PESSIMISTIC MODEL-BASED ACTOR-CRITIC FOR OFFLINE REINFORCEMENT LEARNING: THEORY AND ALGORITHMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model-based offline reinforcement learning (RL) has achieved superior performance than model-free RL in many decision-making problems due to its sample efficiency and generalizability. However, prior model-based offline RL methods in the literature either demonstrate their successes only through empirical studies, or provide algorithms that have theoretical guarantees but are hard to implement in practice. To date, a general computationally-tractable algorithm for model-based offline RL with PAC guarantees is still lacking. To fill this gap, we develop a pessimistic model-based actor-critic (PeMACO) algorithm with general function approximations assuming partial coverage of the offline dataset. Specifically, the critic provides a pessimistic Q function through incorporating uncertainties of the learned transition model, and the actor updates policies by employing approximations of the pessimistic Q function. Under some mild assumptions, we establish theoretical PAC guarantees of the proposed PeMACO algorithm by proving an upper bound on the suboptimality of the returned policy by PeMACO.

1 INTRODUCTION

Reinforcement Learning (RL) has emerged as an effective approach for optimizing sequential decision making by maximizing the expected cumulative rewards to learn the optimal policy. RL algorithms have made significant advances in a wide range of areas such as autonomous driving (Shalev-Shwartz et al., 2016), video games (Torrado et al., 2018), and robotics (Kober et al., 2013). However, applying RL to some real-world problems may require optimizing sequential decisions from pre-collected and static (i.e., offline) datasets because interacting with the environment can be expensive or unethical, such as assigning patients to inferior or toxic treatments in healthcare applications (Gottesman et al., 2019). Therefore, developing offline RL methods has grown rapidly in recent decades to learn the optimal policy from offline datasets without further interactions with the environment (Wu et al., 2019; Kumar et al., 2020; Kidambi et al., 2020; Yu et al., 2020; Levine et al., 2020).

The performance of offline RL methods often rely on the coverage of offline data. Earlier theoretical studies of offline RL usually assume that offline data has full coverage, i.e., every possible policy's state distribution can be covered by the distribution of the behavior policy that generates offline data (Munos & Szepesvári, 2008; Ross & Bagnell, 2012; Uehara et al., 2020; Xie & Jiang, 2021). To relax this restrictive assumption, a number of model-free offline RL methods have been developed recently to consider partial coverage of offline data by incorporating pessimism (Liu et al., 2020; Xie et al., 2021; Zanette et al., 2021). However, most existing model-free methods require the Bellman completeness assumption, which is particularly strong in practice due to the lack of monotonic properties for Bellman completeness.

In contrast, model-based methods for offline RL have attracted increasing attentions because of fewer assumptions in theories and better sample efficiencies in practice. Yu et al. (2020) and Kidambi et al. (2020) proposed model-based offline RL methods by modifying the Markov decision process (MDP) model learned from offline data and introducing pessimism in terms of uncertainties of the transition model. Despite their empirical successes, the uncertainties presented in their work were not analytically quantified in an exact manner. For instance, the penalty term in Yu et al.

(2020) is an upper bound of the point-wise estimation error for the transition model, which was not theoretically studied with finite-sample analysis. Recently, Uehara & Sun (2021) developed a pessimistic model-based offline algorithm with general function approximation and proved an upper bound for the suboptimality gap under partial coverage with PAC (probably approximately correct) guarantees. However, their algorithm cannot be easily implemented in practice. Inspired by Uehara & Sun (2021), Rigter et al. (2022) designed a computationally-tractable algorithm by reformulating a max-min constrained optimization problem as a two-player zero sum game against an adversarial environment model. However, theoretical properties of their algorithm have not been studied. A challenge remains open: *can we design a model-based offline RL algorithm that not only can be implemented in practice but also has PAC guarantees?*

In this work, we fill this gap by proposing PeMACO, a pessimistic model-based actor-critic (AC) algorithm for offline RL. Studying model-based offline RL in the AC framework provides us a convenient way to separately investigate the statistical complexity from the critic and computational complexity from the actor by separating the policy optimization from the policy evaluation. Specifically, in the critic, we find a pessimistic Q function in each iteration t by minimizing the Q function over a constraint set of the transition model P. The critic returns a model P_t such that $P_t(\cdot \mid s, a)$ is close to $P^*(\cdot \mid s, a)$ when the state-action pair (s, a) lies in the support of the offline distribution, where P^* is the ground true transition model. Therefore the estimation error would not increase a lot when (s, a) lies in the support of the occupancy measure induced by some policy $\pi^\dagger$ covered by offline data. Such a design is able to return an accurate estimation of the value function induced by a comparator policy covered by offline data, but a pessimistic value function when the comparator policy is not covered by offline data. In the actor part, we approximate the Q function by a linear span of features in a finite-dimensional space and use the natural policy gradient (Agarwal et al., 2021) to update policy parameters. We study the distribution shift arising from the policy gradient step because of the changing occupancy measure (induced by the transition model P_t and the policy π_t) in each iteration. By introducing a finite concentrability coefficient, we show that the transfer error coming from the linear approximation of Q in the actor can also be controlled.

To summarize, we design a computationally-tractable algorithm under general function approximation for model-based offline RL with theoretical guarantees. Our main contributions are threefold. First, compared to the state-of-the-art theoretical work of Uehara & Sun (2021) for model-based offline RL, the proposed PeMACO algorithm not only has PAC guarantees, but also can be implemented in practice. Second, compared to model-free RL literature, we do not require the assumption of Bellman completeness. Third, the theoretical analysis in this work provides a fundamental framework and opens a door for future development of offline model-based algorithms based on AC. Alternative ways of handling transition models from the statistical perspective and approximating the Q function can be adapted in our theoretical framework.

2 RELATED WORK

Model-free Offline RL: Model-free offline RL algorithms usually learn a near-optimal policy from offline datasets by either constraining the policy space to a neighborhood of the behavior policy (Fujimoto et al., 2019; Wu et al., 2019; Liu et al., 2019; Nachum et al., 2019; Kostrikov et al., 2021), or incorporating uncertainties as a notion of pessimism added to the value function during the training process (Kumar et al., 2020; Xie et al., 2021; Kostrikov et al., 2021; Cheng et al., 2022). Compared to constraining the policy space, the pessimistic methods allow the policy to explore actions outside the constraint set. Theoretically, earlier model-free offline RL methods often require realizability and global coverage (Chen & Jiang, 2019; Duan et al., 2021). However, the assumption of global coverage is too strong in the offline RL setting and may not hold in practice. Motivated by the pessimism idea, some recent work proposed model-free RL methods considering the assumption of partial coverage in tabular or linear MDPs (Jin et al., 2021; Rashidinejad et al., 2021; Zhang et al., 2022). In this work, we consider a pessimistic model-based approach in the offline setting under the assumption of partial coverage, in which general MDPs can be studied with PAC guarantees.

Model-based Offline RL: Model-based methods have been explored relatively sparsely in offline RL. Ross & Bagnell (2012) proposed to learn the dynamics from offline data followed by planning, and demonstrated that it could lead to arbitrarily large sub-optimality. Several model-based online RL methods have been explored in the offline setting by limiting model exploitation (Deisenroth &

Rasmussen, 2011; Chua et al., 2018). In recent years, model-based offline RL methods proposed to incorporate pessimism into the value function by quantifying uncertainties of the learned dynamic model. Yu et al. (2020) and Kidambi et al. (2020) introduced pessimism by modifying the MDP model learned from offline data, provided analytic bounds, and demonstrated empirical successes compared to state-of-the-art model-free offline RL algorithms. Uehara & Sun (2021) developed a model-based offline algorithm called constrained pessimistic policy optimization and established its theoretical PAC guarantee under the assumption of partial coverage for general MDPs. However, their algorithm is not computationally practicable. Rigter et al. (2022) designed a practical model-based algorithm inspired by the max-min optimization framework from Uehara & Sun (2021), but its theoretical guarantee has not been investigated. In our work, we design a computationally-tractable algorithm with a theoretical guarantee for model-based offline RL building upon the natural actor-critic algorithm.

Actor-critic algorithms: Actor-critic (AC) and its variant natural AC have gained great popularity in both online and offline RL (Konda & Tsitsiklis, 1999; Peters & Schaal, 2008). The theoretical guarantee for AC algorithms in the online setting has been investigated (Yang et al., 2018; Agarwal et al., 2021). Several recent work consider the AC algorithm in offline RL and argue that it is simpler to study its theoretical properties by separating the policy optimization from the policy evaluation (Zanette et al., 2021; Chen et al., 2022). In particular, Zanette et al. (2021) proposed a model-free AC algorithm for offline RL by considering linear action-value functions under the assumption of Bellman linear completeness. However, all existing work only study the theoretical guarantee of AC algorithms in the model-free setting. In contrast, we propose the first model-based offline AC algorithm that not only assumes partial coverage but also allows for considering more general MDPs beyond simple tabular and linear MDPs.

3 PRELIMINARIES

Markov Decision Processes and Offline RL: We consider an infinite-horizon Markov Decision Process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0)$, with state space $\mathcal{S}$, action space $\mathcal{A}$, a transition dynamics $P(s' \mid s, a)$ with $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, a discount factor $\gamma \in [0, 1)$, and an initial state distribution μ_0. The reward function r is assumed to be known throughout this work. A policy π maps from state space to a distribution over actions, representing a decision strategy to pick an action with probability $\pi(\cdot|s)$ given the current state s. Given a policy π and a transition dynamics P, the value function $V_P^\pi(s) := \mathbb{E}_{P,\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s]$ denotes the expected cumulative discounted reward of π under the transition dynamics P with an initial state s and a reward function r. We use $V_P^\pi := \mathbb{E}_{s\sim\mu_0} V_P^\pi(s)$ to denote the expected value integrating over $\mathcal{S}$ with an initial distribution μ_0. The action-value function (i.e., Q function) is defined similarly: $Q_P^\pi(s, a) = \mathbb{E}_{P,\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s, a_0 = a]$. Let $d_P^\pi(s, a) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}(s_t = s, a_t = a|s_0 \sim \mu_0)$ be the occupancy measure of the policy π under the dynamics P. Then V_P^π can also be expressed as $\mathbb{E}_{(s,a)\sim d_P^\pi}[r(s, a)]$. Assuming that a static offline dataset is generated by some behavior policy under the ground true transition dynamics P^*, model-based offline RL methods aim to learn an optimal policy that maximizes the value $V_{P^*}^\pi$ through the learning of dynamics from the offline dataset without any further interactions with the environment.

Partial Coverage: One fundamental challenge in offline RL is distribution shift (Levine et al., 2020): the visitation distribution of states (and actions) induced by the learned policy inevitably deviates from the distribution of offline data. The concept of coverage has been introduced to measure the distribution shift using the density ratio (Chen & Jiang, 2019). Denote $\rho(s, a)$ to be the offline distribution that generates the state-action pairs $(s_i, a_i)_{i=1}^n$ in the offline dataset. Full coverage means $\sup_{(s,a)} d_{P^*}^\pi(s, a)/\rho(s, a) < \infty$ for all possible policies π, which may not hold in practice. In contrast, partial coverage only assumes that the offline distribution covers the visitation distribution induced by some comparator policy $\pi^\dagger$ (Xie et al., 2021; Uehara & Sun, 2021), such that $\sup_{s,a} d_{P^*}^{\pi^\dagger}(s, a)/\rho(s, a) < \infty$. The goal of our work is to learn the optimal policy among all polices covered by the offline dataset, i.e., $\Pi_C := \{\pi \in \Pi : \sup_{s,a} d_{P^*}^\pi(s, a)/\rho(s, a) < C\}$, where C is some large constant.

