Robust Offline Reinforcement Learning for Non-Markovian Decision Processes

Ruiquan Huang, Yingbin Liang, *Fellow, IEEE*, and Jing Yang, *Senior Member, IEEE*

***Abstract*—Distributionally robust offline reinforcement learning (RL) aims to find a policy that performs the best under the worst environment within an uncertainty set using an offline dataset collected from a nominal model. While recent advances in robust RL focus on Markov decision processes (MDPs), robust non-Markovian RL is limited to planning problem where the transitions in the uncertainty set are known. In this paper, we study the learning problem of robust offline non-Markovian RL. Specifically, when the nominal model admits a low-rank structure, we propose a new algorithm, featuring a novel dataset distillation and a lower confidence bound (LCB) design for robust values under different types of the uncertainty set. We also derive new dual forms for these robust values in non-Markovian RL, making our algorithm more amenable to practical implementation. By further introducing a novel type-I concentrability coefficient tailored for offline low-rank non-Markovian decision processes, we prove that our algorithm can find an ϵ-optimal robust policy using $O(1/\epsilon^2)$ offline samples. Moreover, we extend our algorithm to the case when the nominal model does not have specific structure. With a new type-II concentrability coefficient, the extended algorithm also enjoys polynomial sample efficiency under all different types of the uncertainty set.**

***Index Terms*—Robust RL, offline RL, POMDP, non-Markovian decision processes.**

I. Introduction

In REINFORCEMENT learning (RL) [1], an agent learns to make decisions by interacting with an unknown environment and maximizes the total received reward. This learning framework has found wide applications such as video games [2], robotics [3], and language model fine-tuning [4]. However, in many real-world scenarios such as autonomous driving and industrial control, direct training in the actual environment and generating data samples in an online manner is often unrealistic or even harmful. Offline RL [5], [6], which leverages dataset collected in the past for learning, thus becomes important. Yet, pre-collected data suffers from either noisy simulation or out-of-date measurements, and causes the so-called *Sim-to-real* problem [7], [8].

Distributionally robust RL [9] was proposed to address the aforementioned issue by learning an optimal policy under the worst environment in an *uncertainty set*. Robust RL has been studied under Markov decision processes (MDPs) with known transition probabilities in [10], [11], and [12]. It has also been studied for MDPs with unknown transitions [13], [14], [15], [16], [17]. Regarding robust RL under non-Markovian decision processes, only robust POMDP with known transitions in the uncertainty set has been studied in [18], [19], and [20]. To the best of our knowledge, there has not been any study on robust non-Markovian RL with unknown transitions.

Several challenges need to be handled for exploring the robust offline non-Markovian RL with unknown transitions. (i) Considering non-Markovian processes in the entire uncertainty set can cause dramatic increase of sample complexity. The challenge then lies in how to exploit the structure conditions of these non-Markovian processes for sample efficient design. (ii) Existing coverage condition for offline non-Markovian RL [21] is strong and can incur more stringent coverage conditions for *robust offline RL*. It is thus critical to relax such a condition to design desirable offline algorithms. These challenges pose the following fundamental open question for us to address:

Q: Can we design a provably efficient algorithm for learning **robust non-Markovian** decision processes using **offline data**, which (i) effectively exploit the structure of non-Markovian decision processes, and (ii) relaxes the coverage condition for an offline dataset?

Our contributions. In this paper, we provide the **first study** of robust non-Markovian RL with unknown transitions, i.e., unknown uncertainty set. We focus on the setting where an offline dataset is provided and is collected from one (unkown) *nominal* environment in an uncertainty set, and the environments are non-Markovian. We summarize our contributions as follows.

- **New algorithm design.** We propose a novel algorithm for robust offline RL when the nominal model admits a low-rank structure of predictive state representations (PSRs). In particular, we consider two types of uncertainty sets, namely $\mathcal{T}$-type and $\mathcal{P}$-type, for non-Markovian decision processes. Our algorithm features two new elements: (i) a novel *dataset distillation,* where the distilled dataset ensures that the estimated model from vanilla maximum likelihood estimation performs well on the distilled data; and (ii) *lower confidence bound design for the robust*

Received 11 November 2024; revised 28 April 2025; accepted 29 June 2025. Date of publication 10 July 2025; date of current version 20 August 2025. The work of Ruiquan Huang and Jing Yang was supported in part by U.S. National Science Foundation under Grant NSF CNS-1956276 and Grant ECCS-2133170. The work of Yingbin Liang was supported in part by U.S. National Science Foundation under Grant RINGS-2148253, Grant CNS-2112471, and Grant ECCS-2413528. *(Corresponding author: Jing Yang.)*

Ruiquan Huang is with the Department of Electrical Engineering, The Pennsylvania State University, University Park, PA 16802 USA.

Yingbin Liang is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA.

Jing Yang is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: yangjing@virginia.edu).

Communicated by H. H. Permuter, Associate Editor for Shannon Theory and Information Measures.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2025.3587509.

