}^{\dagger}\xi}.$$

□

F.5 Proofs for Linear Mixture MDPs

Proof of Corollary 2. Here, letting $P(\theta) = \theta^\top \psi(s,a,s')$, recall

$$\mathcal{M}_{\text{Mix}} = \left\{P(\theta) \mid \theta \in \Theta \subset \mathbb{R}^d, \int \theta^\top \psi(s,a,s')\mathrm{d}(s') = 1 \quad \forall (s,a)\right\}, \mathcal{H} = \left\{\sqrt{\frac{P+P^\star}{2}} \mid P \in \mathcal{M}_{\text{Mix}}\right\}.$$

Upper-bounding $\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2]$**.**
By invoking Theorem 4, we first show

$$\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2] \leq c\{(d/n)\ln^2(nR) + \ln(c/\delta)/n\}.$$

To do that, we calculate the entropy integral with bracketing. First, we have

$$\mathcal{N}_{[]}(\epsilon, \mathcal{H}, d) \leq \mathcal{N}_{[]}(\epsilon, \mathcal{M}_{\text{Mix}}, d'). \tag{19}$$

where

$$d'(a,b) = \mathbb{E}_{(s,a)\sim\rho}\left[\int (a(s,a,s') - b(s,a,s'))^2 \mathrm{d}(s')\right]^{1/2}, \tag{20}$$

$$d(a,b) = \mathbb{E}_{(s,a)\sim\rho}\left[\int (\sqrt{a(s,a,s')} - \sqrt{b(s,a,s')})^2 \mathrm{d}(s')\right]^{1/2}. \tag{21}$$

Here, we use two observations. The first observation is

$$d^2\left(\sqrt{\frac{P(\theta') + P^\star}{2}}, \sqrt{\frac{P(\theta'') + P^\star}{2}}\right) \leq c_1 d'^2(P(\theta'), P(\theta''))$$

due to the mean-value theorem

$$\sqrt{a} - \sqrt{b} \leq \max(1/\sqrt{a}, 1/\sqrt{b})(a-b)$$

and assumption $P^\star(s' \mid s,a) \geq c_0 > 0$. The second observation is when we have $P' < g < P''$, we also have $\sqrt{(P' + P^\star)/2} < \sqrt{(g + P^\star)/2} < \sqrt{(P'' + P^\star)/2}$. Then, (19) is concluded.

Next, by letting $\theta^{(1)}, \cdots, \theta^{(K)}$ be an ϵ-cover of the d-dimensional ball with a radius R, i.e, $B_d(R)$, we have the brackets $\{[P(\theta^{(i)}) - \epsilon, P(\theta^{(i)}) + \epsilon]\}_{i=1}^K$ which cover $\mathcal{M}_{mix}$. This is because for any $P(\theta) \in \mathcal{M}_{mix}$, we can take $\theta^{(i)}$ s.t. $\|\theta - \theta^{(i)}\|_2 \leq \epsilon$, then,

$$P(\theta^{(i)}) - \epsilon < P(\theta) < P(\theta^{(i)}) + \epsilon, \quad \forall (s,a,s')$$

noting

$$|P(\theta)(s,a,s') - P(\theta^{(i)})(s,a,s')| \leq \|\theta - \theta^{(i)}\|_2 \leq \epsilon, \quad \forall (s,a,s') \tag{22}$$

The last equality is from Lemma 10.

The brackets above are size of ϵ. Therefore, we have

$$\mathcal{N}_{[]}(\epsilon, \mathcal{M}_{mix}, \|\cdot\|_2) \leq \mathcal{N}(\epsilon, B_d(cR), \|\cdot\|_2),$$

where $\mathcal{N}(\epsilon, B_d(cR), \|\cdot\|_2)$ is a covering number of $B_d(cR)$ w.r.t $\|\cdot\|_2$. This is upper-bounded by $(cR/\epsilon)^d$ (Wainwright, 2019, Lemma 5.7). Thus, we can calculate the upper bound of the entropy integral $J_B(\delta, \mathcal{M}_{mix}, \|\cdot\|_2)$:

$$\begin{aligned}
\int_0^\delta d^{1/2} \ln^{1/2}(cR/u)\mathrm{d}u &\leq \int_0^\delta d^{1/2}\ln(1/u)\mathrm{d}u + \delta d^{1/2}\ln(c_1 R) \\
&= cd^{1/2}(\delta + \delta\ln(1/\delta)) + \delta d^{1/2}\ln(cR) \\
&\leq cd^{1/2}\delta\ln(cR/\delta).
\end{aligned}$$

By taking $G(x) = d^{1/2}x\ln(cR/x)$ in Theorem 4, $\delta_n = (d/n)^{1/2}\ln(nR)$ satisfies the critical inequality

$$\sqrt{n}\delta_n^2 \geq d^{1/2}\delta_n \ln(cR/\delta_n).$$

Finally, with probability $1-\delta$

$$\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2] \leq \xi', \quad \xi' \coloneqq \{(d/n)\ln^2(nR) + \ln(c/\delta)/n\}. \tag{23}$$

Hereafter, we condition on this event.

Upper bounding $\mathbb{E}_{\mathcal{D}}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2]$. We take an ϵ-cover of the ball $B_d(R)$ in terms of $\|\cdot\|_2$, i.e., $\bar{M} = \{\theta^{(1)}, \cdots, \theta^{(K)}\}$, where $K = (cR/\epsilon)^d$. We take $\epsilon = 1/n$. Then, for any $\theta \in B_d(R)$, there exists $\theta^{(i)}$ s.t. $\forall (s,a)$,

$$
\begin{aligned}
&|\mathrm{TV}(P(\theta)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2 - \mathrm{TV}(P(\theta^{(i)})(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2| \\
&\leq 4|\mathrm{TV}(P(\theta)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a)) - \mathrm{TV}(P(\theta^{(i)})(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))| \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (a^2 - b^2 = (a-b)(a+b)) \\
&\leq 4\mathrm{TV}(P(\theta)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a)) \qquad\qquad\qquad\qquad (|\|a\| - \|b\|| \leq \|a-b\|) \\
&\leq 4\|\theta - \theta^{(i)}\|_2 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{From Lemma 10.}) \\
&\leq 4\epsilon. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (24)
\end{aligned}
$$

Hereafter, we condition on the event:

$$
\begin{aligned}
&|(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho})[\mathrm{TV}(P(\theta)(\cdot \mid s,a), P(\theta^\star)(\cdot \mid s,a))^2]| \qquad (25) \\
&\lesssim \sqrt{\frac{\mathrm{var}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta)(\cdot \mid s,a), P(\theta^\star)(\cdot \mid s,a))^2]\ln(K/\delta)}{n}} + \frac{\ln(K/\delta)}{n}, \quad \forall \theta \in \bar{M}.
\end{aligned}
$$

This event holds with probability $1-\delta$ from Bernstein's inequality.

Here, note for $\hat{\theta}_{\mathrm{MLE}}$, we have $\theta^{(i)}$ s.t. $\|\hat{\theta}_{\mathrm{MLE}} - \theta^{(i)}\|_2 \leq \epsilon$. Then, following the first step of Theorem 1,

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\mathcal{D}}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2 \\
&\lesssim \mathbb{E}_{(s,a)\sim\mathcal{D}}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2 + \epsilon \qquad\qquad (\text{From (24)}) \\
&\lesssim (\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho})\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2 + \epsilon + \mathbb{E}_{(s,a)\sim\rho}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2 \\
&\lesssim \sqrt{\frac{\mathrm{var}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2]\ln(K/\delta)}{n}} + \frac{\ln(K/\delta)}{n} \\
&+ \epsilon + \mathbb{E}_{(s,a)\sim\rho}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2 \qquad\qquad (\text{From (25)}) \\
&\lesssim \sqrt{\frac{\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2]\ln(K/\delta)}{n}} + \frac{\ln(K/\delta)}{n} \\
&+ \epsilon + \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2]. \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2 \leq 4)
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2] \\
&\lesssim \sqrt{\frac{\{\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2] + \epsilon\}\ln(K/\delta)}{n}} + \frac{\ln(K/\delta)}{n} \\
&+ \epsilon + \mathbb{E}_{(s,a)\sim\rho}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2 \\
&\lesssim \sqrt{\frac{\{\xi' + \epsilon\}\ln(K/\delta)}{n}} + \frac{\ln(K/\delta)}{n} + \epsilon + \xi'. \qquad\qquad (\text{From (23)})
\end{aligned}
$$

In the end, by taking $\epsilon = 1/n$, we have with probability $1-\delta$,

$$
\mathbb{E}_{\mathcal{D}}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\hat{\theta}_{\mathrm{MLE}})(\cdot \mid s,a))^2 \leq \xi, \quad \xi = c\{(d/n)\ln^2(nR) + \ln(c/\delta)/n\}.
$$

This also implies with probability $1-\delta$, $P^\star \in \mathcal{M}_{\mathcal{D}}$.