4 PEMACO: PESSIMISTIC MODEL-BASED ACTOR-CRITIC FOR OFFLINE RL

We first give a brief introduction to constrained pessimistic policy optimization (CPPO) (Uehara & Sun, 2021), which motivates our work. CPPO is a model-based offline RL algorithm that induces pessimism via searching for the least favorable transition model in a constraint set under partial coverage assumption. Suppose that the maximum likelihood estimate (MLE) of the transition model can be computed from the offline data $\mathcal{D} := \{(s_i, a_i, r_i, s_i')_{i=1}^n\}$ in a model class $\mathcal{M}$. Let $\mathbb{E}_{\mathcal{D}}$ denote the empirical distribution of the offline data $\{(s_i, a_i)_{i=1}^n\}$, i.e., $\mathbb{E}_{\mathcal{D}} f(s,a) = \frac{1}{n}\sum_{i=1}^n f(s_i, a_i)$, and $\mathrm{TV}(\cdot,\cdot)$ denote the total variation distance. CPPO solves a constrained max-min optimization problem:

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

Under the assumptions of some entropy control for model class $\mathcal{M}$, realizability of the true transition dynamics P^* (i.e., $P^* \in \mathcal{M}$), and partial coverage of any comparator policy $\pi^\dagger$, they show an upper bound on the suboptimality of the policy returned by CPPO π^{CPPO}:

$$V_{P^*}^{\pi^\dagger} - V_{P^*}^{\pi^{\mathrm{CPPO}}} \leq c\sqrt{\frac{C_{\pi^\dagger}}{n}}$$

for some constant c with high probability.

Although CPPO considers a general class of MDPs and does not require the assumption of Bellman completeness, it is difficult to design a practical implementation of CPPO due to the computational complexities in solving the max-min optimization problem in (1). This motivates the development of PeMACO, which is a computationally-tractable algorithm for model-based offline RL under partial coverage with theoretical guarantees.

4.1 OVERVIEW OF PEMACO

Our goal is to develop a model-based offline RL algorithm that not only has theoretical PCA guarantees but also can be implemented in practice. To achieve this goal, we design PeMACO, a pessimistic model-based natural actor-critic (AC) algorithm for offline RL. Under the general AC framework, iteratively the actor performs policy improvement which typically does gradient-acent, and the critic performs policy evaluation which estimates the Q function given the current policy, until the optimal policy is achieved (Konda & Tsitsiklis, 1999). Natural AC is a variant of AC when a pre-conditioner is considered in each update of the actor (Peters & Schaal, 2008). The reason why we consider a natural AC framework is that by separating the policy evaluation from the policy optimization, we can analyze statistical and computational properties of the algorithm separately. In particular, we study the statistical guarantee by analyzing the distance between the true and estimated transition models in the critic, and study the theoretical guarantee of computation in the actor.

In model-based AC algorithms, considering the right model class and policy class is of great importance. Intuitively, the model class for transition models should not be too large so that P^* can be learned efficiently. We consider any general MDPs with a controllable entropy in PeMACO. Many commonly-used MDPs can be adapted to our setting, including tabular MDPs, linear mixture MDPs, kernelized nonlinear regulators, low rank MDPs, and factored MDPs (Uehara et al., 2020). For the policy class, since we study the function approximation setting where $|S|$ can be infinite, we consider the log-linear parametric class:

$$\Pi_\theta = \left\{ \pi_\theta(a \mid s) = \frac{\exp\left(\theta \cdot \phi_{s,a}\right)}{\sum_{a'\in\mathcal{A}} \exp\left(\theta \cdot \phi_{s,a'}\right)} \mid \theta \in \mathbb{R}^d \right\}, \quad (2)$$

where $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a feature map with $\|\phi\| \leq B$ for some constant B. We note that more general parametric policy classes can be adapted here with additional assumptions.

Algorithm 1 summarizes PeMACO, which incorporates the pessimism principle into the AC framework. Specifically, we compute the MLE $\hat{P}_{MLE}$ of the transition model from the offline data, and construct a constraint set $\mathcal{M}_{\mathcal{D}}$, in which the transition model $P(\cdot \mid s,a)$ is forced to be close to $\hat{P}_{MLE}(s,a)$ when (s,a) approximately follows the offline distribution ρ. Then in each iteration t,

given the current policy π_t, the critic finds $P_t \in \mathcal{M}_\mathcal{D}$ such that P_t minimizes $V_P^{\pi_t}$. This procedure introduces pessimism in the sense that $V_{P_t}^{\pi_t} \leq V_{P_*}^{\pi_t}$ with high probability, because $P^* \in \mathcal{M}_\mathcal{D}$ with high probability. $Q_{P_t}^{\pi_t}$ can be obtained from Monte Carlo methods given π_t and P_t. Then the actor updates π_t through natural policy gradient (Agarwal et al., 2021), which involves an ascent direction of exponentiated Q function. Due to the fact that the Q function is general with continuous state space in our setting, we employ a linear approximation for Q, allowing us to update the policy in a finite-dimensional space. We will discuss more details about the critic and actor of PeMACO in Section 4.2 and prove an upper bound of the suboptimality for the policy returned by PeMACO in Section 5.

Algorithm 1: Pessimistic Model-Based Actor-Critic for Offline RL (PeMACO)

Initialize $\pi_0(\cdot|s) = \mathrm{Unif}(\mathcal{A})$.
for $t = 0, ..., T-1$ **do**
  Critic (policy evaluation): $Q_t(s,a) = Q_{P_t}^{\pi_t}(s,a)$, where $P_t = \mathrm{argmin}_{P \in \mathcal{M}_\mathcal{D}} V_P^{\pi_t}$ with
  $\mathcal{M}_\mathcal{D} = \left\{ P \mid P \in \mathcal{M}, \mathbb{E}_\mathcal{D} \left[\mathrm{TV}\left(\widehat{P}_{MLE}(\cdot \mid s,a), P(\cdot \mid s,a) \right)^2 \right] \leq \xi \right\}.$
  Actor (policy improvement): $\theta_{t+1} \leftarrow \theta_t + \eta w_t$, where
  $w_t \in \mathrm{argmin}_{\|w\| \leq W} \mathbb{E}_{s \sim d_{P_t}^{\pi_t}, a \sim \mathrm{Unif}(\mathcal{A})} \left[\left(Q_t(s,a) - w \cdot \phi_{s,a} \right)^2 \right].$
end

4.2 More details on PeMACO

4.2.1 The critic: constrained pessimistic policy evaluation

The critic part of PeMACO is designed to provide a pessimistic estimate of the value function given a policy from the actor. This idea was first explored by Zanette et al. (2021), who developed a pessimistic AC algorithm with the critic lower bounding the Q function. Specifically, they assume that the Q function is a linear combination of features under Bellman restricted closedness and find the minimum Q in a constraint set of the linear coefficient. However, their linear assumption of the Q function and assumption of Bellman restricted closedness are rather strong. In this work, we consider more general MDPs with potentially nonlinear Q functions. Instead of lower bounding the Q function directly, we borrow the idea from Uehara & Sun (2021) to minimize the value function given an appropriate constraint set:

$$\min_{P \in \mathcal{M}_\mathcal{D}} V_P^{\pi_t}, \mathcal{M}_\mathcal{D} = \left\{ P \mid P \in \mathcal{M}, \mathbb{E}_\mathcal{D} \left[\mathrm{TV}\left(\widehat{P}_{MLE}(\cdot \mid s,a), P(\cdot \mid s,a) \right)^2 \right] \leq \xi \right\}. \tag{3}$$

The purpose of introducing the radius of the constraint set $\mathcal{M}_\mathcal{D}$ is to compensate potentially high statistical errors due to insufficient coverage of the offline data.

In tabular MDPs, (3) is a constrained convex optimization problem, which is easy to solve. With more general settings for the transition model P, the value function $V_P^{\pi_t}$ could be nonlinear, which makes computation challenging. To solve (3), we will employ the projected gradient descent (Chong & Zak, 2004), more details of which will be discussed in Section 6. For neatness of the main theorem, we assume that we have access to the oracle of the constrained minimization problem (3).

Assumption 1 (Constrained minimization oracle)**.** *Given any $\pi \in \Pi_\theta$, we can find a $P(\pi)$ such that*

$$V_{P(\pi)}^{\pi} = \min_{P \in \mathcal{M}_\mathcal{D}} V_P^{\pi}. \tag{4}$$

Assume P_t satisfies (4) given π_t, i.e., $P_t = P(\pi_t)$, then the critic can output $Q_{P_t}^{\pi_t}$ given P_t and π_t. Here $Q_{P_t}^{\pi_t}$ can be computed via the Monte Carlo method, i.e., generating a number of trajectories corresponding to P_t and π_t and then estimating the Q function given the desired accuracy.

4.2.2 The actor: Natural policy gradient (NPG)

For policy improvement in the actor, we use the natural policy gradient (NPG) algorithm (Agarwal et al., 2021). Under the tabular setting where $|\mathcal{S}||\mathcal{A}|$ is finite and P^* is known, the update rule of

NPG is $\pi_{t+1}(a \mid s) \propto \pi_t(a \mid s) e^{Q^{\pi_t}_{P^*}(s,a)}$. However, when $|S|$ is infinite, we cannot update $\pi(a|s)$ for each $s \in \mathcal{S}$. To tackle this issue, one can introduce a class of parameterized policies using a finite-dimensional feature space and project the Q function to the same feature space, so that the policy parameters can be updated in a finite-dimensional space. For example, when the underlying transition model P^* is known, Agarwal et al. (2021) update policy parameters as follows:

$$\text{NPG:} \quad \theta_{t+1} \leftarrow \theta_t + \eta w_t, \quad w_t \in \operatorname{argmin}_w \mathbb{E}_{s \sim d^{\pi_t}_{P^*}, a \sim \pi_t(\cdot|s)} \left[\left(Q^{\pi_t}_{P^*}(s,a) - w \cdot \phi_{s,a} \right)^2 \right],$$

where $\pi_t := \pi_{\theta_t}$.

Essentially, with the help of linear approximations of the Q function, one can avoid updating $\pi(a|s)$ for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, which has an infinite cardinality. However, we do not have access to P^* in the offline setting. This motivates us to update the parameter θ using the dynamic model P_t obtained from the critic in the policy improvement step of PeMACO as follows:

$$\theta_{t+1} \leftarrow \theta_t + \eta w_t, \quad w_t \in \operatorname{argmin}_w \mathbb{E}_{s \sim d^{\pi_t}_{P_t}, a \sim \mathrm{Unif}(\mathcal{A})} \left[\left(Q^{\pi_t}_{P_t}(s,a) - w \cdot \phi_{s,a} \right)^2 \right].$$

5 Theoretical Guarantees for PeMACO

In this section, we theoretically analyze PeMACO by establishing the suboptimality of the learned policy π_t with any comparator policy $\pi^\dagger$ covered by the offline data: $V^{\pi^\dagger}_{P^*} - V^{\pi_t}_{P^*}$. We first introduce several assumptions needed for our theoretical study.

The first assumption is related to offline data generation, which is common in theoretical literature of offline RL.

Assumption 2 (Data generation). *The dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i') : i = 1, ..., n\}$ satisfies $(s_i, a_i) \sim \rho$ i.i.d. with $s_i' \sim P^*(\cdot \mid s_i, a_i)$, where ρ denotes the offline distribution induced by the behavior policy under P^*.*

In order to quantify the partial coverage of any comparator policy $\pi^\dagger$, we define the concentrability coefficient following Uehara & Sun (2021), who showed that a finite concentration coefficient is needed to control the distribution shift between the offline distribution and the occupancy measure induced by $\pi^\dagger$.

Definition 1 (Concentrability coefficient).

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

Assumption 3 (Coverage of any comparator policy $\pi^\dagger$).

$$C_{\pi^\dagger} < \infty.$$

To ensure that the Q function can be linearly approximated, we make the below assumption of approximation error so that the "distance" between the function class of Q and the linear function class can be controlled given the model class $\mathcal{M}$ and the policy class Π_θ.