Digital Object Identifier 10.1109/TIT.2025.3587509

0018-9448 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

TABLE I
NOTATION TABLE

Symbol	Description
$\mathcal{O}$	Observation space
$\mathcal{A}$	Action space
π	Policy (mapping from histories to actions)
$\mathcal{T}_h(o_h\|o_{1:h}, a_{1:h})$	One-step transition from history $(o_{1:h}, a_{1:h})$ to the next observation o_{h+1}
$\mathcal{P}(o_{1:H}\|a_{1:H-1})$	Probability of observation sequence $o_{1:H}$ given an action sequence $a_{1:H-1}$
$\mathbb{D}$	Model dynamics, also used to denote the marginal distribution of the entire trajectory
$B^f_{\mathcal{T}}(\theta)$	Uncertainty set of $\mathcal{T}$-type under f-divergence with center (nominal model) θ
$B^f_{\mathcal{P}}(\theta)$	Uncertainty set of $\mathcal{P}$-type under f-divergence with center (nominal model) θ
$V^{\pi}_{\theta,R}$	Value function under policy π, model θ, and reward function R.
τ_h	History up to step h, e.g. $\tau_h = (o_{1:h}, a_{1:h})$
ω_h	Future from step $h+1$, e.g. $\omega_h = (o_{h+1:H}, a_{h+1:H})$
x_h	Available information for the agent at time step h, e.g. $x_h = (o_{1:h}, a_{1:h-1})$
$\mathbf{m}$	Feature of future trajectory
ψ	Feature of history trajectory
$\bar{\psi}$	Normalized feature of history, also known as the predictive state representation of the history
$\mathcal{Q}^A_h$	The set of core action sequence from step h
$C_{\mathrm{p}}(\pi\|\rho), C(\pi\|\rho)$	Concentrability coefficient
C_B	Wellness condition number
$\boldsymbol{\eta}, \gamma, \boldsymbol{\lambda}$	Dual variables for robust optimization

value, which is derived by leveraging the low-rank structure and the newly developed dual forms for robust values under both $\mathcal{T}$-type and $\mathcal{P}$-type uncertainty sets.

- **Novel coverage condition and theory for low-rank non-Markovian RL.** We characterize the near-optimality guarantee of our algorithm based on novel *type-I concentrability coefficient* for offline learning in low-rank non-Markovian decision processes. This coefficient is much more relaxed than that proposed for offline low-rank non-Markovian RL in [21]. This relaxation is due to refined analysis on the distributional shift. We also identify a wellness condition number of the nominal model. Our results suggests that as long as the concentrability coefficient and the condition number are finite, our algorithm can find a robust policy with suboptimal gap vanishing at a rate of $O(1/\sqrt{N})$, where N is the number of offline samples.
- **Algorithm and theory for general non-Markovian RL.** We develop a robust algorithm for general offline non-Markovian RL where the transition probabilities may not have any structure. We show that our algorithm can learn a near-optimal robust policy efficiently provided that the offline dataset has finite *Type-II concentrability coefficient*, which is also a much weaker condition than that proposed in [21].

II. RELATED WORK

Offline RL. Offline RL problem has been investigated extensively in the literature [5], [6], [22], [23], [24], [25], [26]. Many methods such as Q-learning [27], its variant Fitted Q-Iteration (FQI) [26], [28], [29], [30], [31], [32], [33], value iteration [34], [35], policy gradient [36], and model-based approaches [37], [38] have been shown to be effective in offline RL either theoretically or empirically. Among them, the pessimism principle is a popular approach to handle the distribution shift caused by an offline policy. Empirically, pessimistic approaches perform well on simulation control tasks [33], [39], [40], [41], [42]. On the theoretical side, pessimism allows us to obtain the Probably Approximately Correct (PAC) guarantee on various models under a coverage condition [27], [37], [38], [43], [44], [45]. In this work, we also adopt the pessimism principle and is the first to establish its efficiency for robust offline non-Markovian RL. We develop new coverage conditions in order to guarantee the provable efficiency.

Robust RL. Robust MDP (RMDP) was first introduced by [9] and [46]. Most of RMDP research in the literature focuses on the *planning problem* [10], [11], [12], [47], [48], [49], [50], [51], providing computationally efficient algorithms. Based on the data generation process, studies on the *learning problem* in RMDP can be primarily divided into three categories. RMDP with generative model has been studied in [13], [52], and [53], where a generative model can uniformly collect the offline data corresponding to each and every state-action pair. Online RMDP has been studied in [14], [15], [54], and [55], which allows the agent adaptively sample states and actions in the nominal model. Offline RMDP has been studied in [16] and [17], where only a pre-collected dataset is provided, which makes the problem more difficult due to distributional shift [6]. Reference [56] studied robust RL under risk-sensitive setting.

In addition to RMDP, several works have studied robust POMDPs with assumptions such as known transitions in the uncertainty set [18], [19], [20]. In particular, they focus on the planning problem in robust POMDPs, i.e., finding an optimal policy with full information of the uncertainty set, while our work focus on the learning problem that characterizes the sample complexity with unknown uncertainty sets. Moreover, [57] studied offline regular MDP, which is a sub-class of robust offline non-Markovian decision process considered in our work.

Non-Markovian RL. Learning in non-Markovian decision processes is both computationally and statistically intractable in general [58], [59], [60], [61], [62]. Recently, a line of research [21], [63], [64], [65], [66], [67], [68], [69], [70] has obtained polynomial sample complexity with various structural

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

conditions. Among them, low-rank structure [67], [68] has been proved to admit predictive state representations (PSR) [71], [72], and captures and generalizes a rich subclass of non-Markov decision-making problems such as MDPs and POMDPs. However, none of these works studies robust offline non-Markovian RL.