Show $\mathbb{E}_{(s,a)\sim\rho}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2 \lesssim \xi, \forall P(\theta) \in \mathcal{M}_{\mathcal{D}}$.
We show for any $P \in \mathcal{M}_{\mathcal{D}}$, the distance between $P^\star$ is controlled in terms of TV distance. Our goal is showing

$$
\mathbb{E}_{(s,a)\sim\rho}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2 \lesssim \xi, \quad \forall P(\theta) \in \mathcal{M}_{\mathcal{D}}.
$$

First, following the second step of Theorem 1 based on equation 8, we have

$$\mathbb{E}_{(s,a)\sim\rho}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2 \leq A(\theta) + c\xi, \quad \forall P(\theta) \in \mathcal{M}_{\mathcal{D}} \tag{26}$$

where

$$A(\theta) \coloneqq |(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho})\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2|.$$

From now on, we again consider an ϵ-cover of the ball $B_d(R)$ in terms of $\|\cdot\|_2$, i.e., $\bar{M} = \{\theta^{(1)}, \cdots, \theta^{(K)}\}$, where $K = (c_1 R/\epsilon)^d$ ($\epsilon = 1/n$). This also covers the space $\mathcal{M}_{\mathcal{D}}$. We take $\mathcal{M}' = \{\theta^{(i_1)}, \theta^{(i_2)} \cdots, \} \subset \mathcal{M}$ which covers $\mathcal{M}_{\mathcal{D}}$, that is,

$$\mathcal{M}' = \left\{\theta \in \bar{\mathcal{M}} \mid \exists \theta' \text{s.t.} P(\theta') \in \mathcal{M}_{\mathcal{D}}, \|\theta - \theta'\|_2 \leq \epsilon\right\}.$$

Recall Figure 1, which illustrates this definition. Here, for any θ s.t. $\forall P(\theta) \in \mathcal{M}_{\mathcal{D}}$, we can take $\theta' \in \mathcal{M}'$ s.t. $\|\theta - \theta'\|_2 \leq 1/n$. This is because we can take $\theta \in \bar{\mathcal{M}}$ s.t. $\|\theta - \theta'\|_2 \leq \epsilon$ noting $\bar{\mathcal{M}}$ is an ϵ-net, but such θ belongs to $\mathcal{M}'$ from the definition of $\mathcal{M}'$.

Then, we have

$$\mathbb{E}_{(s,a)\sim\rho}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2 \leq A(\theta) + c\xi, \quad \forall \theta \in \mathcal{M}'. \tag{27}$$

This is because for any $\theta^{(i)} \in \mathcal{M}'$, we can take $P(\theta) \in \mathcal{M}_{\mathcal{D}}$ such that

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2 \\
&\leq \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta^{(i)})(\cdot \mid s,a))^2 - \mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2] \\
&+ \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2] \\
&\leq 4\epsilon + \mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2] && (\|\theta - \theta^{(i)}\|_2 \leq \epsilon \text{ and from (24)}) \\
&\lesssim A(\theta) + \xi. && (\text{ From (26)})
\end{aligned}$$

Here, from (25), we have

$$A(\theta) \leq \sqrt{c\frac{\mathrm{var}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2]\ln(K/\delta)}{n}} + \frac{c\ln(K/\delta)}{n}, \quad \forall \theta \in \mathcal{M}'$$

Based on the construction of $\mathcal{M}'$ and (27), we have

$$\mathrm{var}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2] \lesssim A(\theta) + \xi, \quad \forall \theta \in \mathcal{M}'.$$

Then, following the second step of Theorem 1, $A(\theta)$ satisfies

$$A^2(\theta) - A(\theta)B_1 - B_2 \leq 0, \quad B_1 = \frac{\ln(K/\delta)}{n}, B_2 = \xi\frac{\ln(K/\delta)}{n} + \left(\frac{\ln(K/\delta)}{n}\right)^2.$$

Then, we have

$$A(\theta) \leq \frac{\ln(K/\delta)}{n} + \xi^{1/2}\sqrt{\frac{\ln(K/\delta)}{n}} \lesssim \xi, \quad \forall \theta \in \mathcal{M}'. \tag{28}$$

We combine all steps. Recall for any $\forall P(\theta) \in \mathcal{M}_{\mathcal{D}}$, we can take $\theta' \in \mathcal{M}'$ s.t. $\|\theta - \theta'\|_2 \leq 1/n$. Then, for any $P(\theta) \in \mathcal{M}_{\mathcal{D}}$, we have

$$\begin{aligned}
A(\theta) &= |(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho})\mathrm{TV}(P(\theta)(\cdot \mid s,a), P(\theta^\star)(\cdot \mid s,a))^2 \\
&\leq |(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho})[\mathrm{TV}(P(\theta)(\cdot \mid s,a), P(\theta^\star)(\cdot \mid s,a))^2 - \mathrm{TV}(P(\theta')(\cdot \mid s,a), P(\theta^\star)(\cdot \mid s,a))^2] \\
&+ (\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho})[\mathrm{TV}(P(\theta')(\cdot \mid s,a), P(\theta^\star)(\cdot \mid s,a))^2] \\
&\leq 8\epsilon + |(\mathbb{E}_{\mathcal{D}} - \mathbb{E}_{(s,a)\sim\rho})[\mathrm{TV}(P(\theta')(\cdot \mid s,a), P(\theta^\star)(\cdot \mid s,a))^2] && (\text{From equation 24}) \\
&\lesssim \xi. && (\text{From equation 28 and } \theta' \in \mathcal{M}')
\end{aligned}$$

Then, we have with probability $1-\delta$,

$$A(\theta) \lesssim \xi, \quad \forall P(\theta) \in \mathcal{M}_{\mathcal{D}}. \tag{29}$$

Finally, for any $P(\theta) \in \mathcal{M}_{\mathcal{D}}$, with probability $1-\delta$, we have

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2] &\leq A(\theta) + c\xi && (\text{From equation 26}) \\
&\lesssim \xi. && (\text{From equation 29})
\end{aligned}$$

Distribution shift part Here, for $P \in \mathcal{M}_{\mathcal{D}}$ we prove

$$V_{P^\star}^{\pi^*} - V_P^{\pi^*} \lesssim (1-\gamma)^{-2}\sqrt{dC_{\pi^*,\mathrm{mix}}\xi}, \tag{30}$$

$$V_{P^\star}^{\pi^*} - V_P^{\pi^*} \lesssim (1-\gamma)^{-2}\sqrt{C_{\pi^*}^{\dagger}\xi}. \tag{31}$$

Following the third step of the proof of Theorem 5, this immediately concludes the bound

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \lesssim (1-\gamma)^{-2}\sqrt{dC_{\pi^*,\mathrm{mix}}\xi},$$

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \lesssim (1-\gamma)^{-2}\sqrt{C_{\pi^*}^{\dagger}\xi}.$$

Since (31) is obvious from simulation lemma, we only prove (30). To prove (30), we take a distribution $P(\theta) \in \mathcal{M}_{\mathcal{D}}$. First, recall for $P(\theta) \in \mathcal{M}_{\mathcal{D}}$, we have

$$\mathbb{E}_{(s,a)\sim\rho}\mathrm{TV}(P(\theta^\star)(\cdot \mid s,a), P(\theta)(\cdot \mid s,a))^2 \lesssim \xi.$$

From the third statement of Lemma 10, for any $V : \mathcal{S} \to [0,1]$, we have

$$\mathbb{E}_{(s,a)\sim\rho}[|(\theta-\theta^*)^\top \psi_V(s,a)|^2] \lesssim \xi.$$