Assumption 4 (Approximation error).

$$\sup_{P \in \mathcal{M}, \pi \in \Pi_\theta} \inf_{w : \|w\|_2 \le W} \|Q^\pi_P(s,a) - w \cdot \phi_{s,a}\|_{2, d^\pi_P \circ \mathrm{Unif}} \le \varepsilon_{\mathrm{approx}},$$

where $\|f(s,a)\|_{2, d^\pi_P \circ \mathrm{Unif}}$ denotes $(\mathbb{E}_{s \sim d^\pi_P, a \sim \mathrm{Unif}(\mathcal{A})}[f(s,a)^2])^{\frac{1}{2}}$, and $\phi_{s,a}$ comes from the definition of Π_θ in (2).

Finally, we define an estimation error that is related to the offline distribution as well as the usage of offline data for estimating P.

Definition 2 (Estimation error).

$$\varepsilon_{est} := \sup_{P \in \mathcal{M}_\mathcal{D}} (\mathbb{E}_{(s,a) \sim \rho} [\mathrm{TV}(P(\cdot \mid s,a), P^*(\cdot \mid s,a))^2])^{\frac{1}{2}}.$$

To find a policy for which we can provide a non-asymptotic bound on its suboptimality $V_{P^*}^{\pi^\dagger} - V_{P^*}^{\pi^t}$, we decompose the suboptimality into three parts and analyze their bounds separately:

$$V_{P^*}^{\pi^\dagger} - V_{P^*}^{\pi^t} = \underbrace{(V_{P^*}^{\pi^\dagger} - V_{P_t}^{\pi^\dagger})}_{\text{(a)}} + \underbrace{(V_{P_t}^{\pi^\dagger} - V_{P_t}^{\pi_t})}_{\text{(b)}} + \underbrace{(V_{P_t}^{\pi_t} - V_{P^*}^{\pi_t})}_{\text{(c)}}.$$

Previous work has shown that part (a) can be bounded by a term of $O(\frac{\sqrt{C_{\pi^\dagger}}}{\sqrt{n}})$ with the concentrability of a single policy $\pi^\dagger$ (Uehara & Sun, 2021). We will bound part (b) after expressing it as a difference in KL-divergence under P_t using the performance difference lemma (Kakade & Langford, 2002). For part (c), since $P^* \in \mathcal{M}_\mathcal{D}$ with high probability, we have $V_{P_t}^{\pi_t} \leq V_{P^*}^{\pi_t}$ with high probability by the definition of P_t. Pessimism is encoded here because π_t may not be covered by the offline data, i.e., there is no control on C_{π_t}.

Now we are ready to present our main result: the theoretical guarantee of PeMACO. Under assumption 1 for oracle of solving (3), assumption 2 for offline data generation, assumption 3 for sufficient coverage of any comparator policy $\pi^\dagger$ covered by the offline dataset, assumption 4 for good approximation of Q by a linear combination of features, the following theorem provides an upper bound for the suboptimality of the best policy among the iterated policies from Algorithm 1. All proofs are given in the Appendix.

Theorem 1. *Let $\eta = \sqrt{\frac{2\log|\mathcal{A}|}{B^2W^2(T+1)}}$. Let ξ be some constant such that $P^* \in \mathcal{M}_\mathcal{D}$ with probability at least $1-\delta$. Assume the feature mapping is bounded, i.e., $\|\phi_{s,a}\|_2 \leq B$. Then under the assumptions 1-4, there exist some constant c_1 such that*

$$V_{P^*}^{\pi^\dagger} - \max_{0\leq t\leq T} V_{P^*}^{\pi_t} \leq \underbrace{c_1\sqrt{C_{\pi^\dagger}}\,\varepsilon_{est}}_{\text{statistical error}} + \underbrace{\frac{BW}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T+1}}}_{\text{optimization error}} + \underbrace{\frac{2|\mathcal{A}|}{(1-\gamma)^2}\sqrt{\sup_s \frac{d_{P^*}^{\pi^\dagger}(s)}{\mu_0(s)}}\,\varepsilon_{\text{approx}}}_{\text{transfer error}}$$

with probability at least $1-\delta$.

The upper bound includes three error terms: statistical error, optimization error, and transfer error. The **statistical error** comes from two sources: estimating the transition model and the offline distribution from offline data. With finite concentrability $C_{\pi^\dagger}$, the error can be controlled when learning the value of the comparator policy $\pi^\dagger$. We stress that the statistical error comes from both the critic and the actor. In the critic, the estimation error ε_{est} is utilized to introduce the pessimism to the value function induced by the policy that is not covered by the offline data. To explain the involvement of statistical error in the actor, recall that each actor iteration of PeMACO employs a different transition model P_t within the constraint set. We prove that when ε_{est} is small, the actor can approximately improve the policy in each iteration of PeMACO.

The **optimization error** comes from the natural gradient descent in the actor and is bounded by a term of $O(\sqrt{\frac{\log|\mathcal{A}|}{T}})$, which quantifies the decreasing error rate in terms of the number of iterations in the actor. The **transfer error** also comes from the analysis of the actor, and is an unavoidable constant that cannot be reduced by increasing sample size n or the number of iterations T. When $\varepsilon_{\text{approx}}$ is an upper bound for $\sup_{P\in\mathcal{M},\pi\in\Pi_\theta}\inf_{w:\|w\|_2\leq W}\|Q_P^\pi(s,a) - w\cdot\phi_{s,a}\|_{2,d_P^\pi\circ\text{Unif}}$, it implies that $\inf_{w:\|w\|_2\leq W}\|Q_{P_t}^{\pi_t}(s,a) - w\cdot\phi_{s,a}\|_{2,d_{P_t}^{\pi^t}\circ\text{Unif}} \leq \varepsilon_{\text{approx}}$ for each t. In other words, if we can find a perfect $w_t = \inf_{w:\|w\|_2\leq W}\|Q_{P_t}^{\pi_t}(s,a) - w\cdot\phi_{s,a}\|_{2,d_{P_t}^{\pi^t}\circ\text{Unif}}$ in the actor, then $\varepsilon_{\text{approx}}$ upper bounds the error of the function approximation in each iteration. If the Q function can be exactly expressed as a linear combination of the given features, then $\varepsilon_{\text{approx}} = 0$. One appealing result in our analysis is that we quantify the distribution shift $\sup_s \frac{d_{P^*}^{\pi^\dagger}(s)}{\mu_0(s)}$ between the initial distribution and the occupancy measure induced by a single $\pi^\dagger$ under P^*. That being said, the initial distribution μ_0 should have sufficient coverage over the whole state space in order to have a smaller $\sup_s \frac{d_{P^*}^{\pi^\dagger}(s)}{\mu_0(s)}$.

Remark. There are two types of distribution shift here. If we view $d_{P_*}^{\pi^\dagger}$ as a "target" distribution, then $C_{\pi^\dagger}$ captures the discrepancy between the offline distribution and the target distribution, while $\sup_s \frac{d_{P^*}^{\pi^\dagger}(s)}{\mu_0(s)}$ captures the discrepancy between the initial distribution and the target distribution. Eventually, $C_{\pi^\dagger}$ controls the transfer of the estimation error from the offline data to the target distribution, while $\sup_s \frac{d_{P^*}^{\pi^\dagger}(s)}{\mu_0(s)}$ controls the transfer of $\varepsilon_{\text{approx}}$ from the linear approximation of Q under $d_{P_t}^{\pi^t}$ in each iteration t.

Theorem 1 provides a general framework for analyzing the suboptimality for offline model-based actor-critic algorithms. Based on the proposed framework, one can study ε_{est} independently under the traditional supervised learning setting. We take the kernelized nonlinear regulator (KNR) as an example, building upon the analysis of ε_{est} in the section 5.2 of Uehara et al. (2020). A KNR (Kakade et al., 2020) assumes that the underlying true transition model $P^\star(s' \mid s, a)$ is defined by $s' = W^\star \varphi(s,a) + \epsilon$ with $\epsilon \sim \mathcal{N}\left(0, \zeta^2 \mathbf{I}\right)$, where $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a possibly nonlinear feature mapping. Suppose that the underlying true MDP is a KNR, we have the following result.

Corollary 1. *Assume* $\|\varphi(s,a)\|_2 \leq 1, \|\phi_{s,a}\|_2 \leq B, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. *Let* $\eta = \sqrt{\frac{2\log|\mathcal{A}|}{B^2W^2(T+1)}}$, $\xi = \sqrt{2\lambda \|W^\star\|_2^2 + 8\zeta^2\left(d_{\mathcal{S}} \ln(5) + \ln(1/\delta) + \overline{\mathcal{I}}_n\right)}$, *where* $\overline{\mathcal{I}}_n = \ln\left(\det\left(\Sigma_n\right) / \det(\lambda \mathbf{I})\right)$. *Then, under assumptions 1-4, by letting* $\|W^\star\|_2^2 = O(1), \zeta^2 = O(1), \lambda = O(1)$, *with probability* $1-\delta$, *there exist some constants* c_1, c_2 *such that*

$$V_{P^*}^{\pi^\dagger} - \max_{0\leq t\leq T} V_{P^*}^{\pi_t} \leq c_1(1-\gamma)^{-2} \min\left(d^{1/2}, \bar{R}\right) \sqrt{\bar{R}} \sqrt{\frac{d_{\mathcal{S}} C_{\pi^\dagger} \ln(1+n)}{n}} + \frac{BW}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T+1}} + \frac{2}{(1-\gamma)^2}|\mathcal{A}| \sqrt{\sup_s \frac{d_{P^*}^{\pi^\dagger}(s)}{\mu_0(s)}} \varepsilon_{\text{approx}},$$

where $\Sigma_\rho = \mathbb{E}_\rho[\varphi\varphi^T]$, $\bar{R} = \operatorname{rank}\left[\Sigma_\rho\right]\left\{\operatorname{rank}\left[\Sigma_\rho\right] + \ln\left(c_2/\delta\right)\right\}$, *and* $d_{\mathcal{S}}$ *denotes the dimension of* $\mathcal{S}$.

6 A Practical Implementation of PeMACO

The proposed PeMACO algorithm can be applied to many general MDPs with PAC guarantees. In this section, we describe a practical implementation of PeMACO. In the proposed AC framework, solving the constrained optimization problem (3) in the critic is the most computationally challenging part. We propose to use projected gradient descent (PGD) (Jain et al., 2017), and show that under mild assumptions on $V_{P_t}^{\pi_t}$, PGD can achieve a minimizer over $\mathcal{M}_{\mathcal{D}}$. The key idea of PGD is to iteratively improve the objective value through gradient descent and project the updated parameter into a convex constraint set once the updated parameter falls out of the constraint set. In order to implement PGD, we need to be able to construct a convex constraint set and compute the gradient of the objective function.

To compute the gradient of the objective function, we can directly follow Proposition 2 from Rigter et al. (2022), summarized in Lemma 1.

Lemma 1. *(Proposition 2 of Rigter et al. (2022)) Let* $P_\psi(\cdot|s,a)$ *be the transition model parameterized by* $\psi \in \mathbb{R}^d$, *and let* V_ψ^π *denote* $V_{P_\psi}^\pi$, *then for any given policy* π, *we have*

$$\nabla_\psi V_\psi^\pi = \mathbb{E}_{s\sim d_\psi^\pi, a\sim\pi, s'\sim P_\psi}\left[\left(r(s,a) + \gamma V_\psi^\pi(s')\right) \cdot \nabla_\psi \log P_\psi(s' \mid s,a)\right].$$

Combining lemma 1 and the convexity of $\mathcal{M}_{\mathcal{D}}$, we can easily show that PGD can find P_t given some conditions on the value function, following Jain et al. (2017).