III. Preliminaries

Notations. For a vector x, $\|x\|_A$ stands for $\sqrt{x^\top A x}$, and its i-th coordinate is represented as $[x]_i$. For functions $\mathbb{P}$ and $\mathbb{Q}$ (not necessarily probability measures) over a set $\mathcal{X}$, the ℓ_1 distance between them is $\|\mathbb{P}-\mathbb{Q}\|_1 = \sum_x |\mathbb{P}(x)-\mathbb{Q}(x)|$, while the hellinger-squared distance is defined as $D_H^2(\mathbb{P}(x),\mathbb{Q}(x)) = \frac{1}{2}\sum_x(\sqrt{\mathbb{P}(x)}-\sqrt{\mathbb{Q}(x)})^2$. For a sequence $\{x_n, x_{n+1},\ldots,x_m\}$ with $n < m$, we denote it shortly as $x_{n:m}$. For a sequence of sets $\mathcal{S}_1,\mathcal{S}_2,\ldots,\mathcal{S}_n$, the Cartesian product is represented by $\mathcal{S}_1\times\mathcal{S}_2\times\cdots\times\mathcal{S}_n$. When these sets are identical, this product simplifies to $\mathcal{S}_1^n$. $\Delta(\mathcal{S})$ consists of all probability distributions over the set $\mathcal{S}$.

A. Non-Markovian Decision Processes

In this work, we consider finite-horizon episodic (non-Markovian) decision processes with an observation space $\mathcal{O}$, an action space $\mathcal{A}$ and a horizon H. The dynamics of the process is determined by a set of distributions $\{\mathcal{T}_h(\cdot|o_{1:h},a_{1:h}) \in \Delta(\mathcal{O})\}_{h=1}^{H-1}$, where $o_{1:h}\in\mathcal{O}^h$ denotes a sequence of observations up to time h, $a_{1:h}\in\mathcal{A}^h$ denotes a sequence of actions up to time h, and $\mathcal{T}_h(o_{h+1}|o_{1:h},a_{1:h})$ denotes the probability of observing o_{h+1} given history $o_{1:h},a_{1:h}$. A policy π is defined by a collection of distributions $\pi=\{\pi_h:\mathcal{O}^h\times\mathcal{A}^{h-1}\to\Delta(\mathcal{A})\}_{h=1}^H$ chosen by an agent, with each $\pi_h(\cdot|\tau_{h-1},o_h)$ being a distribution over $\mathcal{A}$, where $\tau_{h-1}=(o_1,a_1,\ldots,o_{h-1},a_{h-1})$ is the history up to step $h-1$. Specifically, at the beginning of each episode, the process starts at a fixed observation $o_1\in\mathcal{O}$ at time step 1. After observing o_1, the agent samples (takes) an action $a_1\in\mathcal{A}$ according to the distribution (policy) $\pi_1(\cdot|o_1)$, and the process transits to $o_2\in\mathcal{O}$, which is sampled according to the distribution $\mathcal{T}_1(\cdot|o_1,a_1)$. Then, at any time step $h\geq 2$, due to the non-Markovian nature, the agent samples an action a_h based on all past information $(o_{1:h},a_{1:h-1})$, and the process transits to o_{h+1}, sampled from $\mathcal{T}_h(o_{h+1}|o_{1:h},a_{1:h})$. The interaction terminates after time step H.

Notably, the general non-Markovian decision-making problem subsumes not only fully observable MDPs but also POMDPs as special cases. In MDPs, $\mathcal{T}_{h-1}(o_h|\tau_{h-1}) = \mathcal{T}_{h-1}(o_h|o_{h-1},a_{h-1})$, and in POMDPs, $\mathcal{T}_{h-1}(o_h|\tau_{h-1})$ can be factorized as $\mathcal{T}_{h-1}(o_h|\tau_{h-1}) = \sum_s \mathbb{O}_h(o_h|s_h)\mathbb{T}_{h-1}(s_h|s_{h-1},a_{h-1})\cdots\mathbb{T}_1(s_2|s_1,a_1)$, where $\{s_t\}_{t=1}^H$ represents unobserved states, $\{\mathbb{O}_t\}_{t=1}^H$ and $\{\mathbb{T}_t\}_{t=1}^H$ are called emission distributions and transition distributions, respectively.

We note that an alternative equivalent description of the process dynamics is the set of probabilities $\{\mathcal{P}(o_{1:H}|a_{1:H-1}) = \mathcal{T}_{H-1}(o_H|o_{1:H-1},a_{1:H-1})\cdots\mathcal{T}_1(o_2|o_1,a_1),\forall o_{1:H},a_{1:H-1}\}$, which is also frequently used in this paper. In fact, both $\mathcal{P}$ and $\mathcal{T}$ fully determine the dynamics with an additional constraint for $\mathcal{P}$, that is, $\sum_{o_{h+1:H}}\mathcal{P}(o_{1:H}|a_{1:H-1})$ is constant with respect to $a_{h:H-1}$ for all action sequence $a_{h:H-1}$. Hence, we use $\mathbb{D}=\{\mathcal{T}_h,\mathcal{P}\}$ to represent the model dynamics. To make the presentation more explicit, we introduce a unified parameterization $\theta\in\Theta\subset\mathbb{R}^D$ for the process dynamics, written as $\mathbb{D}_\theta=\{\mathcal{T}_h^\theta,\mathcal{P}^\theta\}$. Here, Θ is the complete set of model parameters and D is usually a large integer. We also use θ to refer to a model. With a little abuse of notation, we use $\mathbb{D}_\theta^\rho(\tau_h)$ to represent the marginal probability of the trajectory τ_h under the policy ρ. Mathematically, we have

$$\mathbb{D}_\theta^\rho(\tau_H)=\prod_{h'=1}^{h}\left(\mathcal{P}_h^\theta(o_{h'}|o_1,a_1,\ldots,o_{h'-1},a_{h'-1})\rho(a_{h'}|o_1,a_1\ldots,o_{h'})\right).$$