Thus,

$$\forall V : \mathcal{S} \to [0,1], \quad (\theta-\theta^*)^\top \Sigma_{\rho,V}(\theta-\theta^*) \lesssim \xi, \quad \Sigma_{\rho,V} = \mathbb{E}_{(s,a)\sim\rho}[\psi_V(s,a)\psi_V^\top(s,a)].$$

Here, we have

$$V_{P^\star}^{\pi^*} - V_P^{\pi^*} \leq (1-\gamma)^{-1}\left|\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[\int \{P(s' \mid s,a) - P^\star(s' \mid s,a)\}V_P^{\pi^*}(s')\mathrm{d}(s')\right]\right|$$

(Simulation lemma, Lemma 5)

$$\leq (1-\gamma)^{-1}\left|\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[(\theta-\theta^*)\psi_{V_P^{\pi^*}}(s,a)\right]\right|$$

(Recall $\psi_V = \int \psi(s,a,s')V_P^{\pi^*}(s')\mathrm{d}(s')$)

$$\leq (1-\gamma)^{-1}\underbrace{\|\theta-\theta^*\|_{\lambda I+\Sigma_{\rho,V_P^{\pi^*}}}}_{(a)}\underbrace{\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[\|\psi_{V_P^{\pi^*}}(s,a)\|_{(\Sigma_{\rho,V_P^{\pi^*}}+\lambda I)^{-1}}\right]}_{(b)}.$$

(CS inequality)

The first term (a) is upper-bounded by $\sqrt{\{(1-\gamma)^{-2}\xi + \lambda R^2\}}$ noting $0 \leq V_P^{\pi^*} \leq (1-\gamma)^{-1}$. The term (b) is upper-bounded by

$$\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[\|\psi_{V_P^{\pi^*}}(s,a)\|_{(\Sigma_{\rho,V_P^{\pi^*}}+\lambda I)^{-1}}\right] \leq \mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[\|\psi_{V_P^{\pi^*}}(s,a)\|^2_{(\Sigma_{\rho,V_P^{\pi^*}}+\lambda I)^{-1}}\right]^{1/2}$$

(Jensen's inequality)

$$= \sqrt{\mathrm{Tr}(\Sigma_{d_{P^\star}^{\pi^*},V_P^{\pi^*}}(\lambda I + \Sigma_{\rho,V_P^{\pi^*}})^{-1})}$$

$$\leq \sqrt{C_{\pi^*,\mathrm{mix}}\,\mathrm{Tr}(\Sigma_{\rho,V_P^{\pi^*}}(\lambda I + \Sigma_{\rho,V_P^{\pi^*}})^{-1})}$$

(From Lemma 11)

$$\leq \sqrt{C_{\pi^*,\mathrm{mix}}\mathrm{rank}(\Sigma_{\rho,V_P^{\pi^*}})} \leq \sqrt{C_{\pi^*,\mathrm{mix}}d}.$$

By taking λ s.t. $\lambda R^2 \lesssim (1-\gamma)^{-2}\xi$, (30) is proved.

For linear MDPs, from the fourth statement of Lemma 10, $C_{\pi^*,\mathrm{mix}} \leq \bar{C}_{\pi^*}$. Then, the statement is concluded. □

F.6 PROOFS FOR KNRS

Proof of Corollary 3. We prove in a similar way as Theorem 1.

First Step Recall

$$\xi = \sqrt{2\lambda\|W^\star\|_2^2 + 8\zeta^2\left(d_{\mathcal{S}}\ln(5) + \ln(1/\delta) + \bar{\mathcal{I}}_n\right)}, \quad \bar{\mathcal{I}}_n = \ln\left(\det(\Sigma_n)/\det(\lambda\mathbf{I})\right).$$

Thus, from Lemma 8, with probability $1-\delta$, we can show $W^* \in \mathcal{W}_{\mathcal{D}}$ since

$$\left\|\left(\widehat{W}_{\mathrm{MLE}} - W^\star\right)(\Sigma_n)^{1/2}\right\|_2 \le \xi.$$

Hereafter, we condition on this event.

Second step For any $W \in \mathcal{W}_{\mathcal{D}}$, with probability $1-\delta$, we have

$$\left\|(W - W^\star)(\Sigma_n)^{1/2}\right\|_2 \le \left\|\left(W - \widehat{W}\right)(\Sigma_n)^{1/2}\right\|_2 + \left\|\left(W^* - \widehat{W}\right)(\Sigma_n)^{1/2}\right\|_2 \le \xi.$$

Third step Note $P^\star = P(W^*)$. Then,

$$\begin{aligned}
V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} &\le V_{P^\star}^{\pi^*} - \min_{W\in\mathcal{W}_{\mathcal{D}}} V_{P(W)}^{\pi^*} + \min_{W\in\mathcal{W}_{\mathcal{D}}} V_{P(W)}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \\
&\le V_{P^\star}^{\pi^*} - \min_{W\in\mathcal{W}_{\mathcal{D}}} V_{P(W)}^{\pi^*} + \min_{W\in\mathcal{W}_{\mathcal{D}}} V_{P(W)}^{\hat{\pi}} - V_{P^\star}^{\hat{\pi}} && \text{(definition of } \hat{\pi}) \\
&\le V_{P^\star}^{\pi^*} - \min_{W\in\mathcal{W}_{\mathcal{D}}} V_{P(W)}^{\pi^*}. && \text{(Fist step, } W^* \in \mathcal{W}_{\mathcal{D}})
\end{aligned}$$

Then, by setting $W' = \arg\min_{W\in\mathcal{M}_{\mathcal{D}}} V_{P(W)}^{\pi^*}$, we have

$$\begin{aligned}
V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} &\le (1-\gamma)^{-2}\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[\|P'(s,a) - P^\star(s,a)\|_{\mathrm{TV}}\right] \\
&\le \frac{(1-\gamma)^{-2}}{\zeta}\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[\|(W' - W^\star)\phi(s,a)\|_2\right] && \text{(Lemma 9)} \\
&\le \frac{(1-\gamma)^{-2}}{\zeta}\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[\left\|(W' - W^\star)(\Sigma_n)^{1/2}\right\|_2 \|\phi(s,a)\|_{\Sigma_n^{-1}}\right] && \text{(CS inequality)} \\
&\le \frac{(1-\gamma)^{-2}}{\zeta}\xi\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[\|\phi(s,a)\|_{\Sigma_n^{-1}}\right] && \text{(Second step)}
\end{aligned}$$

From Chang et al. (2021, Theorem 20), with probability $1-\delta$, we have

$$\xi \le c_1\sqrt{\|W^*\|_2 + d_{\mathcal{S}}\min(\mathrm{rank}(\Sigma_\rho)\{\mathrm{rank}(\Sigma_\rho) + \ln(c_2/\delta)\}, d)\ln(1+n)}.$$

In addition, from Chang et al. (2021, Theorem 21), with probability $1-\delta$, we also have

$$\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}\left[\|\phi(s,a)\|_{\Sigma_n^{-1}}\right] \le c_1\sqrt{\frac{\bar{C}_{\pi^*}\mathrm{rank}[\Sigma_\rho]\{\mathrm{rank}[\Sigma_\rho] + \ln(c_2/\delta)\}}{n}}.$$

Finally, by combining all things, we have

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \le c_1(1-\gamma)^{-2}\min(d^{1/2}, \bar{R})\sqrt{\bar{R}}\sqrt{\frac{d_{\mathcal{S}}\bar{C}_{\pi^*}\ln(1+n)}{n}}, \bar{R} = \mathrm{rank}[\Sigma_\rho]\{\mathrm{rank}[\Sigma_\rho] + \ln(c_2/\delta)\}.$$

□

Proof of Corollary 4. We prove in a similar way as Theorem 1.

First step Recall

$$\xi = \sqrt{d_{\mathcal{S}}\{2 + 150\ln^3(d_{\mathcal{S}}n/\delta)\mathcal{I}_n\}}, \quad \mathcal{I}_n = \ln(\det(\mathbf{I} + \zeta^{-2}\mathbf{K}_n)).$$

From Chang et al. (2021, Leemma 14), with probability $1-\delta$, we can show $g^* \in \mathcal{G}_{\mathcal{D}}$ since

$$\sum_{i=1}^{d_{\mathcal{S}}}\|\hat{g}_i - g_i^*\|_{k_n}^2 \le \xi^2.$$

Hereafter, we condition on this event.