Proposition 1. *Suppose that* $V_\psi^\pi : \mathcal{M}_{\mathcal{D}} \to \mathbb{R}$ *is convex and 1-Lipschitz. Assume* $\mathcal{M}_{\mathcal{D}}$ *is convex. Let* $\psi_1, \psi_2, ..., \psi_T$ *be the iterative output according to* $\psi_{t+1} = \operatorname{Proj}_{\mathcal{M}_{\mathcal{D}}}(\psi_t - \eta \nabla_\psi V_{\psi_t}^\pi)$. *Define* $\eta = \frac{1}{\sqrt{T}}$. *Let* $\hat{\psi} := \frac{1}{T}\sum_{i=1}^T \psi_i$, *then*

$$V_{\hat{\psi}}^\pi - \min_{P_\psi \in \mathcal{M}_{\mathcal{D}}} V_\psi^\pi \leq \frac{1}{\sqrt{T}}.$$

The constraint set $\mathcal{M}_{\mathcal{D}}$ constructed from many commonly-used MDPs can be shown to be convex, such as tabular MDPs and KNRs. Recall that a KNR has the transition model defined by $s' \sim N(W\varphi(s,a), \zeta^2 I)$ where $\varphi(s,a) \in \mathbb{R}^d$ is a feature map. Then the transition model is parameterized by $W \in \mathbb{R}^d$. Following Devroye et al. (2018), $\mathrm{TV}\left(P_{W_1}(\cdot \mid s,a), P_{W_2}(\cdot \mid s,a)\right)^2 = \Theta\left(\left\|(W_1 - W_2)^T \varphi(s,a)\right\|_2^2\right)$. This implies

$$\mathcal{M}_{\mathcal{D}} = \{W \in \mathbb{R}^d | (W - \widehat{W}_{MLE})^T \Sigma_{\mathcal{D},\varphi} (\phi - \widehat{W}_{MLE}) \leq \xi\},$$

where $\Sigma_{\mathcal{D},\varphi} := \frac{1}{n}\sum_{(s,a)\in\mathcal{D}} \varphi(s,a)\varphi(s,a)^T$. Note that $(W - \widehat{W}_{MLE})^T \Sigma_{\mathcal{D},\varphi} (W - \widehat{W}_{MLE})$ can also be written as $\|W - \widehat{W}_{MLE}\|_{\Sigma_{\mathcal{D},\varphi}}$, and $\mathcal{M}_{\mathcal{D}}$, as a $\sqrt{\xi} - \|\cdot\|_{\Sigma_{\mathcal{D},\varphi}}$ ball, is convex with respect to W.

Remark: The convexity of the constraint set in PGD is important in order to achieve the stationary point of the objective function (Jain et al., 2017). When the constraint set is not convex, we can consider generalized projected gradient descent (Jain et al., 2017) for solving (3) . Also, the convexity assumption of $V^{\pi}_{P_W}$ sometimes can be strong in practice, preventing PGD to achieve the global optimum within the constraint set. When the objective function is not convex, PGD can still find the stationary points within the constraint set (Dunn, 1987). In practice, we recommend to run PGD multiple times with different initialization values when the objective function is not convex.

With the above discussion, we propose a practical implementation of PeMACO in Algorithm 2.

Algorithm 2: A Practical Implementation of PeMACO

Initialize $\pi_0(\cdot|s) = \mathrm{Unif}(\mathcal{A})$.
for $t = 0, ..., T-1$ **do**
  Critic (policy evaluation):
  for $k = 0, ..., K-1$ **do**
    $\phi_{k+1} = \mathrm{Proj}_{\mathcal{M}_{\mathcal{D}}}(\phi_k - \eta \nabla_{\phi} V^{\pi_t}_{\phi_k})$.
  end
  Let $P_t := P_{\phi^t}$ where $\phi^t = \frac{\sum_{k=1}^{K}\phi_k}{K}$. Let $Q_t := Q^{\pi_t}_{P_t}$.
  Actor (policy improvement): $\theta_{t+1} \leftarrow \theta_t + \eta w_t, \quad where \quad w_t \in \mathrm{argmin}_{\|w\|\leq W} \mathbb{E}_{s\sim d^{\pi_t}_{P_t}, a\sim\mathrm{Unif}(\mathcal{A})}\left[\left(Q_t(s,a) - w\cdot\phi_{s,a}\right)^2\right]$.
end

7 Discussion

We developed PeMACO, a pessimistic model-based AC algorithm for offline RL, with general function approximation under the assumption of partial coverage. PeMACO can be practically implemented with PAC guarantees. By separating the policy optimization from the policy evaluation under the AC framework, we theoretically analyzed PeMACO and proved an upper bound for the suboptimality of the policy returned by PeMACO.

The proposed framework has several possible extensions. First, we employ a linear approximation for the Q function. Alternative ways to approximate Q can be further explored for more flexibility. Second, we construct the constraint set $\mathcal{M}_D$ using the MLE $\hat{P}_{MLE}$ of the transition model computed from offline data. We will consider other estimators other than MLE, such as posterior mean from applying a Bayesian method for estimating the transition model, and study its theoretical properties. Lastly, we consider the log-linear policy class in our work. More general policy classes such as those parameterized by neural networks can be explored and theoretical properties corresponding to neural policy gradient methods will be investigated (Wang et al., 2019).

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning*

Research, 22(98):1–76, 2021.

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

Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022.

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

Edwin KP Chong and Stanislaw H Zak. *An introduction to optimization*. John Wiley & Sons, 2004.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

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, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 2892–2902. PMLR, 2021.

John C Dunn. On the convergence of projected gradient processes to singular critical points. *Journal of Optimization Theory and Applications*, 55(2):203–216, 1987.

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.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

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

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

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

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

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

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

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

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. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 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.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

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

Marc Rigter, Bruno Lacerda, and Nick Hawes. Rambo-rl: Robust adversarial model-based offline reinforcement learning. *arXiv preprint arXiv:2204.12581*, 2022.

Stephane Ross and J Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. *arXiv preprint arXiv:1203.1007*, 2012.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Ruben Rodriguez Torrado, Philip Bontrager, Julian Togelius, Jialin Liu, and Diego Perez-Liebana. Deep reinforcement learning for general video game ai. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pp. 1–8. IEEE, 2018.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. *arXiv e-prints*, pp. arXiv–2107, 2021.

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

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

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. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

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

Zhuoran Yang, Kaiqing Zhang, Mingyi Hong, and Tamer Başar. A finite sample analysis of the actor-critic algorithm. In *2018 IEEE conference on decision and control (CDC)*, pp. 2759–2764. IEEE, 2018.

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

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

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5757–5773. PMLR, 2022.

A Proof of Theorem 1

Proof of Theorem 1. Our goal is to provide an upper bound for the suboptimality of the best policy among the iterated policies from algorithm 1. We first use the fact that maximum is lower bounded by the mean, then split it into the sum of three parts. We will deal with the three parts separately.

$$\begin{aligned} V_{P^\dagger}^{\pi^\dagger} - \max_{0\le t\le T} V_{P^*}^{\pi_t} &\le \frac{1}{T+1}\sum_{t=0}^{T}(V_{P^*}^{\pi^\dagger} - V_{P^*}^{\pi_t}) \\ &= \frac{1}{T+1}\sum_{t=0}^{T}(V_{P^*}^{\pi^\dagger} - V_{P_t}^{\pi^\dagger} + V_{P_t}^{\pi^\dagger} - V_{P_t}^{\pi_t} + V_{P_t}^{\pi_t} - V_{P^*}^{\pi_t}) \\ &= \underbrace{\frac{1}{T+1}\sum_{t=0}^{T}(V_{P^*}^{\pi^\dagger} - V_{P_t}^{\pi^\dagger})}_{(a)} + \underbrace{\frac{1}{T+1}\sum_{t=0}^{T}(V_{P_t}^{\pi^\dagger} - V_{P_t}^{\pi_t})}_{(b)} + \underbrace{\frac{1}{T+1}\sum_{t=0}^{T}(V_{P_t}^{\pi_t} - V_{P^*}^{\pi_t})}_{(c)}. \end{aligned}$$

Bound (a). In algorithm 1 , we have $P_t \in \mathcal{M}_\mathcal{D}$ for all t. To bound (a), we need to provide a uniform upper bound of $V_{P^*}^{\pi^\dagger} - V_P^{\pi^\dagger}$ for all $P \in \mathcal{M}_\mathcal{D}$, which is given by the following Lemma 2.

Lemma 2.

$$V_{P^*}^{\pi^\dagger} - V_P^{\pi^\dagger} \le \gamma(1-\gamma)^{-2}\sqrt{C_{\pi^\dagger}}\,\varepsilon_{est}$$

holds for all $P \in \mathcal{M}_\mathcal{D}$.

Then it immediately follows that

$$(a) = \frac{1}{T+1}\sum_{t=0}^{T}(V_{P^*}^{\pi^\dagger} - V_{P_t}^{\pi^\dagger}) \le \gamma(1-\gamma)^{-2}\sqrt{C_{\pi^\dagger}}\,\varepsilon_{\text{est}}. \tag{5}$$

Bound (c). As we assume that $P^* \in \mathcal{M}_\mathcal{D}$ with high probability in theorem 1. Then by the critic step in algorithm 1, $V_{P_t}^{\pi_t}$ achieves the minimum among all models in $\mathcal{M}_\mathcal{D}$. So $V_{P_t}^{\pi_t} \le V_{P^*}^{\pi_t}$ for all t, which directly implies

$$(c) = \frac{1}{T+1}\sum_{t=0}^{T}(V_{P_t}^{\pi_t} - V_{P^*}^{\pi_t}) \le 0. \tag{6}$$

Bound (b). The term (b) measures the difference between policies π_t and $\pi^\dagger$ under the same transition model P_t, which reminds us of the following useful lemma.

Lemma 3 (Performance difference lemma)**.** *Suppose $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0)$ is fixed. π and π' are two policies, then*

$$V^\pi - V^{\pi'} = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi} A^{\pi'}(s,a),$$

where $V^{\pi} = \mathbb{E}_{s\sim\mu} V^{\pi}(s)$, $V^{\pi'} = \mathbb{E}_{s\sim\mu} V^{\pi'}(s)$.

Now we can proceed to deal with (b). Applying lemma 3 to the transition model P_t, policies $\pi^\dagger$ and π_t, we get

$$V_{P_t}^{\pi^\dagger} - V_{P_t}^{\pi_t} = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}} A_{P_t}^{\pi_t}(s,a).$$

Then we decompose (b) into three terms:

$$\begin{aligned}(b) = \frac{1}{T+1}\sum_{t=0}^{T}(V_{P_t}^{\pi^\dagger} - V_{P_t}^{\pi_t}) &= \frac{1}{T+1}\sum_{t=0}^{T}\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}} A_{P_t}^{\pi_t}(s,a) \\ &= \frac{1}{(T+1)(1-\gamma)}\sum_{t=0}^{T}(a_t + b_t + c_t),\end{aligned}$$

where

$$a_t := \mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}}\left[w_t \cdot \nabla_\theta \log \pi_t(a|s)\right] - \mathbb{E}_{(s,a)\sim d_{P^*}^{\pi^\dagger}}\left[w_t \cdot \nabla_\theta \log \pi_t(a|s)\right]. \tag{7}$$

$$b_t := \mathbb{E}_{(s,a)\sim d_{P^*}^{\pi^\dagger}}\left[w_t \cdot \nabla_\theta \log \pi_t(a|s)\right]. \tag{8}$$

$$c_t := \mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}}\left[A_{P_t}^{\pi_t}(s,a) - w_t \cdot \nabla_\theta \log \pi_t(a|s)\right]. \tag{9}$$

We will present a sequence of lemmas to control a_t, b_t, c_t. For each a_t, it can be upper bounded by the following lemma 4.

Lemma 4. *The following holds for any* $0 \le t \le T$*,*

$$a_t \le \frac{4\gamma BW}{(1-\gamma)^2}\sqrt{C_{\pi^\dagger}}\varepsilon_{est}.$$

Lemma 5. *Under the assumption* $\|\phi_{s,a}\|_2 \le B$*,*

$$\frac{1}{T+1}\sum_{t=0}^{T} b_t \le \frac{\eta B^2 W^2}{2} + \frac{\log|\mathcal{A}|}{\eta(T+1)}.$$

By lemma 5, and recall that $\eta = \sqrt{\frac{2\log|\mathcal{A}|}{B^2W^2(T+1)}}$, we get

$$\frac{1}{T+1}\sum_{t=0}^{T} b_t \le \frac{\eta B^2 W^2}{2} + \frac{\log|\mathcal{A}|}{\eta(T+1)} = BW\sqrt{\frac{2\log|\mathcal{A}|}{T+1}}.$$

The following lemma 6 provides an upper bound for each c_t.