For a given reward function $R:(\mathcal{O}\times\mathcal{A})^H\to[0,1]$, and any policy π, the value of the policy π under the dynamics $\mathbb{D}_\theta$ is denoted by $V_{\theta,R}^\pi=\mathbb{E}_{\tau_H\sim\mathbb{D}_\theta^\pi}[R(o_{1:H},a_{1:H})]$. We remark here that the goal of an agent in standard RL is to find an ϵ-optimal policy $\hat{\pi}$ such that $\max_\pi V_{\theta,R}^\pi - V_{\theta,R}^{\hat{\pi}}\leq\epsilon$ for some small error $\epsilon>0$ under a single model θ.

B. Robust RL

While standard RL assumes that the system dynamics remains the same, robust RL diverges from this setting by focusing on the 'worst-case' or *robust value* in situations where the parameter θ can change within a defined uncertainty set $B\subset\mathbb{R}^D$ (where B will be defined below). Mathematically, given a policy π, the robust value under the uncertainty set B is defined by $V_{B,R}^\pi=\min_{\theta\in B}V_{\theta,R}^\pi$.

In this paper, we consider two types of uncertainty sets defined respectively on $\mathcal{T}$ and $\mathcal{P}$ via f-divergence $D_f(P,P')=\int f(dP/dP')dP'$, where f is a convex function for measuring the uncertainty. Given a *nominal* model θ^*, a $\mathcal{T}$-type uncertainty set under f-divergence with center θ^* and radius ξ is defined as

$$B_{\mathcal{T}}^f(\theta^*):=\left\{\theta: D_f\left(\mathcal{T}_h^\theta(\cdot|\tau_h),\mathcal{T}_h^{\theta^*}(\cdot|\tau_h)\right)\leq\xi,\forall h,\tau_h\right\}. \quad (1)$$

Similarly, a $\mathcal{P}$-type uncertainty set under f-divergence with center θ^* and radius ξ is defined as

$$B_{\mathcal{P}}^f(\theta^*):=\left\{\theta: D_f\left(\mathcal{P}^\theta(\cdot|a_{1:H}),\mathcal{P}^{\theta^*}(\cdot|a_{1:H})\right)\leq\xi,\forall a_{1:H}\right\}. \quad (2)$$

For simplicity, we set ξ to be a constant throughout the paper. Our proposed algorithm and analysis can be easily extended to the case when ξ is a function of τ_h or $a_{1:H}$. Notably, the $\mathcal{T}$-type uncertainty set reduces to the $\mathcal{S}\times\mathcal{A}$-rectangular uncertainty set [9] if $\mathbb{D}$ reduces to a Markov decision process, i.e., $\mathcal{T}_h(\cdot|\tau_h)=\mathcal{T}_h(\cdot|o_h,a_h)$. However, due to the non-Markovian nature of our problem, we also consider the $\mathcal{P}$-type uncertainty set as $\mathcal{P}$ also fully determines the process.

We study two example types of f-divergence, which corresponds to two classical divergences. (i) If $f_1(t)=|t-1|/2$, then D_{f_1} is called *total variation distance*; (ii) If $f_2(t)=t\log t$, then D_{f_2} is called *Kullback–Leibler (KL) divergence*. We also use B^i to represent B^{f_i} for simplicity.

We focus on **robust offline RL**, where the agent is provided with an offline dataset $\mathcal{D}=\{\tau_H^n\}_{n=1}^N$ that contains N sample trajectories collected via a *behavior* policy ρ under the nominal model θ^*. Given the type of the uncertain set $B\in\{B_{\mathcal{T}}^i,B_{\mathcal{P}}^i\}_{i=1}^2$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

(without knowing the center θ^*), the learning goal of the agent is to find a near-optimal robust policy $\hat{\pi}$ using this offline dataset $\mathcal{D}$ such that

$$\max_{\pi} \min_{\theta \in B} V^{\pi}_{\theta,R} - \min_{\theta \in B} V^{\hat{\pi}}_{\theta,R} \leq \epsilon.$$

C. Low-Rank Non-Markovian Decision Process

We first introduce some additional notations that are helpful to describe low-rank non-Markovian decision processes. Given a historical trajectory at time step h as $\tau_h := (o_{1:h}, a_{1:h})$, we denote a future trajectory as $\omega_h := (o_{h+1:H}, a_{h+1:H})$. The set of all τ_h is denoted by $\mathcal{H}_h = (\mathcal{O} \times \mathcal{A})^h$ and the set of all future trajectories is denoted by $\Omega_h = (\mathcal{O} \times \mathcal{A})^{H-h}$. In addition, we add a superscript o or a to either historical trajectory τ or a future trajectory to represent the observation sequence and the action sequence contained in that trajectory, respectively. For example, $\omega^o_h = o_{h+1:H}$ and $\omega^a_h = a_{h+1:H}$. To better describe a policy, we also use the notation $x_h = (\tau_{h-1}, o_h) = (o_{1:h}, a_{1:h-1})$ as the available information for the learning agent at time step h. We remark that, having τ_h, x_h in the same equation implies that $\tau_h = (x_h, a_h)$.