Second step For any $g \in \mathcal{G}_{\mathcal{D}}$, with probability $1-\delta$, we have

$$\sum_{i=1}^{d_{\mathcal{S}}} \|g_i - g_i^*\|_{k_n}^2 \le \sum_{i=1}^{d_{\mathcal{S}}} \|g_i - \hat{g}_i\|_{k_n}^2 + \sum_{i=1}^{d_{\mathcal{S}}} \|g_i^* - \hat{g}_i\|_{k_n}^2 \le 2\xi.$$

Third step Note $P^\star = P(g^*)$. Then,

$$\begin{aligned}
V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} &\le V_{P^\star}^{\pi^*} - \min_{g \in \mathcal{G}_{\mathcal{D}}} V_{P(g)}^{\pi^*} + \min_{g \in \mathcal{G}_{\mathcal{D}}} V_{P(g)}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \\
&\le V_{P^\star}^{\pi^*} - \min_{g \in \mathcal{G}_{\mathcal{D}}} V_{P(g)}^{\pi^*} + \min_{g \in \mathcal{W}_{\mathcal{D}}} V_{P(g)}^{\hat{\pi}} - V_{P^\star}^{\hat{\pi}} && \text{(definition of } \hat{\pi}) \\
&\le V_{P^\star}^{\pi^*} - \min_{g \in \mathcal{G}_{\mathcal{D}}} V_{P(g)}^{\pi^*}. && \text{(Fist step, } g^* \in \mathcal{G}_{\mathcal{D}})
\end{aligned}$$

Then, by setting $g' = \arg\min_{g \in \mathcal{G}_{\mathcal{D}}} V_{P(g)}^{\pi^*}$, we have

$$\begin{aligned}
V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} &\le (1-\gamma)^{-2} \mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}} [\|P'(s,a) - P^\star(s,a)\|_{\mathrm{TV}}] \\
&\le \frac{(1-\gamma)^{-2}}{\zeta} \mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}} [\|g'(s,a) - g(s,a)\|_2] && \text{(Lemma 9)} \\
&\le \frac{(1-\gamma)^{-2}}{\zeta} \mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}} \left[\sqrt{k_n((s,a),(s,a))} \left(\sum_{i=1}^{d_{\mathcal{S}}} \|g_i' - g_i\|_{k_n}^2 \right)^{1/2} \right] \\
& && \text{(CS inequality)} \\
&\le \frac{(1-\gamma)^{-2}\xi}{\zeta} \mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}} [\sqrt{k_n((s,a),(s,a))}]. && \text{(Second step)}
\end{aligned}$$

From Chang et al. (2021, Theorem 24), with probability $1-\delta$, we have

$$\xi \le c_1 \sqrt{d_{\mathcal{S}} \ln^3(c_2 d_{\mathcal{S}} n/\delta)\{d^* + \ln(c_2/\delta)\} d^* \ln(1+n)}.$$

In addition, from Chang et al. (2021, Theorem 25), with probability $1-\delta$, we have

$$\mathbb{E}_{x\sim d_{P^\star}^{\pi^*}} [\sqrt{k_n(x,x)}] \le c_1 \sqrt{\frac{\bar{C}_{\pi^*} d^* \{d^* + \ln(c_2/\delta)\}}{n}}.$$

Combining all things together, with probability $1-\delta$, we have

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \le c_1 (1-\gamma)^{-2} \{d^* + \ln(c_2/\delta)\} d^* \sqrt{\frac{d_{\mathcal{S}} \bar{C}_{\pi^*} \ln^3(c_2 d_{\mathcal{S}} n/\delta) \ln(1+n)}{n}}.$$

□

F.7 Proofs for Low-rank MDPs

Proof of Theorem 2. Until the second step, we can perform the same analysis as Theorem 1. More concretely, with probability $1-\delta$, we have $P^\star \in \mathcal{M}_{\mathcal{D}}$ and

$$\mathbb{E}_{s,a\sim\rho}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))^2] \le \xi, \quad \forall P \in \mathcal{M}_{\mathcal{D}}, \xi := c\frac{\ln(|\mathcal{M}|/\delta)}{n}. \tag{32}$$

Hereafter, we condition on the above event.

Letting $f(s,a) = \mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))$, we use Lemma 4. Then,

$$\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi}} [f(s,a)] \le \mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi}} [\|\phi^\star(s,a)\|_{\Sigma_{\rho,\phi^\star}^{-1}}] \sqrt{n\gamma\omega_\pi \mathbb{E}_\rho[f^2(s,a)] + 4\gamma^2\lambda d} + \sqrt{(1-\gamma)\omega_\pi \mathbb{E}_\rho[f^2(s,a)]}$$

where $\Sigma = n\mathbb{E}_\rho[\phi^\star \phi^{\star\top}] + \lambda I$. We consider how to bound $\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[\|\phi^\star(s,a)\|_{\Sigma^{-1}_{\rho,\phi^\star}}]$. This is upper-bounded by

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[\|\phi^\star(s,a)\|_{\Sigma^{-1}_{\rho,\phi^\star}}] &\leq \sqrt{\mathrm{tr}(\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[\phi^\star \phi^{\star\top}]\Sigma^{-1}_{\rho,\phi^\star})} \\
&\leq \sqrt{\bar{C}_{\pi^*,\phi^\star}\, \mathrm{tr}(\mathbb{E}_{(s,a)\sim\rho}[\phi^\star \phi^{\star\top}]\Sigma^{-1}_{\rho,\phi^\star})} \qquad \text{(From Lemma 11)} \\
&\leq \sqrt{\bar{C}_{\pi^*,\phi^\star}\mathrm{rank}(\Sigma_\rho)/n}.
\end{aligned}$$

Here, in the last line, by letting the SVD of $\Sigma_\rho = \mathbb{E}_\rho[\phi\phi^\top]$ be $U\tilde{\Sigma}_\rho U^\top$ where $\tilde{\Sigma}_\rho$ is a $d\times d$ diagonal matrix and U is a $d\times d$ orthogonal matrix , we use

$$\begin{aligned}
\mathrm{tr}\left(\Sigma_\rho \Sigma^{-1}_{\rho,\phi^\star}\right) &= \mathrm{tr}(U\tilde{\Sigma}_\rho U^\top \{nU\tilde{\Sigma}_\rho U^\top + \lambda I\}^{-1}) = \mathrm{tr}(\tilde{\Sigma}_\rho U^\top \{nU\tilde{\Sigma}_\rho U^\top + \lambda I\}^{-1} U) \\
&= \mathrm{tr}(\tilde{\Sigma}_\rho U^\top \{U\{n\tilde{\Sigma}_\rho + \lambda I\}U^\top\}^{-1} U) \\
&= \mathrm{tr}(\tilde{\Sigma}_\rho U^\top U\{n\tilde{\Sigma}_\rho + \lambda I\}^{-1} U^\top U) \\
&= \mathrm{tr}(\tilde{\Sigma}_\rho \{n\tilde{\Sigma}_\rho + \lambda I\})^{-1} \leq \mathrm{rank}(\Sigma_\rho)/n.
\end{aligned}$$

Hence, when $P \in \mathcal{M}_\mathcal{D}$, by setting λ s.t. $\lambda d \lesssim n\omega_\pi \xi$, we have

$$\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[f(s,a)] \leq \sqrt{\frac{\gamma \bar{C}_{\pi^*,\phi^\star}\mathrm{rank}(\Sigma_\rho)\omega_\pi \ln(|\mathcal{M}|/\delta)}{n}} + \sqrt{\frac{(1-\gamma)\omega_\pi \ln(|\mathcal{M}|/\delta)}{n}}.$$

We use (32) here.

Finally,

$$\begin{aligned}
&V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} \\
&\leq V^{\pi^*}_{P^\star} - \min_{P\in\mathcal{M}_\mathcal{D}} V^{\pi^*}_P \qquad \text{(Recall the proof of the third step in the proof of Theorem 1)} \\
&\leq (1-\gamma)^{-2}\mathbb{E}_{s,a\sim d^{\pi^\star}_{P^\star}} \mathrm{TV}(P'(s,a), P^\star(\cdot \mid s,a)) \qquad (P' = \arg\min_{P\in\mathcal{M}_\mathcal{D}} V^{\pi^*}_P) \\
&\lesssim \sqrt{\frac{\bar{C}_{\pi^*,\phi^\star}\mathrm{rank}(\Sigma_\rho)\omega_{\pi^*}\ln(|\mathcal{M}|/\delta)}{n(1-\gamma)^4}}.
\end{aligned}$$

□

The following inequality is an important lemma to connect $\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}\{f(s,a)\}$ with an elliptical potential $\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{P^\star}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}}$.