Lemma 6. *Under assumption 4, the following holds for any* $0 \le t \le T$*,*

$$c_t \le \frac{2|\mathcal{A}|}{1-\gamma}\left\|\frac{d_{P^*}^{\pi^\dagger}(s)}{\mu_0(s)}\right\|_\infty^{\frac{1}{2}}\varepsilon_{\text{approx}} + \frac{4\gamma|\mathcal{A}|}{(1-\gamma)^2}\left(WB + \frac{\gamma}{(1-\gamma)}\right)\sqrt{C_{\pi^\dagger}}\,\varepsilon_{est}.$$

Then we obtain an upper bound of (b) by combining the analyses for a_t,b_t,c_t together:

$$
\begin{aligned}
(b) &= \frac{1}{(T+1)(1-\gamma)} \sum_{t=0}^{T} (a_t + b_t + c_t) \\
&\leq \frac{4\gamma BW}{(1-\gamma)^3} \sqrt{C_{\pi^\dagger}}\, \varepsilon_{\text{est}} + \frac{BW}{1-\gamma} \sqrt{\frac{2\log|\mathcal{A}|}{T+1}} \\
&\quad + \frac{2|\mathcal{A}|}{(1-\gamma)^2} \left\| \frac{d^{\pi^\dagger}_{P^*}(s)}{\mu_0(s)} \right\|_\infty^{\frac{1}{2}} \varepsilon_{\text{approx}} + \frac{4\gamma|\mathcal{A}|}{(1-\gamma)^3} \left(WB + \frac{\gamma}{(1-\gamma)} \right) \sqrt{C_{\pi^\dagger}}\, \varepsilon_{est}.
\end{aligned} \tag{10}
$$

Finally, we combine the analyses for (a),(b),(c), i.e., (5)+(6)+(10):

$$
\begin{aligned}
V^{\pi^\dagger}_{P^*} - \max_{0\leq t\leq T} V^{\pi_t}_{P^*} &\leq (c) + (a) + (b) \\
&\leq 0 + \gamma(1-\gamma)^{-2} \sqrt{C_{\pi^\dagger}}\, \varepsilon_{\text{est}} + \frac{4\gamma BW}{(1-\gamma)^3} \sqrt{C_{\pi^\dagger}}\, \varepsilon_{\text{est}} + \frac{BW}{1-\gamma} \sqrt{\frac{2\log|\mathcal{A}|}{T+1}} \\
&\quad + \frac{2|\mathcal{A}|}{(1-\gamma)^2} \left\| \frac{d^{\pi^\dagger}_{P^*}(s)}{\mu_0(s)} \right\|_\infty^{\frac{1}{2}} \varepsilon_{\text{approx}} + \frac{4\gamma|\mathcal{A}|}{(1-\gamma)^3} \left(WB + \frac{\gamma}{(1-\gamma)} \right) \sqrt{C_{\pi^\dagger}}\, \varepsilon_{est} \\
&= C\sqrt{C_{\pi^\dagger}}\, \varepsilon_{\text{est}} + \frac{BW}{1-\gamma} \sqrt{\frac{2\log|\mathcal{A}|}{T+1}} + \frac{2|\mathcal{A}|}{(1-\gamma)^2} \left\| \frac{d^{\pi^\dagger}_{P^*}(s)}{\mu_0(s)} \right\|_\infty^{\frac{1}{2}} \varepsilon_{\text{approx}},
\end{aligned}
$$

which concludes the proof.

B PROOF OF LEMMAS IN SECTION A

In this section, we prove lemmas we used in section A, i.e. lemma 2, 3, 4, 5, 6.

Proof of lemma 2. Note that $0 \leq V^{\pi}_{P} \leq (1-\gamma)^{-1}$, directly applying lemma 7 to P^*, P, $\pi^\dagger$ leads to

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

By definition of concentration coefficient, we have

$$
\mathbb{E}_{(s,a)\sim d^{\pi^\dagger}_{P^*}} \left[\mathrm{TV}(P(\cdot|s,a), P^*(\cdot|s,a))^2 \right] \leq C_{\pi^\dagger} \mathbb{E}_{(s,a)\sim\rho} \left[\mathrm{TV}(P(\cdot|s,a), P^*(\cdot|s,a))^2 \right]. \tag{11}
$$

By definition 2,

$$
\mathbb{E}_{(s,a)\sim\rho} \left[\mathrm{TV}(P(\cdot|s,a), P^*(\cdot|s,a))^2 \right] \leq \varepsilon^2_{\text{est}}.
$$

So we can combine the three inequalities above together:

$$
V^{\pi^\dagger}_{P^*} - V^{\pi^\dagger}_{P} \leq \gamma(1-\gamma)^{-2} \sqrt{C_{\pi^\dagger}\, \varepsilon^2_{\text{est}}},
$$

which concludes the proof. □

Proof of lemma 3. The lemma 3.2 in Agarwal et al. (2021) provides a similar result for the value function evaluated at each state s_0:

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^{\pi}(\cdot,\cdot|s_0)} A^{\pi'}(s,a).$$

Then it naturally extends to our version by taking expectation with respect to the initial distribution:

$$\begin{aligned} V^{\pi} - V^{\pi'} = \mathbb{E}_{s_0\sim\mu}\left[V^{\pi}(s_0) - V^{\pi'}(s_0)\right] &= \frac{1}{1-\gamma}\mathbb{E}_{s_0\sim\mu}\mathbb{E}_{(s,a)\sim d_P^{\pi}(\cdot,\cdot|s_0)} A^{\pi'}(s,a) \\ &= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^{\pi}} A^{\pi'}(s,a). \end{aligned}$$

Recall that under our notations, d_P^{π} is the visitation measure conditioning on the initial distribution μ. □

Proof of lemma 4. Recall the definition of a_t: (see (7))

$$a_t = \mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}}\left[w_t \cdot \nabla_\theta \log \pi_t(a|s)\right] - \mathbb{E}_{(s,a)\sim d_{P^*}^{\pi^\dagger}}\left[w_t \cdot \nabla_\theta \log \pi_t(a|s)\right].$$

We first rewrite the visitation measure above. For clarity, we introduce some notations for the distribution of trajectories. Denote the state-action trajectory as τ, i.e. $\tau = (s_1, a_1, s_2, a_2, \dots)$. Let $P_*^{\pi^\dagger}(\cdot|s_0 \sim \mu)$ denote the distribution of trajectory generated by transition model P^* and policy $\pi^\dagger$, conditioning on an initial state $s_0 \sim \mu$. Let $P_t^{\pi^\dagger}(\cdot|s_0 \sim \mu)$ denote the one corresponding to transition model P_t. Then for any function $f(s,a)$, we have

$$\mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}}\left[f(s,a)\right] = \mathbb{E}_{\tau\sim P_t^{\pi^\dagger}(\cdot|s_0\sim\mu)}\left[\sum_{h=0}^{\infty}\gamma^h f(s_h,a_h)\right].$$

$$\mathbb{E}_{(s,a)\sim d_{P^*}^{\pi^\dagger}}\left[f(s,a)\right] = \mathbb{E}_{\tau\sim P_*^{\pi^\dagger}(\cdot|s_0\sim\mu)}\left[\sum_{h=0}^{\infty}\gamma^h f(s_h,a_h)\right].$$

Also note that in the special case $f(s,a) = r(s,a)$, we have $\mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}}\left[r(s,a)\right] = V_{P_t}^{\pi^\dagger}$, $\mathbb{E}_{(s,a)\sim d_{P^*}^{\pi^\dagger}}\left[r(s,a)\right] = V_{P^*}^{\pi^\dagger}$.

Now consider two MDPs: $\widetilde{M}_t = (\mathcal{S}, \mathcal{A}, P_t, \widetilde{r}, \mu, \gamma)$ and $\widetilde{M}^* = (\mathcal{S}, \mathcal{A}, P^*, \widetilde{r}, \mu, \gamma)$, where $\widetilde{r}(s,a) = w_t \cdot \nabla_\theta \log \pi_t(a|s)$ which is a function defined on $\mathcal{S}\times\mathcal{A}$. We focus on the policy $\pi^\dagger$, and evaluate it under both MDPs.

With these notations, we can rewrite a_t as:

$$\begin{aligned} a_t &= \mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}}\left[w_t \cdot \nabla_\theta \log \pi_t(a|s)\right] - \mathbb{E}_{(s,a)\sim d_{P^*}^{\pi^\dagger}}\left[w_t \cdot \nabla_\theta \log \pi_t(a|s)\right] \\ &= \mathbb{E}_{\tau\sim P_t^{\pi^\dagger}(\cdot|s_0\sim\mu)}\left[\sum_{h=0}^{\infty}\gamma^h \widetilde{r}(s_h,a_h)\right] - \mathbb{E}_{\tau\sim P_*^{\pi^\dagger}(\cdot|s_0\sim\mu)}\left[\sum_{h=0}^{\infty}\gamma^h \widetilde{r}(s_h,a_h)\right] \\ &= V_{\widetilde{M}_t}^{\pi^\dagger} - V_{\widetilde{M}^*}^{\pi^\dagger}. \end{aligned}$$

Note that the two MDPs share all the settings except transition probabilities, and the value functions are evaluated for the same policy $\pi^\dagger$. Then we can apply the second part of lemma 7 to them:

$$a_t = V_{\widetilde{M}_t}^{\pi^\dagger} - V_{\widetilde{M}^*}^{\pi^\dagger} \le \left|V_{\widetilde{M}_t}^{\pi^\dagger} - V_{\widetilde{M}^*}^{\pi^\dagger}\right| \le 2C\frac{\gamma}{1-\gamma}\mathbb{E}_{s,a\sim d_{P^*}^{\pi^\dagger}}\left[\mathrm{TV}(P_t(\cdot \mid s,a), P_*(\cdot \mid s,a))\right], \quad (12)$$

where C should satisfy $-C \leq V^{\pi^\dagger}_{\widetilde{M}_t}(s) \leq C,\ \forall s \in \mathcal{S}$.

By lemma 8, $\nabla_\theta \log \pi_t(a|s) = \phi_{s,a} - \mathbb{E}_{a' \sim \pi_t(\cdot|s)} \phi_{s,a'}$. By assumption in theorem 1, $\|\phi_{s,a}\|_2 \leq B$ for any (s,a). So $\|\nabla_\theta \log \pi_t(a|s)\|_2 \leq 2B$. By algorithm 1, $\|w_t\|_2 \leq W$. Then Cauchy inequality implies

$$|\widetilde{r}(s,a)| = |w_t \cdot \nabla_\theta \log \pi_t(a|s)| \leq \|w_t\|_2 \, \|\nabla_\theta \log \pi_t(a|s)\|_2 \leq 2BW.$$

The bound of reward function immediately leads to a bound for value function: $|V^{\pi^\dagger}_{\widetilde{M}_t}(s)| \leq \frac{2BW}{1-\gamma},\ \forall s \in \mathcal{S}$. So we can set $C = \frac{2BW}{1-\gamma}$ in (12):

$$\begin{aligned} a_t &\leq \frac{4\gamma BW}{(1-\gamma)^2} \mathbb{E}_{s,a \sim d^{\pi^\dagger}_{P^*}} \left[\mathrm{TV}(P_t(\cdot \mid s,a), P_*(\cdot \mid s,a)) \right] \\ &\leq \frac{4\gamma BW}{(1-\gamma)^2} \sqrt{\mathbb{E}_{s,a \sim d^{\pi^\dagger}_{P^*}} \left[\mathrm{TV}(P_t(\cdot \mid s,a), P_*(\cdot \mid s,a))^2 \right]}. \end{aligned} \tag{13}$$

By definition of concentration coefficient, we have

$$\mathbb{E}_{(s,a) \sim d^{\pi^\dagger}_{P^*}} \left[\mathrm{TV}(P_t(\cdot|s,a), P^*(\cdot|s,a))^2 \right] \leq C_{\pi^\dagger} \mathbb{E}_{(s,a) \sim \rho} \left[\mathrm{TV}(P_t(\cdot|s,a), P^*(\cdot|s,a))^2 \right]. \tag{14}$$

By definition 2,

$$\mathbb{E}_{(s,a) \sim \rho} \left[\mathrm{TV}(P_t(\cdot|s,a), P^*(\cdot|s,a))^2 \right] \leq \varepsilon^2_{\text{est}}. \tag{15}$$

Combine (13),(14),(15) together:

$$a_t \leq \frac{4\gamma BW}{(1-\gamma)^2} \sqrt{C_{\pi^\dagger}}\, \varepsilon_{\text{est}}.$$

□

Proof of lemma 5. By lemma 9, $\log \pi_\theta(a|s)$ is β-smooth, and β can be set as B^2.