Recall that a non-Markov decision process θ can be fully determined by $\mathcal{P}^{\theta}$. We introduce a collection of dynamic matrices $\{\mathbb{M}^{\theta}_h \in \mathbb{R}^{|\mathcal{H}_h| \times |\Omega_h|}\}_{h=1}^{H}$. The entry at the τ_h-th row and ω_h-th column of $\mathbb{M}^{\theta}_h$ is $\mathcal{P}^{\theta}(\omega^o_h, \tau^o_h | \tau^a_h, \omega^a_h)$.

Definition 3.1 (Rank-r non-Markov decision process): A non-Markov decision process θ is rank r, if for any h, the model dynamic matrix $\mathbb{M}^{\theta}_h$ has rank r.

It has been proved that any low-rank sequential decision-making problem θ admits Predictive State Representations (PSR) [68] if for each h, there exists a set of future trajectories, namely, core tests known to the agent, $\mathcal{Q}_h = \{q_{h,1}, \ldots, q_{h,d}\} \subset \Omega_h$, such that the submatrix restricted to these tests (columns) $\mathbb{M}^{\theta}_h[\mathcal{Q}_h]$ has rank r, where $d \geq r$ is a positive integer.[1] We consider all low-rank problems with the same set of core tests $\{\mathcal{Q}_h\}_h^H$, and have the following representations.

$$\mathcal{P}^{\theta}(o_H, \ldots, o_1 | a_1, \ldots, a_H) = m(\omega_h)^{\top} \psi(\tau_h),$$
$$\phi_h = \sum_{o_H, \ldots, o_{h+1}} m(\omega_h).$$

where $m(\omega_h) \in \mathbb{R}^d$, and $\psi(\tau_h) \in \mathbb{R}^d$. We denote $\theta = (\phi_h, \psi(\tau_h))$ to be the representation. In particular, Θ is the set of all low-rank problems with the same core tests $\{\mathcal{Q}_h\}_{h=1}^H$. We finally remark that, $\phi_0 = \sum_{o_H, \ldots, o_1} m(\omega_0) = \sum_{o_H, \ldots, o_1} m(o_H, a_H, \ldots, o_1, a_1)$. Note that here $(a_H, \ldots, a_1)$ can arbitrary action sequence, because $\mathbf{m}$ is invariant against actions. We also note that

$\mathcal{P}^{\theta}(o_H, \ldots, o_1 | a_1, \ldots, a_H) = m(\omega_0)^{\top} \psi_0$.

We assume ψ_0 is known to the agent [21], and employ a *bar* over a feature $\psi(\tau_h)$ to denote its normalization $\bar{\psi}(\tau_h) = \psi(\tau_h)/\phi_h^{\top}\psi(\tau_h)$. $\bar{\psi}(\tau_h)$ is also known as the prediction vector [71] or prediction feature of τ_h, since $[\bar{\psi}(\tau_h)]_{\ell}$ is the probability of observing $q^{\mathrm{o}}_{h,\ell}$ given the history τ_h and taking action sequence $q^{\mathrm{a}}_{h,\ell}$.

Moreover, let $\mathcal{Q}^A_h = \{q^{\mathrm{a}}_{h,\ell}\}_{\ell=1}^d$ be the set of action sequences that are part of core tests, constructed by eliminating any repeated action sequence. $\mathcal{Q}^A_h$ known as the set of core action sequences.

We further assume that PSRs studied in this paper are well-conditioned, as specified in Assumption 3.2. Such an assumption and its variants are commonly adopted in the study of PSRs [21], [68], [69], [70].

Assumption 3.2 (γ-Well-Conditioned PSR): A PSR θ is said to be γ-well-conditioned if

$$\forall h, \quad \max_{x \in \mathbb{R}^{d_h} : \|x\|_1 \leq 1} \max_{\pi, \tau_h} \sum_{\omega_h} \pi(\omega_h | \tau_h) |$$
$$m(\omega_h)^{\top} x| \leq \frac{1}{\gamma},$$

where $\pi(\omega_h|\tau_h) = \pi(a_H|x_H) \cdots \pi(a_{h+1}|x_{h+1})$.

Assumption 3.2 requires that the error of estimating θ does not significantly blow up when the estimation error x of estimating the probability of core tests is small.

IV. Robust Offline RL With Low-Rank Nominal Model

In this section, we study the setting when the nominal model has *low-rank* structure (with rank r) [21], [66], [68], [69]. To unify the notations, we use $B^i_{\mathrm{u}}(\theta)$ to denote the uncertainty set, where u is either symbol $\mathcal{T}$ or $\mathcal{P}$.

Algorithm 1 Robust Offline RL With Low-Rank Nominal Model

1: **Input:** offline dataset $\mathcal{D}$, divergence type i
2: Estimate a model $\hat{\theta}$ according to Equation (3).
3: Obtain distilled dataset $\mathcal{D}^{\mathrm{g}}$ according to Equation (4).
4: Construct $\hat{b}$ according to Equation (5).
5: Output $\hat{\pi}$ such that $\hat{\pi} = \arg\max_{\pi \in \Pi} \left(V^{\pi}_{B^i_{\mathrm{u}}(\hat{\theta}),R} - C^i_{\mathrm{u}} V^{\pi}_{\hat{\theta},\hat{b}} \right)$.

A. Algorithm Design

We present our algorithm for robust offline RL when the nominal model enjoys low-rank structure. Our algorithm features **two main novel elements:** (a) dataset distillation, which allows the estimated prediction features of distilled data being close to the true features, and thus enables us to construct (b) a lower confidence bound design for the robust value $V^{\pi}_{B^i_{\mathrm{u}}(\theta^*),R}$. In particular, we develop a new dual form for the robust value $V^{\pi}_{B^i_{\mathcal{P}}(\theta),R}$ under $\mathcal{P}$-type uncertainty set, that is more amenable to practical implementation. The main steps of the algorithm are explained in detail as follows, and the pseudo-code is presented in Algorithm 1.