Lemma 4 (One-step back inequality)**.** *Take any $f \subset \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ s.t. $\|f\|_\infty \leq B$ and $0 < \lambda \in \mathbb{R}$. Letting $\omega = \max_{s,a}(\pi(a \mid s)/\pi_b(a \mid s))$, for any policy π, we have*

$$\begin{aligned}
|\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}\{f(s,a)\}| \leq \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{P^\star}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}}\sqrt{\left\{n\omega_\pi\gamma\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right]\right\} + \gamma^2\lambda d B^2} \\
+ \sqrt{(1-\gamma)\omega_\pi\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right]}.
\end{aligned}$$

where $\Sigma = n\mathbb{E}_{(s,a)\sim\rho}[\phi^\star(s,a)\phi^{\star\top}(s,a)] + \lambda I$.

Proof of Lemma 4. First, we have an equality:

$$\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}\{f(s,a)\} = \gamma\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{P^\star}, s\sim P^\star(\tilde{s},\tilde{a})}\{f(s,a)\} + (1-\gamma)\mathbb{E}_{s\sim d_0, a\sim\pi(s_0)}\{f(s,a)\}. \tag{33}$$

The second term in (33) is upper-bounded by

$$\mathbb{E}_{s\sim d_0,a\sim\pi(s_0)}\{f(s,a)\} \leq \mathbb{E}_{s\sim d_0,a\sim\pi(s_0)}\left\{f^2(s,a)\right\}^{1/2} = \sqrt{\omega_\pi\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right]/(1-\gamma)}.$$

Next we consider the first term in (33). By CS inequality, we have

$$\left|\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}, s\sim P^\star(\tilde{s},\tilde{a})}\left\{f(s,a)\right\}\right| = \left|\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}} \phi^\star(\tilde{s},\tilde{a})^\top \int \hat{\mu}(s)\pi(a\mid s) f(s,a) d(s,a)\right|$$
$$\leq \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}} \| \int \hat{\mu}(s)\pi(a\mid s) f(s,a) d(s,a)\|_{\Sigma_{\rho,\phi^\star}}.$$

Then,

$$\begin{aligned}
&\| \int \hat{\mu}(s)\pi(a\mid s) f(s,a) d(s,a)\|^2_{\Sigma_{\rho,\phi^\star}} \\
&\leq \left\{\int \hat{\mu}(s)\pi(a\mid s) f(s,a) d(s,a)\right\}^\top \left\{n\mathbb{E}_{(s,a)\sim\rho}[\phi^\star \phi^{\star\top}] + \lambda I\right\} \left\{\int \hat{\mu}(s)\pi(a\mid s) f(s,a) d(s,a)\right\} \\
&\leq n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho}\left[\int \hat{\mu}(s)^\top \phi^\star(\tilde{s},\tilde{a})\pi(a\mid s) f(s,a) d(s,a)\right]\right\}^2 + B^2\lambda d \\
&\qquad\qquad \text{(Use the assumption } \|f(s,a)\|_\infty \leq B \text{ and } \|\textstyle\int \hat{\mu}(s)\mathrm{d}(s)\|_2 \leq \sqrt{d}) \\
&= n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho, s\sim P^\star(\tilde{s},\tilde{a}), a\sim\pi(s)}\left[f(s,a)\right]\right\}^2 + B^2\lambda d \\
&\leq n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho, s\sim P^\star(\tilde{s},\tilde{a}), a\sim\pi(s)}\left[f^2(s,a)\right]\right\} + B^2\lambda d. \qquad \text{(Jensen)}
\end{aligned}$$

Finally, the the first term in (33) is upper-bounded by

$$\begin{aligned}
&n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho, s\sim P^\star(\tilde{s},\tilde{a}), a\sim\pi(s)}\left[f^2(s,a)\right]\right\} + \lambda d B^2 \\
&\leq n\omega_\pi\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho, s\sim P^\star(\tilde{s},\tilde{a}), a\sim\pi_b(s)}\left[f^2(s,a)\right]\right\} + \lambda d B^2 \qquad \text{(Importance sampling)} \\
&\leq n\omega_\pi\left\{\frac{1}{\gamma}\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right]\right\} + \lambda d B^2. \qquad \text{(Definition of } \rho)
\end{aligned}$$

The final statement is immediately concluded.

□

F.8 Proofs for Factored MDPs

Proof of Theorem 3. We denote the constrained set as $\mathcal{M}_{\mathcal{D}}$:

$$\mathcal{M}_{\mathcal{D}} = \left\{P = \prod_i P_i \mid \mathbb{E}_{\mathcal{D}}\left[\mathrm{TV}(\widehat{P}_{\mathrm{MLE},i}(\cdot\mid s[pa_i],a), P_i(\cdot\mid s[pa_i],a))^2\right] \leq \xi_i, \forall i\in[1,\cdots,d]\right\}.$$

Following the first step in the proof of Corollary 1, with probability $1-\delta$, the product $\prod_i P_i^\star$ is in $\mathcal{M}_{\mathcal{D}}$, i.e.,

$$\mathbb{E}_{s,a\sim\mathcal{D}}\left[\mathrm{TV}(\widehat{P}_{\mathrm{MLE},i}(\cdot\mid s[pa_i],a), P_i^\star(\cdot\mid s[pa_i],a))^2\right] \leq \xi_i, \forall i\in[1,\cdots,d], \quad \xi_i = \sqrt{\tfrac{L_i\log(L_i d/\delta)}{n}}.$$

Note d comes from the union bound. Besides, following the second step in the proof of Corollary 1, for any $P\in\mathcal{M}_{\mathcal{D}}$, with probability $1-\delta$,

$$\mathbb{E}_{s,a\sim\rho}\left[TV(\widehat{P}_i(\cdot\mid s[pa_i],a), P_i^\star(\cdot\mid s[pa_i],a))^2\right] \leq \xi_i, \forall i\in[1,\cdots,d].$$

After conditioning on the above two events, then, for any $P \in \mathcal{M}_{\mathcal{D}}$ and $\pi^\star \in \Pi$, we have

$$
\begin{aligned}
V_{P^\star}^{\pi^*} - V_P^{\pi^*} &\leq (1-\gamma)^{-2}\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}[\mathrm{TV}(P(\cdot \mid s,a), P^\star(\cdot \mid s,a))] \\
&\hspace{20em} \text{(Simulation lemma, Lemma 5)} \\
&\leq (1-\gamma)^{-2}\mathbb{E}_{(s,a)\sim d_{P^\star}^{\pi^*}}[\sum_i \mathrm{TV}(P_i(\cdot \mid s[pa_i],a), P_i^\star(\cdot \mid s[pa_i],a))] \\
&\leq (1-\gamma)^{-2}\sum_i \sqrt{\mathbb{E}_{(s,a)\sim\rho}\left[\left(\frac{d_{P^\star}^{\pi^*}(s[pa_i],a)}{\rho(s[pa_i],a)}\right)^2\right]\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P_i(\cdot \mid s[pa_i],a), P_i^\star(\cdot \mid s[pa_i],a))^2]} \\
&\hspace{20em} \text{(CS inequality)} \\
&\leq (1-\gamma)^{-2}\sum_i \sqrt{\ddot{C}_{\pi^*,2}\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P_i(\cdot \mid s,a), P_i^\star(\cdot \mid s,a))^2]} \leq (1-\gamma)^{-2}\sum_i\sqrt{\ddot{C}_{\pi^*,2}\xi_i} \\
&\leq (1-\gamma)^{-2}\sqrt{d\ddot{C}_{\pi^*,2}\sum_i \xi_i} \qquad \text{(CS inequality)} \\
&\leq c(1-\gamma)^{-2}\sqrt{d\ddot{C}_{\pi^*,2}\frac{L\ln(Lnd/\delta)}{n}}.
\end{aligned}
$$

Here, recall

$$
\ddot{C}_{\pi^*,2} = \max_{i\in[1,\cdots,d]} \mathbb{E}_{(s,a)\sim\rho}\left[\left(\frac{d_{P^\star}^{\pi^*}(s[pa_i],a)}{\rho(s[pa_i],a)}\right)^2\right].
$$

Following the third step in the proof of Corollary 1, the statement is concluded. □

Next, we show that $\ddot{C}_{\pi^*,\infty} \leq C_{\pi^*,P^\star} = \max_{s,a}\frac{d_{P^\star}^{\pi^*}(s,a)}{\rho(s,a)}$.