By property of β-smooth function, we have

$$\left| \log \pi_{\theta_{t+1}}(a|s) - \log \pi_{\theta_t}(a|s) - \nabla_\theta \log \pi_{\theta_t}(a|s) \cdot (\theta_{t+1} - \theta_t) \right| \leq \frac{B^2}{2} \|\theta_{t+1} - \theta_t\|_2^2.$$

Rearrange it to:

$$\nabla_\theta \log \pi_{\theta_t}(a|s) \cdot (\theta_{t+1} - \theta_t) \leq \log \pi_{\theta_{t+1}}(a|s) - \log \pi_{\theta_t}(a|s) + \frac{B^2}{2} \|\theta_{t+1} - \theta_t\|_2^2.$$

Note that under our notations, π_t is the shorthand of π_{θ_t}. Recall that $\theta_{t+1} - \theta_t = \eta w_t$, then by the definition of b_t: (see (8))

$$\begin{aligned} b_t &= \mathbb{E}_{(s,a) \sim d^{\pi^\dagger}_{P^*}} \left[w_t \cdot \nabla_\theta \log \pi_t(a|s) \right] = \frac{1}{\eta} \mathbb{E}_{(s,a) \sim d^{\pi^\dagger}_{P^*}} \left[(\theta_{t+1} - \theta_t) \cdot \nabla_\theta \log \pi_t(a|s) \right] \\ &\leq \frac{1}{\eta} \mathbb{E}_{(s,a) \sim d^{\pi^\dagger}_{P^*}} \left[\log \pi_{t+1}(a|s) - \log \pi_t(a|s) + \frac{B^2}{2} \eta^2 \|w_t\|_2^2 \right] \\ &\leq \frac{\eta B^2 W^2}{2} + \frac{1}{\eta} \mathbb{E}_{(s,a) \sim d^{\pi^\dagger}_{P^*}} \left[\log \pi_{t+1}(a|s) - \log \pi_t(a|s) \right] \\ &= \frac{\eta B^2 W^2}{2} + \frac{1}{\eta} \mathbb{E}_{s \sim d^{\pi^\dagger}_{P^*}} \left[\mathbb{E}_{a \sim \pi^\dagger(\cdot|s)} \log \pi_{t+1}(a|s) - \mathbb{E}_{a \sim \pi^\dagger(\cdot|s)} \log \pi_t(a|s) \right] \\ &= \frac{\eta B^2 W^2}{2} + \frac{1}{\eta} \mathbb{E}_{s \sim d^{\pi^\dagger}_{P^*}} \left[\mathrm{KL}\left(\pi^\dagger(\cdot|s) \,||\, \pi_t(\cdot|s) \right) - \mathrm{KL}\left(\pi^\dagger(\cdot|s) \,||\, \pi_{t+1}(\cdot|s) \right) \right]. \end{aligned}$$

Take average for $0 \le t \le T$:

$$\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T} b_t &\le \frac{\eta B^2 W^2}{2} \\
&\quad + \frac{1}{\eta(T+1)}\sum_{t=0}^{T} \mathbb{E}_{s\sim d_{P^*}^{\pi^\dagger}} \left[\mathrm{KL}\left(\pi^\dagger(\cdot|s) \,||\, \pi_t(\cdot|s)\right) - \mathrm{KL}\left(\pi^\dagger(\cdot|s) \,||\, \pi_{t+1}(\cdot|s)\right)\right] \\
&= \frac{\eta B^2 W^2}{2} \\
&\quad + \frac{1}{\eta(T+1)}\mathbb{E}_{s\sim d_{P^*}^{\pi^\dagger}} \left[\mathrm{KL}\left(\pi^\dagger(\cdot|s) \,||\, \pi_0(\cdot|s)\right) - \mathrm{KL}\left(\pi^\dagger(\cdot|s) \,||\, \pi_{T+1}(\cdot|s)\right)\right] \\
&\le \frac{\eta B^2 W^2}{2} + \frac{1}{\eta(T+1)}\mathbb{E}_{s\sim d_{P^*}^{\pi^\dagger}} \left[\mathrm{KL}\left(\pi^\dagger(\cdot|s) \,||\, \pi_0(\cdot|s)\right)\right],
\end{aligned}$$

where the last step is because KL divergence is non-negative.

In algorithm 1, we set $\pi_0(\cdot|s) = \mathrm{Uniform}(\mathcal{A})$. Then for any probability measure q on $\mathcal{A}$,

$$\begin{aligned}
\mathrm{KL}\left(q(\cdot) \,||\, \pi_0(\cdot|s)\right) = \mathbb{E}_{a\sim q(\mathcal{A})}\left[\log\frac{q(a)}{\pi_0(a|s)}\right] &= \mathbb{E}_{a\sim q(\mathcal{A})}\left[\log(q(a)|\mathcal{A}|)\right] \\
&\le \mathbb{E}_{a\sim q(\mathcal{A})}\left[\log|\mathcal{A}|\right] = \log|\mathcal{A}|.
\end{aligned}$$

So

$$\frac{1}{T+1}\sum_{t=0}^{T} b_t \le \frac{\eta B^2 W^2}{2} + \frac{\log|\mathcal{A}|}{\eta(T+1)}.$$

□

Proof of lemma 6. Recall the definition of c_t: (see (9))

$$c_t = \mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}}\left[A_{P_t}^{\pi_t}(s,a) - w_t \cdot \nabla_\theta \log \pi_t(a|s)\right].$$

By lemma 8,

$$\nabla_\theta \log \pi_t(a|s) = \phi_{s,a} - \mathbb{E}_{a'\sim\pi_t(\cdot|s)}\phi_{s,a'}.$$

We also have

$$A_{P_t}^{\pi_t}(s,a) = Q_{P_t}^{\pi_t}(s,a) - V_{P_t}^{\pi_t}(s) = Q_{P_t}^{\pi_t}(s,a) - \mathbb{E}_{a'\sim\pi_t(\cdot|s)}Q_{P_t}^{\pi_t}(s,a').$$

So we can rewrite c_t as

$$\begin{aligned}
c_t &= \mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}}\left[Q_{P_t}^{\pi_t}(s,a) - w_t\cdot\phi_{s,a}\right] + \mathbb{E}_{(s,a)\sim d_{P_t}^{\pi^\dagger}}\mathbb{E}_{a'\sim\pi_t(\cdot|s)}\left[w_t\cdot\phi_{s,a'} - Q_{P_t}^{\pi_t}(s,a')\right] \\
&= \mathbb{E}_{s\sim d_{P_t}^{\pi^\dagger}, a\sim\pi^\dagger(\cdot|s)}\left[Q_{P_t}^{\pi_t}(s,a) - w_t\cdot\phi_{s,a}\right] + \mathbb{E}_{s\sim d_{P_t}^{\pi^\dagger}, a'\sim\pi_t(\cdot|s)}\left[w_t\cdot\phi_{s,a'} - Q_{P_t}^{\pi_t}(s,a')\right] \\
&\le \mathbb{E}_{s\sim d_{P_t}^{\pi^\dagger}, a\sim\pi^\dagger(\cdot|s)}\left[|Q_{P_t}^{\pi_t}(s,a) - w_t\cdot\phi_{s,a}|\right] + \mathbb{E}_{s\sim d_{P_t}^{\pi^\dagger}, a'\sim\pi_t(\cdot|s)}\left[|w_t\cdot\phi_{s,a'} - Q_{P_t}^{\pi_t}(s,a')|\right] \\
&\le 2|\mathcal{A}|\mathbb{E}_{s\sim d_{P_t}^{\pi^\dagger}, a\sim\mathrm{Unif}(\mathcal{A})}\left[|Q_{P_t}^{\pi_t}(s,a) - w_t\cdot\phi_{s,a}|\right] \\
&= 2|\mathcal{A}|\mathbb{E}_{s\sim d_{P^*}^{\pi^\dagger}, a\sim\mathrm{Unif}(\mathcal{A})}\left[f(s,a)\right] \\
&\quad + (2|\mathcal{A}|\mathbb{E}_{s\sim d_{P_t}^{\pi^\dagger}, a\sim\mathrm{Unif}(\mathcal{A})}\left[f(s,a)\right] - 2|\mathcal{A}|\mathbb{E}_{s\sim d_{P^*}^{\pi^\dagger}, a\sim\mathrm{Unif}(\mathcal{A})}\left[f(s,a)\right]),
\end{aligned}$$

where $f(s,a) := |Q_{P_t}^{\pi_t}(s,a) - w_t\cdot\phi_{s,a}|$.

Note that in the inequality above, we used the result that for any non-negative function $f(a)$ and any probability distribution q on $\mathcal{A}$,

$$\mathbb{E}_{a\sim q} f(a) = \sum_{a\in\mathcal{A}} f(a)q(a) \leq \sum_{a\in\mathcal{A}} f(a) = |\mathcal{A}|\mathbb{E}_{a\sim \text{unif}(\mathcal{A})} f(a).$$

For the term $\mathbb{E}_{s\sim d^{\pi^\dagger}_{P_t}, a\sim \text{Unif}(\mathcal{A})}[f(s,a)] - \mathbb{E}_{s\sim d^{\pi^\dagger}_{P^*}, a\sim \text{Unif}(\mathcal{A})}[f(s,a)]$, we first take expectation with respect to $a \sim \text{Unif}(\mathcal{A})$, then we have $\mathbb{E}_{s\sim d^{\pi^\dagger}_{P_t}}\left[\frac{1}{|\mathcal{A}|}\sum_a f(s,a)\right]$. Let $\widetilde{f}(s) := \frac{1}{|\mathcal{A}|}\sum_a f(s,a)$. By viewing $\widetilde{f}(s)$ as a reward function $\widetilde{r}(s,a)$ which is the same for each a, i.e., $\widetilde{r}(s,a) = \widetilde{f}(s)$ for all a, then we have

$$\mathbb{E}_{s\sim d^{\pi^\dagger}_{P_t}}\left[\widetilde{f}(s)\right] = \mathbb{E}_{s\sim d^{\pi^\dagger}_{P_t}}\left[\widetilde{r}(s,a)\right] = \mathbb{E}_{s\sim d^{\pi^\dagger}_{P_t}, a\sim\pi^\dagger}\left[\widetilde{r}(s,a)\right].$$

Then we can use the same technique we used in the proof of lemma 4, consider a new MDP with a reward function $\widetilde{r}$, then use simulation lemma.

By using the bound $|\widetilde{f}| \leq WB + \frac{1}{(1-\gamma)}$, we immediately have the value function $\widetilde{V}$ induced by reward $\widetilde{r}$ satisfies $|\widetilde{V}| \leq \frac{1}{1-\gamma}(WB + \frac{1}{(1-\gamma)})$. Apply the simulation lemma (lemma 7) to the term $\mathbb{E}_{s\sim d^{\pi^\dagger}_{P_t}}\widetilde{f}(s) - \mathbb{E}_{s\sim d^{\pi^\dagger}_{P^*}}\widetilde{f}(s)$, we get

$$\frac{2\gamma}{(1-\gamma)^2}\left(WB + \frac{1}{(1-\gamma)}\right)\mathbb{E}_{(s,a)\sim d^{\pi^\dagger}_{P^*}}\left[\text{TV}(P_t(\cdot|s,a), P^*(\cdot|s,a))\right],$$

which is further bounded by $\frac{2\gamma}{(1-\gamma)^2}(WB + \frac{\gamma}{(1-\gamma)})\sqrt{C_{\pi^\dagger}}\,\varepsilon_{est}$ (see 11).