Model estimation and dataset reform. First, we perform maximum likelihood estimation of the nominal model θ^* by minimizing the following negative log-likelihood loss function

$$\mathcal{L}(\theta|\mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} - \log \mathbb{D}^o_{\theta}(\tau^n_H), \tag{3}$$

which can be done using a supervised learning oracle, e.g. stochastic gradient descent. Let $\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta|\mathcal{D})$.

Novel dataset distillation. Motivated by [21], to construct a desirable lower confidence bound for any value function under the nominal model, the estimated model $\hat{\theta}$ needs to

[1] Setting the same d for different h is for notation simplicity.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

assign non-negligible probabilities on every sample trajectory. Differently from [21] that introduces additional constraints to the optimization oracle, we propose a novel idea that only keeps the "good" sample trajectories whose estimated probability is greater than a pre-defined small value $p_{\min}$ under the estimated model $\hat{\theta}$ and collect those sample trajectories in set $\mathcal{D}^{\mathrm{g}}$, defined as follows.

$$\mathcal{D}^{\mathrm{g}} = \left\{\tau_H \in \mathcal{D}, \quad \mathbb{D}^{\rho}_{\hat{\theta}}(\tau_H) \geq p_{\min}\right\}. \tag{4}$$

We prove in Lemma 2.1 that, with high probability $|\mathcal{D}^{\mathrm{g}}| = \Omega(N)$, which still guarantees sufficient number of samples for learning. Then, $\mathcal{D}^{\mathrm{g}}$ is randomly and evenly divided into H datasets $\mathcal{D}^{\mathrm{g}}_0, \ldots, \mathcal{D}^{\mathrm{g}}_{H-1}$, in order to separately learn $\bar{\psi}^*(\tau_h)$ for each $h \in \{0, 1, \ldots, H-1\}$.

LCB of robust value. Given the estimated model $\hat{\theta}$ and divergence type i, Algorithm 1 constructs a lower confidence bound (LCB) of $V^{\pi}_{B^i_{\mathrm{u}}(\theta^*),R}$ for all policy π through a bonus function $\hat{b}$ defined as:

$$\hat{U}_h = \lambda I + \sum_{\tau_h \in \mathcal{D}^{\mathrm{g}}_h} \bar{\hat{\psi}}(\tau_h)\bar{\hat{\psi}}(\tau_h)^{\top},$$
$$\hat{b}(\tau_H) = \min\left\{\alpha \sqrt{\sum_h \left\|\bar{\hat{\psi}}(\tau_h)\right\|^2_{(\hat{U}_h)^{-1}}}, 1\right\}, \tag{5}$$

where λ and α_i are pre-defined parameters. Recall that $\bar{\hat{\psi}}(\tau_h)$ is the prediction feature described in Section III-C. Then, we can show that $V^{\pi}_{B^i_{\mathrm{u}}(\hat{\theta}),R} - C^i_{\mathrm{u}} V^{\pi}_{\hat{\theta},\hat{b}}$ is an LCB of $V^{\pi}_{B^i_{\mathrm{u}}(\theta^*),R}$, where C^i_{u} (defined in Section IV-B) is a *scaling parameter* that depends on the type of uncertainty set.

Novel dual form of robust value. For the $\mathcal{T}$-type uncertainty sets, define robust value function as:

$$V^{\pi}_{B(\theta^*),R,h}(x_h) = \inf_{\theta \in B(\theta^*)} \mathbb{E}^{\pi}_{\theta}\left[\sum_{h'=h}^{H} R(\boldsymbol{\tau}_{h'}) \middle| x_h\right],$$
$$Q^{\pi}_{B(\theta^*),R,h}(\tau_h) = \inf_{\theta \in B(\theta^*)} \mathbb{E}^{\pi}_{\theta}\left[\sum_{h'=h}^{H} R(\boldsymbol{\tau}_{h'}) \middle| \tau_h\right],$$

where $\tau_h = (x_h, a_h)$. Then, the following Bellman-type equations hold for robust value functions.

Lemma 4.1: For any θ, let $V^{\pi}_{B(\theta),R,H+1}(\tau_H) = 0$. Then, we have

$$V^{\pi}_{B(\theta),R,h}(x_h) = \sum_{a_h} \pi(a_h|x_h) Q^{\pi}_{B(\theta),R,h}(\tau_h)$$
$$Q^{\pi}_{B(\theta),R,h}(\tau_h) = R(x_h) + \inf_{\theta' \in B(\theta^*)} \sum_{o_{h+1}} \mathcal{T}^{\theta'}_h(o_{h+1}|\tau_h) V^{\pi}_{B(\theta),R,h+1}(x_{h+1}).$$

Therefore, robust value under $\mathcal{T}$-type uncertainty set can be calculated using the recursive formula above. Notably, the calculation of $Q^{\pi}_{B(\theta),R,h}(\tau_h)$ involves solving an optimization problem in the form of $\inf_{P: D_f(P,P^*)} \mathbb{E}_P[g(X)]$, where P and P^* are two probability distributions over a set $\mathcal{X}$. This optimization problem can be solved efficiently as shown in the literature [16], [17]. We also provide its dual form in Section A.