Proposition 1 (Comparison of L^∞-density-ratio based concentrabiliity coefficient between factored MDPs and non-factored MDPs). *For any π^*, we have:*

$$
\ddot{C}_{\pi^*,\infty} \leq C_{\pi^*,\infty}. \tag{34}
$$

Proof. From now on, for any $i \in [1,\cdots,d]$, by defining $\mathcal{S}_i'$ s.t. $\mathcal{S} = \mathcal{S}_i \times \mathcal{S}_i'$, we prove

$$
\max_{s_i\in\mathcal{S}_i,a\in\mathcal{A}}\frac{d_{P^\star}^{\pi^*}(s_i,a)}{\rho(s_i,a)} \leq \max_{s\in\mathcal{S}_i,s_i'\in\mathcal{S}_i,a\in\mathcal{A}}\frac{d_{P^\star}^{\pi^*}(s_i,s_i',a)}{\rho(s_i,s_i',a)} = C_{\pi^*,\infty}.
$$

Then, (34) is easily proved.

First, for any $s_i \in \mathcal{S}_i, a\in\mathcal{A}$, we have

$$
\max_{s_i'}\frac{d_{P^\star}^{\pi^*}(s_i,s_i',a)}{\rho(s_i,s_i',a)} = \max_{s_i'}\frac{d_{P^\star}^{\pi^*}(s_i,a)d_{P^\star}^{\pi^*}(s_i' \mid s_i,a)}{\rho(s_i,a)\rho(s_i' \mid s_i,a)} = \frac{d_{P^\star}^{\pi^*}(s_i,a)}{\rho(s_i,a)}\max_{s_i'}\frac{d_{P^\star}^{\pi^*}(s_i' \mid s_i,a)}{\rho(s_i' \mid s_i,a)} \geq \frac{d_{P^\star}^{\pi^*}(s_i,a)}{\rho(s_i,a)}. \tag{35}
$$

Here, we use

$$
1 \leq \max_{s_i'}\frac{d_{P^\star}^{\pi^*}(s_i' \mid s_i,a)}{\rho(s_i' \mid s_i,a)},
$$

which is proved by the contradiction argument, that is, if $1 > \max_{s_i'}\frac{d_{P^\star}^{\pi^*}(s_i'|s_i,a)}{\rho(s_i'|s_i,a)}$, both ρ and $d_{P^\star}^{\pi^*}$ cannot be probability mass functions since we would get

$$
1 = \sum_{s_i'} d_{P^\star}^{\pi^*}(s_i' \mid s_i,a) \leq \max_{s_i'}\left(\frac{d_{P^\star}^{\pi^*}(s_i' \mid s_i,a)}{\rho(s_i' \mid s_i,a)}\right)\sum_{s_i'}\rho(s_i' \mid s_i,a) < \sum_{s_i'}\rho(s_i' \mid s_i,a).
$$

Then, by taking the maximum over $s_i \in \mathcal{S}_i, a\in\mathcal{A}$ for both sides on (35), we have

$$
\max_{s_i,a}\frac{d_{P^\star}^{\pi^*}(s_i,a)}{\rho(s_i,a)} \leq \max_{s_i,s_i',a}\frac{d_{P^\star}^{\pi^*}(s_i,s_i',a)}{\rho(s_1,s_i',a)}.
$$

□

Next, we show a similar statement for the density-ratio based concentrabilitty coefficients in terms of L2 norm. We define

$$C_{\pi^*,2} = \mathbb{E}_{(s,a)\sim\rho}\left[\left(\frac{d^{\pi^*}_{P^\star}(s,a)}{\rho(s,a)}\right)^2\right], \ddot{C}_{\pi^*,2} = \max_{j\in[1,\cdots,d]} \mathbb{E}_{(s,a)\sim\rho}\left[\left(\frac{d^{\pi^*}_{P^\star}(s[\mathrm{pa}_j],a)}{\rho(s[\mathrm{pa}_j],a)}\right)^2\right].$$

Note $C_{\pi^*,2}$ is always greater than 1 noting $C_{\pi^*,2} - 1$ is Chi-square divergence between $d^{\pi^*}_{P^\star}$ and ρ. Recall chi-square divergence between p and q is defined as

$$\chi(p \mid q) = \int \left(\left(\frac{p(x)}{q(x)}\right)^2 - 1\right) q(x)\mathrm{d}(x).$$

Similar argument is applied to $\ddot{C}_{\pi^*,2}$; and we have $\ddot{C}_{\pi^*,2} \geq 1$.

Finally, we prove $\ddot{C}_{\pi^*,2} \leq C_{\pi^*,2}$. This states that the density-ratio based concentrabilitty coefficient for a factored MDP in terms of L2 norm is smaller than the one of a non-factored tabular MDP.

Proposition 2 (Comparison of L^2-density-ratio based concentrabiliity coefficient of factored MDPs and non-factored MDPs)**.**

$$\ddot{C}_{\pi^*,2} \leq C_{\pi^*,2}.$$

Proof. By denoting $s_j = s[pa_i]$, we prove for any i,

$$\mathbb{E}_{(s,a)\sim\rho}\left[\left(\frac{d^{\pi^*}_{P^\star}(s_i,a)}{\rho(s_i,a)}\right)^2\right] \leq C_{\pi^*,2}.$$

Here, letting s'_i be a value s.t. $s = (s_i, s'_i)$, we have

$$\begin{aligned}
C_{\pi^*,2} &= \mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}\left[\frac{d^{\pi^*}_{P^\star}(s,a)}{\rho(s,a)}\right] = \mathbb{E}_{(s_i,s'_i,a)\sim d^{\pi^*}_{P^\star}}\left[\frac{d^{\pi^*}_{P^\star}(s_i,s'_i,a)}{\rho(s_i,s'_i,a)}\right] \\
&= \mathbb{E}_{(s_i,a)\sim d^{\pi^*}_{P^\star}}\left[\mathbb{E}_{s'_i\sim d^{\pi^*}_{P^\star}(s_i,a)}\left[\frac{d^{\pi^*}_{P^\star}(s_i,s'_i,a)}{\rho(s_i,s'_i,a)}\right]\right] \\
&= \mathbb{E}_{(s_i,a)\sim d^{\pi^*}_{P^\star}}\left[\frac{d^{\pi^*}_{P^\star}(s_i,a)}{\rho(s_i,a)}\mathbb{E}_{s'_i\sim d^{\pi^*}_{P^\star}(s_i,a)}\left[\frac{d^{\pi^*}_{P^\star}(s'_i \mid s_i,a)}{\rho(s'_i \mid s_i,a)}\right]\right] \\
&\geq \mathbb{E}_{(s_i,a)\sim d^{\pi^*}_{P^\star}}\left[\frac{d^{\pi^*}_{P^\star}(s_i,a)}{\rho(s_i,a)}\right] = \mathbb{E}_{(s,a)\sim\rho}\left[\left(\frac{d^{\pi^*}_{P^\star}(s_i,a)}{\rho(s_i,a)}\right)^2\right].
\end{aligned}$$

In the above inequality, we use

$$\mathbb{E}_{s'_i\sim d^{\pi^*}_{P^\star}(s_i,a)}\left[\frac{d^{\pi^*}_{P^\star}(s'_i \mid s_i,a)}{\rho(s'_i \mid s_i,a)}\right] - 1 \geq 0, \forall s_i \in \mathcal{S}_i, \forall a \in \mathcal{A}$$

as this is Chi-square divergence between two conditional distributions. □

G AUXILIARY LEMMAS

Lemma 5 (Simulation Lemma)**.** *Consider any two transitions P and $\widehat{P}$, and any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. We have:*

$$\begin{aligned}
|V^\pi_P - V^\pi_{\widehat{P}}| &\leq |(1-\gamma)^{-1}\mathbb{E}_{s,a\sim d^\pi_P}[\mathbb{E}_{s'\sim P(s,a)}[V^\pi_{\widehat{P}}(s')] - \mathbb{E}_{s'\sim P(s,a)}[V^\pi_{\widehat{P}}(s')]]| \\
&\leq (1-\gamma)^{-2}\mathbb{E}_{s,a\sim d^\pi_P}\left[\mathrm{TV}(P(\cdot|s,a), \widehat{P}(\cdot|s,a))\right].
\end{aligned}$$