For the first term

$$\mathbb{E}_{s\sim d^{\pi^\dagger}_{P^*}, a\sim\text{Unif}(\mathcal{A})}\left[f(s,a)\right],$$

we have

$$\begin{aligned}
\mathbb{E}_{s\sim d^{\pi^\dagger}_{P^*}, a\sim\text{Unif}(\mathcal{A})}\left[f(s,a)\right] &\leq \sqrt{\mathbb{E}_{s\sim d^{\pi^\dagger}_{P^*}, a\sim\text{Unif}(\mathcal{A})}\left[f(s,a)^2\right]} \\
&\leq \left\|\frac{d^{\pi^\dagger}_{P^*}(s)\circ \text{Unif}_{\mathcal{A}}(a)}{d^{\pi_t}_{P_t}(s)\circ \text{Unif}_{\mathcal{A}}(a)}\right\|_\infty^{\frac{1}{2}} \sqrt{\mathbb{E}_{s\sim d^{\pi_t}_{P_t}, a\sim\text{Unif}(\mathcal{A})}\left[f(s,a)^2\right]} \\
&\leq \left\|\frac{d^{\pi^\dagger}_{P^*}(s)}{d^{\pi_t}_{P_t}(s)}\right\|_\infty^{\frac{1}{2}} \varepsilon_{\text{approx}} \\
&\leq \frac{1}{1-\gamma}\left\|\frac{d^{\pi^\dagger}_{P^*}(s)}{\mu_0(s)}\right\|_\infty^{\frac{1}{2}} \varepsilon_{\text{approx}}.
\end{aligned}$$

Thus we have

$$c_t \leq \frac{2|\mathcal{A}|}{1-\gamma}\left\|\frac{d^{\pi^\dagger}_{P^*}(s)}{\mu_0(s)}\right\|_\infty^{\frac{1}{2}} \varepsilon_{\text{approx}} + \frac{4\gamma|\mathcal{A}|}{(1-\gamma)^2}\left(WB + \frac{\gamma}{(1-\gamma)}\right)\sqrt{C_{\pi^\dagger}}\,\varepsilon_{est}.$$

□

C Proof of Auxiliary lemmas

The following lemma is a helpful supporting lemma. Simulation lemma basically focuses on the difference in value functions (or Q functions) of fixed policy under different transition models. It is common in RL literature, and there are different version of variants. We will prove the following version, which will be used in the proof of lemma 2 and lemma 4.

Lemma 7 (A generalization of simulation lemma)**.** *Suppose $\mathcal{S}$, $\mathcal{A}$, r, γ, μ_0 are all fixed. Here $\mathcal{S}$ and $\mathcal{A}$ can be infinite sets, and $r : \mathcal{S} \to \mathbb{R}$ can be any real value function. For two arbitrary transition models P and $\widehat{P}$, and any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, we have*

$$V_P^\pi - V_{\widehat{P}}^\pi = \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi}\left[\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V_{\widehat{P}}^\pi(s')\right] - \mathbb{E}_{s'\sim \widehat{P}(\cdot|s,a)}\left[V_{\widehat{P}}^\pi(s')\right]\right].$$

If $V_{\widehat{P}}^\pi(s)$ is bounded, i.e. $-C \le V_{\widehat{P}}^\pi(s) \le C,\ \forall s \in \mathcal{S}$, then we further have

$$\left|V_P^\pi - V_{\widehat{P}}^\pi\right| \le 2C\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi}\left[\mathrm{TV}(P(\cdot|s,a), \widehat{P}(\cdot|s,a))\right].$$

If $V_{\widehat{P}}^\pi(s)$ is positive and bounded, i.e. $0 \le V_{\widehat{P}}^\pi(s) \le C,\ \forall s \in \mathcal{S}$, then

$$\left|V_P^\pi - V_{\widehat{P}}^\pi\right| \le C\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi}\left[\mathrm{TV}(P(\cdot|s,a), \widehat{P}(\cdot|s,a))\right].$$

Proof. We first prove the first part of the lemma.

Let $d_P^\pi(\cdot,\cdot|s_0,a_0)$ denote the visitation measure over (s,a) conditioning on $(S_0 = s_0, A_0 = a_0)$ under transition model P, i.e. $d_P^\pi(\cdot,\cdot|s_0,a_0) = (1-\gamma)\sum_{t=0}^\infty \gamma^t P^\pi(S_t = \cdot\,, A_t = \cdot\,|s_0,a_0)$.

Then we have for any (s_0,a_0),

$$Q_P^\pi(s_0,a_0) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}[r(s,a)]. \tag{16}$$

By Bellman equation, for any (s,a),

$$Q_P^\pi(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi(\cdot|s')}\left[Q_P^\pi(s',a')\right]. \tag{17}$$

$$Q_{\widehat{P}}^\pi(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim \widehat{P}(\cdot|s,a),a'\sim\pi(\cdot|s')}\left[Q_{\widehat{P}}^\pi(s',a')\right]. \tag{18}$$

Substitute the $r(s,a)$ in (16) by the $r(s,a)$ in (18):

$$Q_P^\pi(s_0,a_0) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}\left[Q_{\widehat{P}}^\pi(s,a) - \gamma\mathbb{E}_{s'\sim\widehat{P}(\cdot|s,a),a'\sim\pi(\cdot|s')}Q_{\widehat{P}}^\pi(s',a')\right]. \tag{19}$$

By (16) and (17), we first apply (17) to the $Q_P^\pi(s_0,a_0)$ in (16), then apply (16) iteratively:

$$\begin{aligned}
&\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}[r(s,a)] \\
&= Q_P^\pi(s_0,a_0) \\
&= r(s_0,a_0) + \gamma\mathbb{E}_{s\sim P(\cdot|s_0,a_0),a\sim\pi(\cdot|s)}\left[Q_P^\pi(s,a)\right] \\
&= r(s_0,a_0) + \gamma\mathbb{E}_{s\sim P(\cdot|s_0,a_0),a\sim\pi(\cdot|s)}\left[\frac{1}{1-\gamma}\mathbb{E}_{(s',a')\sim d_P^\pi(\cdot,\cdot|s,a)}[r(s',a')]\right].
\end{aligned}$$

Rearrange it as

$$\begin{aligned}
-r(s_0,a_0) = \frac{\gamma}{1-\gamma}&\mathbb{E}_{s\sim P(\cdot|s_0,a_0),a\sim\pi(\cdot|s)}\left[\mathbb{E}_{(s',a')\sim d_P^\pi(\cdot,\cdot|s,a)}[r(s',a')]\right] \\
&- \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}[r(s,a)].
\end{aligned}$$

Note that the equation above holds for any real function $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, so we can replace $r(\cdot,\cdot)$ by $Q_{\widehat{P}}^\pi(\cdot,\cdot)$

$$
\begin{aligned}
-r(s_0,a_0) = &\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim P(\cdot|s_0,a_0),a\sim\pi(\cdot|s)}\left[\mathbb{E}_{(s',a')\sim d_P^\pi(\cdot,\cdot|s,a)}[r(s',a')]\right] \\
&- \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}[r(s,a)].
\end{aligned} \tag{20}
$$

(19)+(20):

$$
\begin{aligned}
Q_P^\pi(s_0,a_0) - Q_{\widehat{P}}^\pi(s_0,a_0) = &\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim P(\cdot|s_0,a_0),a\sim\pi(\cdot|s)}\left[\mathbb{E}_{(s',a')\sim d_P^\pi(\cdot,\cdot|s,a)}Q_{\widehat{P}}^\pi(s',a')\right] \\
&- \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}\left[\mathbb{E}_{s'\sim\widehat{P}(\cdot|s,a),a'\sim\pi(\cdot|s')}Q_{\widehat{P}}^\pi(s',a')\right].
\end{aligned} \tag{21}
$$

Consider the first term on right hand side:

$$
\begin{aligned}
\mathbb{E}_{s\sim P(\cdot|s_0,a_0),a\sim\pi(\cdot|s)}\mathbb{E}_{(s',a')\sim d_P^\pi(\cdot,\cdot|s,a)}[\cdot] &= \mathbb{E}_{(s',a')\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}[\cdot] \\
&= \mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi(\cdot|s')}[\cdot].
\end{aligned}
$$

So (21) can be rewritten as

$$
\begin{aligned}
&Q_P^\pi(s_0,a_0) - Q_{\widehat{P}}^\pi(s_0,a_0) \\
&= \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}\left[\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi(\cdot|s')}Q_{\widehat{P}}^\pi(s',a') - \mathbb{E}_{s'\sim\widehat{P}(\cdot|s,a),a'\sim\pi(\cdot|s')}Q_{\widehat{P}}^\pi(s',a')\right] \\
&= \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}\left[\mathbb{E}_{s'\sim P(\cdot|s,a)}V_{\widehat{P}}^\pi(s') - \mathbb{E}_{s'\sim\widehat{P}(\cdot|s,a)}V_{\widehat{P}}^\pi(s')\right].
\end{aligned}
$$

Finally, consider $V_P^\pi(s_0)$, $V_{\widehat{P}}^\pi(s_0)$ and the initial distribution μ. Recall that d_P^π is the visitation measure conditioning on the initial distribution μ. So we have

$$
\begin{aligned}
V_P^\pi - V_{\widehat{P}}^\pi &= \mathbb{E}_{s_0\sim\mu}\left[V_P^\pi(s_0) - V_{\widehat{P}}^\pi(s_0)\right] \\
&= \mathbb{E}_{s_0\sim\mu,a_0\sim\pi(\cdot|s_0)}\left[Q_P^\pi(s_0,a_0) - Q_{\widehat{P}}^\pi(s_0,a_0)\right] \\
&= \frac{\gamma}{1-\gamma}\mathbb{E}_{s_0\sim\mu,a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{(s,a)\sim d_P^\pi(\cdot,\cdot|s_0,a_0)}\left[\mathbb{E}_{s'\sim P(\cdot|s,a)}V_{\widehat{P}}^\pi(s') - \mathbb{E}_{s'\sim\widehat{P}(\cdot|s,a)}V_{\widehat{P}}^\pi(s')\right] \\
&= \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi}\left[\mathbb{E}_{s'\sim P(\cdot|s,a)}V_{\widehat{P}}^\pi(s') - \mathbb{E}_{s'\sim\widehat{P}(\cdot|s,a)}V_{\widehat{P}}^\pi(s')\right],
\end{aligned}
$$

which finishes the first part of the lemma.

Then we prove the second part: first note that

$$
\begin{aligned}
\left|V_P^\pi - V_{\widehat{P}}^\pi\right| &= \frac{\gamma}{1-\gamma}\left|\mathbb{E}_{(s,a)\sim d_P^\pi}\left[\mathbb{E}_{s'\sim P(\cdot|s,a)}V_{\widehat{P}}^\pi(s') - \mathbb{E}_{s'\sim\widehat{P}(\cdot|s,a)}V_{\widehat{P}}^\pi(s')\right]\right| \\
&\le \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d_P^\pi}\left|\mathbb{E}_{s'\sim P(\cdot|s,a)}V_{\widehat{P}}^\pi(s') - \mathbb{E}_{s'\sim\widehat{P}(\cdot|s,a)}V_{\widehat{P}}^\pi(s')\right|.
\end{aligned} \tag{22}
$$

Suppose q_1, q_2 are two arbitrary probability distributions, and C is a constant satisfying $-C \le f(x) \le C$. By property of total variation distance, $\mathrm{TV}(q_1,q_2) = \frac{1}{2}\|q_1 - q_2\|_1$.