On the other hand, when the uncertainty set is $\mathcal{P}$-type, we develop a new dual form of the robust value $V^{\pi}_{B^1_{\mathcal{P}}(\theta^*),R}$ and $V^{\pi}_{B^2_{\mathcal{P}}(\theta^*),R}$ in Equation (6) and Equation (7), respectively. Note that, in those equations, $f(x_H) = \sum_{a_H} R(\tau_H)\pi(\tau_{H-1})$.

$$V^{\pi}_{B^1_{\mathcal{P}}(\theta^*),R} = \max_{\boldsymbol{\gamma} \geq 0, \boldsymbol{\lambda}} \Bigg\{ \sum_{x_H} (f(x_H) - \gamma_{x_H}) \mathcal{P}^{\theta^*}(x_H) - \sum_{a_{1:H-1}} \max_{o_{1:H}} \left|\gamma_{x_H} - f(x_H) - \lambda_{a_{H-1},x_{H-1}}\right| \xi \Bigg\} \tag{6}$$

$$V^{\pi}_{B^2_{\mathcal{P}}(\theta^*),R} = \max_{\boldsymbol{\eta} \geq 0} \Bigg\{ -\sum_{\tau_{H-1}} \eta_{a_{1:H-1}} \mathcal{P}^{\theta^*}(\tau_{H-1}) \log \mathbb{E}_{o_H \sim \mathcal{T}^{\theta^*}_{H-1}} \left[e^{-\frac{f(x_H)}{\eta_{a_{1:H-1}}}}\right] - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}} \xi \Bigg\} \tag{7}$$

We remark that although solving the dual forms exactly may be computationally intensive, in practice we utilize approximate solutions that are sufficient for ensuring theoretical guarantees. For example, we could use gradient-based method to solve dual forms, where we update the dual parameters $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ through one-step gradient at each iteration.

Output policy design. Finally, Algorithm 1 outputs a policy $\hat{\pi} = \arg\max_{\pi \in \Pi} \left(V^{\pi}_{B^i_{\mathrm{u}}(\hat{\theta}),R} - C^i_{\mathrm{u}} V^{\pi}_{\hat{\theta},\hat{b}}\right)$, which maximizes a *lower confidence bound* of $V^{\pi}_{B^i_{\mathrm{u}}(\theta^*),R}$. An illustration of our algorithm in provided in Figure 1.

B. Theoretical Guarantee

In this section, we present the theoretical results for Algorithm 1. Before we state the theorem, we need to introduce several definitions and condition coefficients.

Bracketing number. First, we introduce a complexity measure of the model class Θ, which is standard in the literature [21], [68].

Definition 4.2 (ε-Bracketing Number of Θ): Suppose for each $\theta \in \Theta$, there exists a function $F_\theta : \mathcal{X} \to \mathbb{R}_+$ parameterized by θ. Given two functions l and $g : \mathcal{X} \to \mathbb{R}_+$, the bracket $[l, g]$ is the set of all $\theta \in \Theta$ satisfying $l \leq F_\theta \leq g$. An ε-bracket is a bracket $[l, g]$ with $\|g - l\|_1 < \varepsilon$. The bracketing number $\mathcal{N}_\varepsilon(\Theta)$ is the minimum number of η-brackets needed to cover Θ.

Note that if Θ is the parameter space for tabular PSRs with rank r, we have $\log \mathcal{N}_\varepsilon(\Theta) \leq r^2|\mathcal{O}||\mathcal{A}|H^2 \log \frac{H|\mathcal{O}||\mathcal{A}|}{\varepsilon}$ (see Theorem 4.7 in [68]).

Novel concentrability coefficient. Under the low-rank structure and PSRs, we introduce the following novel concentrability coefficient $C_{\mathrm{p}}(\pi|\rho)$, namely **Type-I concentrability coefficient**, of a behavior policy ρ against a robust policy π.

$$C_{\mathrm{p}}(\pi|\rho) = \max_h \max_x \frac{x^{\top} \mathbb{E}_{\tau_h \sim \mathbb{D}^{\pi}_{\theta^*}} [\bar{\psi}^*(\tau_h)\bar{\psi}^*(\tau_h)^{\top}] x}{x^{\top} \mathbb{E}_{\tau_h \sim \mathbb{D}^{\rho}_{\theta^*}} [\bar{\psi}^*(\tau_h)\bar{\psi}^*(\tau_h)^{\top}] x}.$$

Remark 4.3: Requiring a finite Type-I concentrability coefficient is much weaker than requiring a finite concentrability coefficient C defined in [21], in which a point-to-point probability ratio is used, i.e., $C \geq \mathbb{D}^{\pi}_{\theta^*}(\tau_h)/\mathbb{D}^{\rho}_{\theta^*}(\tau_h)$. It is obvious that $C_{\mathrm{p}}(\pi|\rho) \leq C$. Moreover, even if $\mathbb{D}^{\rho}_{\theta^*}(\tau_h)$ is exponentially small, indicating C is exponentially large, the minimum eigenvalue of the covariance matrix of $\bar{\psi}^*(\tau_h)$ under ρ (i.e., $\mathbb{E}_{\tau_h \sim \mathbb{D}^{\rho}_{\theta^*}}[\bar{\psi}^*(\tau_h)\bar{\psi}^*(\tau_h)^{\top}]$) could be bounded way from zero. By Assumption 3.2, we have $\|\bar{\psi}^*(\tau_h)\|_1/\gamma \geq \sum_{\omega_h} \pi(\omega_h|\tau_h) m(\omega_h)^{\top} \bar{\psi}^*(\tau_h) = 1$, which implies that the inverse