Proof. Such simulation lemma is standard in model-based RL literature and the derivation can be found, for instance, in the proof of Lemma 10 from Sun et al. (2019). □

Lemma 6 (MLE guarantee)**.** *Given a set of models $\mathcal{M} = \{P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$ with $P^\star \in \mathcal{M}$, and a dataset $\mathcal{D} = \{s_i, a_i, s_i'\}_{i=1}^n$ with $s_i, a_i \sim \rho$, and $s_i' \sim P^\star(s_i, a_i)$, let $\widehat{P}_{\mathrm{MLE}}$ be*

$$\widehat{P}_{\mathrm{MLE}} = \arg\min_{P \in \mathcal{M}} \sum_{i=1}^n -\ln P(s_i'|s_i, a_i).$$

With probability at least $1-\delta$, we have:

$$\mathbb{E}_{s,a\sim\rho}\mathrm{TV}(\widehat{P}_{\mathrm{MLE}}(\cdot|s,a), P^\star(\cdot|s,a))^2 \lesssim \frac{\ln(|\mathcal{M}|/\delta)}{n}.$$

Proof. Refer to (Agarwal et al., 2020b, Section E) □

Lemma 7 (MLE guarantee for tabular models)**.**

$$\mathbb{E}_{\mathcal{D}}\left[\mathrm{TV}(P(\cdot|s,a), \widehat{P}_{\mathrm{MLE}}(\cdot|s,a))^2\right] \leq \frac{|\mathcal{S}|\mathcal{A}|\{|\mathcal{S}|\ln 2 + \ln(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{2n}.$$

Proof. From Chang et al. (2021, Lemma 12) , with probability $1-\delta$,

$$\mathrm{TV}(P(\cdot|s,a), \widehat{P}_{\mathrm{MLE}}(\cdot|s,a))^2 \leq \frac{|\mathcal{S}|\ln 2 + \ln(2|\mathcal{S}||\mathcal{A}|/\delta)}{2N(s,a)} \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A},$$

where $N(s,a)$ is the number of visiting times for (s,a). Then,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\mathrm{TV}(P(\cdot|s,a), \widehat{P}_{\mathrm{MLE}}(\cdot|s,a))^2\right] &\leq \mathbb{E}_{\mathcal{D}}\left[\frac{|\mathcal{S}|\ln 2 + \ln(2|\mathcal{S}||\mathcal{A}|/\delta)}{2N(s,a)}\right] \\
&\leq \sum_{(s,a)}\left[\frac{|\mathcal{S}|\ln 2 + \ln(2|\mathcal{S}||\mathcal{A}|/\delta)}{2n}\right] \\
&= \frac{|\mathcal{S}|\mathcal{A}|\{|\mathcal{S}|\ln 2 + \ln(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{2n}.
\end{aligned}$$

□

Lemma 8 (MLE guarantee for KNRs)**.**

$$\left\|\left(\widehat{W}_{\mathrm{MLE}} - W^\star\right)(\Sigma_n)^{1/2}\right\|_2 \leq \beta_n.$$

Proof. The proof directly follows the confidence ball construction and proof from (Kakade et al., 2020). □

Lemma 9 (ℓ_1 Distance between two Gaussians)**.** *Consider two Gaussian distributions $P_1 := \mathcal{N}(\mu_1, \zeta^2\mathbf{I})$ and $P_2 := \mathcal{N}(\mu_2, \zeta^2\mathbf{I})$. We have:*

$$\mathrm{TV}(P_1, P_2) \leq \frac{1}{\zeta}\|\mu_1 - \mu_2\|_2.$$

Proof. This lemma is proved by Pinsker's inequality and the closed-form of the KL divergence between P_1 and P_2. Refer to (Kakade et al., 2020). □

Lemma 10 (Property of linear mixture MDPs)**.** *Let $P(\theta) = \theta^\top\psi(s,a,s')$. Suppose $P(\theta) \in \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$. For any function $V \in \mathcal{S} \to [0,1]$, letting $\psi_V(s,a) = \int \psi(s,a,s')V(s')\mathrm{d}(s')$, we suppose $\|\psi_V(s,a)\|_2 \leq 1$. The following theorems hold:*

1. *For any (s,a,s'), we have $|P(\theta)(s,a,s') - P(\theta')(s,a,s')| \leq \|\theta-\theta'\|_2$.*
2. *For any (s,a), we have $\mathrm{TV}(P(\theta)(s,a,\cdot), P(\theta')(s,a,\cdot)) \leq \|\theta-\theta'\|_2$. Besides, for any $V : \mathcal{S} \to [0,1]$, we have*

$$|(\theta-\theta')\psi_V(s,a)| \leq \mathrm{TV}(P(\theta)(s,a,\cdot), P(\theta')(s,a,\cdot)).$$

3.

$$C^{\dagger}_{\pi^\star,P^\star} = \sup_x \frac{x^\top \mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[\psi_{V_{(s,a,x)}}(s,a)\psi^\top_{V_{(s,a,x)}}(s,a)]x}{x^\top \mathbb{E}_{(s,a)\sim\rho}[\psi_{V_{(s,a,x)}}(s,a)\psi^\top_{V_{(s,a,x)}}(s,a)]x},$$
$$V_{(s,a,x)} = \operatorname*{arg\,max}_{V:\mathcal{S}\to[0,1]} \left| x^\top \int \phi(s,a,s')V(s')\mathrm{d}(s') \right|.$$

4. In linear MDPs (i.e., $\psi(s,a,s') = \phi(s,a)\otimes\mu(s')$), we have

$$\sup_{V\in\{\mathcal{S}\to[0,1]\}} \sup_x \frac{x^\top \mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[\psi_V(s,a)\psi_V^\top(s,a)]x}{x^\top \mathbb{E}_{(s,a)\sim\rho}[\psi_V(s,a)\psi_V^\top(s,a)]x} = \sup_x \frac{x^\top \mathbb{E}_{d^{\pi^*}_{P^\star}}[\phi(s,a)\phi(s,a)^\top]x}{x^\top \mathbb{E}_{\rho}[\phi(s,a)\phi(s,a)^\top]x}.$$

Proof. We prove the first statement. This is proved by

$$|P(\theta) - P(\theta')| = |(\theta-\theta')\psi(s,a,s')| \le \|\theta-\theta'\|_2\|\psi(s,a,s')\|_2 \le \|\theta-\theta'\|_2,$$

Here, we use $\|\psi(s,a,s')\|_2 \le 1$ which is proved by the assumption by setting $V(s) = I(s' = s)$ for any s'.

Next, we prove the second statement. For fixed $\theta\in\mathbb{R}^d$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}
\mathrm{TV}(P(\theta)(s,a,\cdot), P(\theta^\star)(s,a,\cdot)) &= \sup_{V:\mathcal{S}\to[0,1]} |\int (\theta-\theta^\star)^\top\psi(s,a,s')V(s')\mathrm{d}(s')| \\
&= \sup_{V:\mathcal{S}\to[0,1]} |(\theta-\theta^\star)^\top \int \psi(s,a,s')V(s')\mathrm{d}(s')| \\
&= |(\theta-\theta^\star)^\top \int \psi(s,a,s')V_{(s,a,\theta)}(s')\mathrm{d}(s')| \\
&= |(\theta-\theta^\star)^\top \psi_{V_{(s,a,\theta)}}(s,a)|.
\end{aligned}$$

In the third line, we define $V(s,a,\theta) = \arg\max_{V:\mathcal{S}\to[0,1]} |(\theta-\theta^\star)^\top \int \psi(s,a,s')V(s')\mathrm{d}(s')|$.