By Hölder inequality

$$
\begin{aligned}
|\mathbb{E}_{x\sim q_1} f(x) - \mathbb{E}_{x\sim q_2} f(x)| &= \left| \int f(x)(q_1(x) - q_2(x))dx \right| \\
&= \|f(q_1 - q_2)\|_1 \le \|f\|_\infty \|q_1 - q_2\|_1 \le 2C\mathrm{TV}(q_1, q_2).
\end{aligned} \tag{23}
$$

Apply (23) to the right hand side of (22):

$$
\left| V_P^\pi - V_{\widehat{P}}^\pi \right| \le 2C \frac{\gamma}{1-\gamma} \mathbb{E}_{(s,a)\sim d_P^\pi} \left[\mathrm{TV}(P(\cdot|s,a), \widehat{P}(\cdot|s,a)) \right],
$$

which concludes the second part.

Third part: Consider the special case that $0 \le f(x) \le C$, then we can improve the upper bound in (23)

$$
\begin{aligned}
&|\mathbb{E}_{x\sim q_1} f(x) - \mathbb{E}_{x\sim q_2} f(x)| \\
&= \left| \int f(x)(q_1(x) - q_2(x))dx \right| \\
&= \left| \int f(x)(q_1(x) - q_2(x))\mathbf{1}\{q_1(x) > q_2(x)\}dx - \int f(x)(q_2(x) - q_1(x))\mathbf{1}\{q_1(x) \le q_2(x)\}dx \right|.
\end{aligned}
$$

Note that on the right hand side, the two terms inside the absolute value sign are both non-negative, so

$$
\begin{aligned}
&|\mathbb{E}_{x\sim q_1} f(x) - \mathbb{E}_{x\sim q_2} f(x)| \\
&\le \max\left\{ \int f(x)(q_1(x) - q_2(x))\mathbf{1}\{q_1(x) > q_2(x)\}dx, \int f(x)(q_2(x) - q_1(x))\mathbf{1}\{q_1(x) \le q_2(x)\}dx \right\} \\
&\le C \max\left\{ \int (q_1(x) - q_2(x))\mathbf{1}\{q_1(x) > q_2(x)\}dx, \int (q_2(x) - q_1(x))\mathbf{1}\{q_1(x) \le q_2(x)\}dx \right\} \\
&= C\mathrm{TV}(q_1, q_2),
\end{aligned}
$$

where the last step is an equivalent definition of total variation distance (for two probability distributions).

So the factor 2 on the right hand side in (23) can be improved to 1 in this case. □

The following two lemmas (lemma 8 and 9) are useful properties for the log-linear parametric class (defined in (2)).

Lemma 8. *For any policy π_θ in the log-linear parametric class (see (2)), the following holds for any (s, a):*

$$
\nabla_\theta \log \pi_\theta(a|s) = \phi_{s,a} - \mathbb{E}_{a'\sim\pi_\theta(\cdot|s)} \phi_{s,a'}.
$$

Proof. Recall that

$$
\pi_\theta(a|s) = \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a'\in\mathcal{A}} \exp(\theta \cdot \phi_{s,a'})},
$$

$$
\log \pi_\theta(a|s) = \theta \cdot \phi_{s,a} - \log\left(\sum_{a'\in\mathcal{A}} \exp(\theta \cdot \phi_{s,a'}) \right).
$$

So we have

$$\begin{aligned}\nabla_\theta \log \pi_\theta(a|s) &= \phi_{s,a} - \nabla_\theta \log \left(\sum_{a' \in \mathcal{A}} \exp(\theta \cdot \phi_{s,a'}) \right) \\ &= \phi_{s,a} - \frac{\sum_{a'' \in \mathcal{A}} \exp(\theta \cdot \phi_{s,a''}) \phi_{s,a''}}{\sum_{a' \in \mathcal{A}} \exp(\theta \cdot \phi_{s,a'})} \\ &= \phi_{s,a} - \sum_{a'' \in \mathcal{A}} \pi_\theta(a''|s) \phi_{s,a''}.\end{aligned}$$

□

Lemma 9. *For any policy π_θ in the log-linear parametric class (see (2)), if $\|\phi_{s,a}\|_2 \leq B$ for any (s,a), then $\log \pi_\theta(a|s)$ is a B^2-smooth function (as a function of θ) for any (s,a), i.e.*

$$\|\nabla_\theta \log \pi_{\theta_1}(a|s) - \nabla_\theta \log \pi_{\theta_2}(a|s)\|_2 \leq B^2 \|\theta_1 - \theta_2\|_2$$

for any s,a,θ_1,θ_2.

Proof. By lemma 8,

$$\nabla_\theta \log \pi_\theta(a|s) = \phi_{s,a} - \sum_{a' \in \mathcal{A}} \pi_\theta(a'|s) \phi_{s,a'}.$$

For convenience, let

$$g(\theta) := \sum_{a' \in \mathcal{A}} \pi_\theta(a'|s) \phi_{s,a'}.$$

Then we only need to prove $\|g(\theta_1) - g(\theta_2)\|_2 \leq B^2 \|\theta_1 - \theta_2\|_2$ for any θ_1,θ_2.

Rewrite $g(\theta)$ into an explicit form of θ:

$$g(\theta) = \sum_{a \in \mathcal{A}} \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})} \phi_{s,a} = \frac{\sum_a \phi_{s,a} \exp(\theta \cdot \phi_{s,a})}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})}.$$

Note that $g : \mathbb{R}^d \to \mathbb{R}^d$, so the Jacobian matrix of $g(\theta)$, i.e. $J(\theta)$, is $d \times d$:

$$\begin{aligned}J(\theta) &= \frac{1}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})} \sum_a \phi_{s,a} \phi_{s,a}^T \exp(\theta \cdot \phi_{s,a}) \\ &\quad - \frac{\sum_a \phi_{s,a} \exp(\theta \cdot \phi_{s,a})}{(\sum_{a'} \exp(\theta \cdot \phi_{s,a'}))^2} \sum_{a'} \phi_{s,a'}^T \exp(\theta \cdot \phi_{s,a'}) \\ &= \sum_a \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})} \phi_{s,a} \phi_{s,a}^T \\ &\quad - \left(\sum_a \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})} \phi_{s,a} \right) \left(\sum_a \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})} \phi_{s,a} \right)^T \\ &= A_1 - A_2,\end{aligned} \tag{24}$$

where

$$A_1 = \sum_a \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})} \phi_{s,a} \phi_{s,a}^T$$

$$A_2 = \left(\sum_a \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})} \phi_{s,a} \right) \left(\sum_a \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})} \phi_{s,a} \right)^T = \phi_s^{(\theta)} \phi_s^{(\theta)T}$$

$$\phi_s^{(\theta)} = \sum_a \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a'} \exp(\theta \cdot \phi_{s,a'})} \phi_{s,a}.$$

Now consider the induced norm of matrix:

$$\|A\|_{2,2} = \sup_{\|x\|_2=1} \|Ax\|_2.$$

By the property of $\|\cdot\|_{2,2}$, it is indeed a matrix norm, so it satisfies triangular inequality.

Suppose $\psi \in \mathbb{R}^d$ and $\|\psi\|_2 \le B$, then

$$\|\psi\psi^T\|_{2,2} = \sup_{\|x\|_2=1} \|\psi\psi^T x\|_2 = \sup_{\|x\|_2=1} \sqrt{x^T\psi\psi^T\psi\psi^T x} = \|\psi\|_2 \sup_{\|x\|_2=1} |\psi^T x| \le \|\psi\|_2^2 \le B^2,$$

where the first inequality is by Cauchy inequality: $|\psi^T x| \le \|\psi\|_2\|x\|_2$.

By the assumption on $\|\phi_{s,a}\|_2$, we have $\|\phi_{s,a}\phi_{s,a}^T\|_{2,2} \le B^2$ for any (s,a). A_1 is a weighted average (with non-negative weights) of such matrices, so triangular inequality implies $\|A_1\|_{2,2} \le B^2$.

$\phi_s^{(\theta)}$ is a weighted average (with non-negative weights) of some $\phi_{s,a}$. So $\|\phi_s^{(\theta)}\|_2 \le B$, $\|A_2\|_{2,2} \le B^2$.

Suppose A is an arbitrary real symmetric matrix. Let $\sigma(A)$ denote the spectrum of A. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A. Then

$$\|A\|_{2,2} = \max_{\lambda\in\sigma(A)} |\lambda| = \max\{|\lambda_{\min}(A)|, |\lambda_{\max}(A)|\}.$$

Note that both A_1 and A_2 are real symmetric, and positive semi-definite. So we have

$$\begin{aligned}
\lambda_{\min}(J(\theta)) &= \min_{\|x\|_2=1} x^T J(\theta) x = \min_{\|x\|_2=1} x^T(A_1 - A_2)x \\
&\ge \min_{\|x\|_2=1} x^T(-A_2)x = -\max_{\|x\|_2=1} x^T A_2 x = -\lambda_{\max}(A_2) \\
&\ge -\|A_2\|_{2,2} \ge -B^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\lambda_{\max}(J(\theta)) = \max_{\|x\|_2=1} x^T J(\theta) x &= \min_{\|x\|_2=1} x^T(A_1 - A_2)x \\
&\le \min_{\|x\|_2=1} x^T A_1 x = \lambda_{\min}(A_1) \le \|A_1\|_{2,2} \le B^2.
\end{aligned}$$

So $\|J(\theta)\|_{2,2} = \max\{|\lambda_{\min}(J(\theta))|, |\lambda_{\max}(J(\theta))|\} \le B^2$, for any θ.

By the "$\mathbb{R}^d \to \mathbb{R}^d$ version" Mean Value Theorem,

$$g(\theta_1) - g(\theta_2) = \left(\int_0^1 J(\theta_2 + t(\theta_1 - \theta_2))dt\right)(\theta_1 - \theta_2),$$

where the right hand side is a $d \times d$ matrix multiply a $d \times 1$ column vector, and the integral is entry-wise.

Take $\|\cdot\|_2$:

$$
\begin{aligned}
\|g(\theta_1) - g(\theta_2)\|_2 &= \left\| \left(\int_0^1 J(\theta_2 + t(\theta_1 - \theta_2))dt \right) (\theta_1 - \theta_2) \right\|_2 \\
&\leq \|\theta_1 - \theta_2\|_2 \left\| \int_0^1 J(\theta_2 + t(\theta_1 - \theta_2))dt \right\|_{2,2} \\
&\leq \|\theta_1 - \theta_2\|_2 \int_0^1 \|J(\theta_2 + t(\theta_1 - \theta_2))\|_{2,2}\, dt \\
&\leq \|\theta_1 - \theta_2\|_2 \int_0^1 B^2 dt = B^2 \|\theta_1 - \theta_2\|_2.
\end{aligned}
$$

This finishes the proof. □

D PROOF OF COROLLARY 1

Proof. The proof of corollary 1 is straightforward following Uehara et al. (2020), who proved that $\varepsilon_{est} \leq \xi \frac{1}{\sqrt{C_{\pi^\dagger}}} \mathbb{E}_{(s,a)\sim d^{\pi^*}_{P^*}} \left[\|\phi(s,a)\|_{\Sigma_n^{-1}} \right] \leq c_1 \sqrt{\frac{\text{rank}[\Sigma_\rho]\{\text{rank}[\Sigma_\rho] + \ln(c_2/\delta)\}}{n}}$. By setting an appropriate ξ, the estimation error ε_{est} can be upper bounded by $c_1(1-\gamma)^{-2} \min\left(d^{1/2}, \bar{R}\right) \sqrt{\bar{R}} \sqrt{\frac{d_{\mathcal{S}} \ln(1+n)}{n}}$, $\bar{R} = \text{rank}\left[\Sigma_\rho\right] \{\text{rank}\left[\Sigma_\rho\right] + \ln\left(c_2/\delta\right)\}$.

Combining the upper bound of ε_{est} and appropriate design of ξ such that $P^* \in \mathcal{M}_{\mathcal{D}}$ with high probability, and we can easily obtain the result based on Theorem 1. □