Then, from CS inequality,

$$\mathrm{TV}(P(\theta)(s,a,\cdot), P(\theta^\star)(s,a,\cdot)) \le \|(\theta-\theta^\star\|_2\|\psi_{V_{(s,a,\theta)}}(s,a)|\|_2 \le \|\theta-\theta^\star\|_2.$$

We use the assumption $\|\psi_{V_{(s,a,\theta)}}(s,a)\|_2 \le 1$. This concludes the second statement. Besides, for any $V:\mathcal{S}\to[0,1]$, we have

$$\begin{aligned}
|(\theta-\theta')\psi_V(s,a)| &\le |(\theta-\theta^\star)^\top\psi_{V_{(s,a,\theta)}}(s,a)| \\
&\le \mathrm{TV}(P(\theta)(s,a,\cdot), P(\theta')(s,a,\cdot)).
\end{aligned}$$

The third statement is immediately concluded by

$$\frac{\mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[\mathrm{TV}(P(\theta)(s,a,\cdot), P(\theta^\star)(s,a,\cdot))^2]}{\mathbb{E}_{(s,a)\sim\rho}[\mathrm{TV}(P(\theta)(s,a,\cdot), P(\theta^\star)(s,a,\cdot))^2]} = \frac{\mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[|(\theta-\theta^\star)^\top\psi_{V_{(s,a,\theta)}}(s,a)|^2]}{\mathbb{E}_{(s,a)\sim\rho}[|(\theta-\theta^\star)^\top\psi_{V_{(s,a,\theta)}}(s,a)|^2]}. \quad (36)$$

Finally, we prove the fourth statement. Suppose $\psi(s,a,s') = \phi(s,a)\otimes\mu(s')$ ($\otimes$ denotes kronerker product). Then, $\phi_V(s,a,s') = \phi(s,a)\otimes\int\mu(s')V(s')\mathrm{d}(s')$. Then, by defining a vector $\mu(V) = \int\mu(s')V(s')\mathrm{d}(s')$, we immediately have

$$\frac{x^\top \mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[\psi_V(s,a)\psi_V^\top(s,a)]x}{x^\top \mathbb{E}_{(s,a)\sim\rho}[\psi_V(s,a)\psi_V^\top(s,a)]x} = \sup_x \frac{x^\top \mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^\star}}[(\phi(s,a)\otimes\mu(V))(\phi(s,a)\otimes\mu(V))^\top]x}{x^\top \mathbb{E}_{(s,a)\sim\rho}[(\phi(s,a)\otimes\mu(V))(\phi(s,a)\otimes\mu(V))^\top]x}. \quad (37)$$

Here, we have

$$\begin{aligned}
\mathbb{E}_\rho[(\phi(s,a)\otimes\mu(V))(\phi(s,a)\otimes\mu(V))^\top] &= \mathbb{E}_\rho[(\phi(s,a)\otimes\mu(V))(\phi(s,a)^\top\otimes\mu(V)^\top)] \\
&= \mathbb{E}_\rho[(\phi(s,a)\phi(s,a)^\top)]\otimes(\mu(V)\mu(V)^\top).
\end{aligned}$$

We notice

$$\{\mathbb{E}_\rho[(\phi(s,a)\phi(s,a)^\top)] \otimes (\mu(V)\mu(V)^\top)\}^{1/2} = \mathbb{E}_\rho[\phi(s,a)\phi(s,a)^\top]^{1/2} \otimes (\mu(V)\mu(V)^\top)^{1/2}.$$

This is because the square root of a matrix is unique and we have $(A^{1/2} \otimes B^{1/2})(A^{1/2} \otimes B^{1/2}) = AB$ for symmetric matrices A and B. Then, by denoting $F_\rho = \mathbb{E}_\rho[\phi(s,a)\phi(s,a)^\top], F_{d^\pi_{P^\star}} = \mathbb{E}_{d^\pi_{P^\star}}[\phi(s,a)\phi(s,a)^\top]$ and denoting the pseudo inverse of F as F^+, we can see (37) is equal to

$$\begin{aligned}
&\{F_\rho^{1/2} \otimes (\mu(V)\mu(V)^\top)^{1/2}\}^+ \{F_{d^\pi_{P^\star}} \otimes (\mu(V)\mu(V)^\top)\}\{F_\rho^{1/2} \otimes (\mu(V)\mu(V)^\top)^{1/2}\}^+ \\
&= \{F_\rho^{-1/2} \otimes (\mu(V)\mu(V)^\top)^{-1/2}\}\{F_{d^\pi_{P^\star}} \otimes (\mu(V)\mu(V)^\top)\}\{F_\rho^{-1/2} \otimes (\mu(V)\mu(V)^\top)^{-1/2}\} \\
&= \{F_\rho^{-1/2} F_{d^\pi_{P^\star}} F_\rho^{-1/2}\} \otimes \{(\mu(V)\mu(V)^\top)^{-1/2}(\mu(V)\mu(V)^\top)(\mu(V)\mu(V)^\top)^{-1/2}\} \\
&= \{F_\rho^{-1/2} F_{d^\pi_{P^\star}} F_\rho^{-1/2}\} \otimes I_k \, (k = \mathrm{rank}(\mu(V)\mu(V)^\top)).
\end{aligned}$$

Here, I_k is a diagonal matrix s.t. $k \in \mathbb{N}^+$ values in the diagonal entries are 1 and the rest of values are 0. Then, the maximum singular value of $\{F_\rho^{-1/2} F_{d^\pi_{P^\star}} F_\rho^{-1/2}\} \otimes I_k$ is equal to the one of $\{F_\rho^{-1/2} F_{d^\pi_{P^\star}} F_\rho^{-1/2}\}$. This is equal to

$$\sup_x \frac{x^\top F_{d^\pi_{P^\star}} x}{x^\top F_\rho x}$$

Hence, the fourth statement is concluded.

□

Lemma 11 (Distribution shift lemma)**.** *Suppose A_1, A_2, A_3 are semipositive definite matrices:*

$$\mathrm{Tr}(A_1 A_2) \le \sigma_{\max}(A_3^{-1/2} A_1 A_3^{-1/2}) \, \mathrm{Tr}(A_3 A_2).$$

Note

$$\sigma_{\max}(A_3^{-1/2} A_1 A_3^{-1/2}) = \sup_{x \in \mathbb{R}^d} \frac{x^\top A_1 x}{x^\top A_3 x}.$$

Proof.

$$\begin{aligned}
&\mathrm{Tr}(A_1 A_2) = \mathrm{Tr}(A_1^{1/2} A_2 A_1^{1/2}) = \mathrm{Tr}(A_1^{1/2} A_3^{-1/2} A_3^{1/2} A_2 A_3^{1/2} A_3^{-1/2} A_1^{1/2}) \\
&= \mathrm{Tr}(A_3^{-1/2} A_1 A_3^{-1/2} A_3^{1/2} A_2 A_3^{1/2}).
\end{aligned}$$

In addition, for any semipositive definite matrices A, B we have

$$\mathrm{Tr}(AB) = \mathrm{Tr}(U\Lambda U^\top B) = \mathrm{Tr}(\Lambda U^\top B U) \le \sigma_{\max}(\Lambda) \, \mathrm{Tr}(U^\top B U) = \sigma_{\max}(A) \, \mathrm{Tr}(B),$$

where $U\Lambda U^\top$ is the SVD decomoposition of A. This concludes that

$$\mathrm{Tr}(A_1 A_2) \le \sigma_{\max}(A_3^{-1/2} A_1 A_3^{-1/2}) \, \mathrm{Tr}(A_3 A_2).$$

□

The following lemma is useful to obtain the generalized result of Theorem 1. The proof is given in (Wainwright, 2019, Theorem 3.27). We first define

$$\begin{aligned}
Z &= \sup_{f \in \mathcal{F}} |\{\mathbb{E}_\mathcal{D} - \mathbb{E}_\rho\}[f] \\
\Sigma^2 &= \sup_{f \in \mathcal{F}} \mathbb{E}_\mathcal{D}[\{f(s,a) - \mathbb{E}_\rho[f(s,a)]\}^2], \; \sigma^2 = \sup_{f \in \mathcal{F}} \mathrm{var}[f(s,a)].
\end{aligned}$$

Lemma 12 (Functional Bernstein's inequality: Talagrand concentration inequality for empirical process)**.** *Suppose $\|f\|_\infty \le B$. With probability $1-\delta$,*

$$|Z - \mathbb{E}[Z]| \le \Sigma^2 \sqrt{\frac{\log(c/\delta)}{n}} + \frac{B \log(c/\delta)}{n}.$$

As an immediate corollary,

$$|Z - \mathbb{E}[Z]| \le \{\sigma^2 + B\mathbb{E}[Z]\} \sqrt{\frac{\log(c/\delta)}{n}} + \frac{B \log(c/\delta)}{n}.$$