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

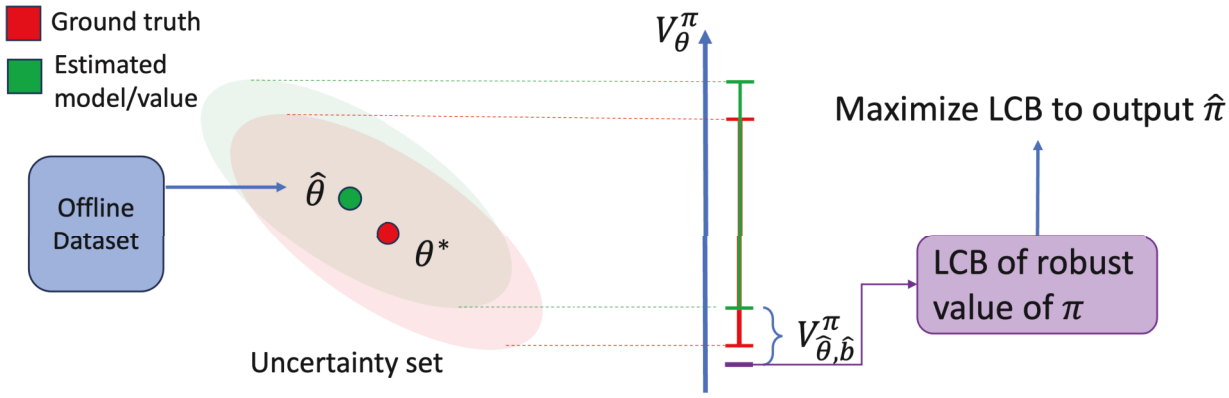

Fig. 1. Illustration of Algorithm 1.

of the norm of the prediction feature scales in $\mathrm{Poly}(d, 1/\gamma)$. Thus, with a good behavior policy ρ, the scaling of $C_{\mathrm{p}}(\pi|\rho)$ is likely to be polynomial.

Wellness condition of the nominal model. When the uncertainty set is $\mathcal{T}$-type, we further introduce a wellness condition number C_B of the nominal model θ^* defined as follows.

$$C_B = \max_{\theta\in B(\theta^*)} \max_h \frac{\mathbb{D}_\theta^\pi(\tau_h)}{\mathbb{D}_{\theta^*}^\pi(\tau_h)} = \max_{\theta\in B(\theta^*)} \max_h \frac{\mathcal{P}^\theta(\tau_h)}{\mathcal{P}^{\theta^*}(\tau_h)}.$$

Intuitively, this condition number is called the maximum likelihood ratio between the distributions $\mathcal{P}^{\theta^*}(\cdot|a_{1:h})$ and $\mathcal{P}^\theta(\cdot|a_{1:h})$ and only depends on the property of the nominal model and the shape of the uncertainty set. In particular, when $B = B^2_{\mathcal{T}}$, we must have

$$\mathbb{E}_\theta^\pi\left[\log\frac{\mathcal{P}^\theta(\tau_h)}{\mathcal{P}^{\theta^*}(\tau_h)}\right] = \mathbb{E}_\theta^\pi\left[\sum_{h'\le h}\log\frac{\mathcal{T}^\theta(o_{h'}|\tau_{h'-1})}{\mathcal{T}^{\theta^*}(o_{h'}|\tau_{h'-1})}\right] \le \xi h,$$

which should hold for all policy π, and thus implies that C_B is highly likely to be small.

Scaling parameters. Finally, we briefly introduce the scaling parameter C_{u}^i that is used to construct the LCB of robust value. Equipped with the dual forms, we define $C_{\mathcal{P}}^1 = 1$, $C_{\mathcal{T}}^1 = C_B$. When uncertainty set is determined by KL divergence, we make the following assumption.

Assumption 4.4: Suppose that $\boldsymbol{\eta}^* \in \mathbb{R}^{|\mathcal{A}|^{H-1}}$ is an optimal solution of Equation (7), and $\boldsymbol{\eta}^*_{\hat{\theta}} \in \mathbb{R}^{|\mathcal{A}|^{H-1}}$ is an optimal solution of Equation (7) when θ^* is replaced by $\hat{\theta}$. We assume that the agent has an estimate of a lower bound η^* of the coordinates of both $\boldsymbol{\eta}^*$ and $\boldsymbol{\eta}^*_{\hat{\theta}}$, i.e., $\eta^* \le \min_{a_{1:H-1}}\{\eta^*_{a_{1:H-1}}, \eta^*_{\hat{\theta},a_{1:H-1}}\}$.

Assumption 4.5: $\lambda^*_{\theta\|\theta^*}(\tau_h)$ is an optimal solution to the following optimization problem:

$$\sup_{\lambda\ge 0}\left\{-\lambda\log\mathbb{E}_{\mathcal{T}_h^\theta}[\exp(-V^\pi_{B^2_{\mathcal{T}}(\theta^*),R,h+1}(\tau_h,o)/\lambda)] - \lambda\xi\right\}.$$

We assume that the agent has an estimate of a lower bound λ^* of all $\lambda^*_{\theta^*\|\hat{\theta}}(\tau_h)$ and $\lambda^*_{\hat{\theta}\|\theta^*}(\tau_h)$, i.e., $\lambda^* \le \min_{h,\tau_h}\{\lambda^*_{\theta^*\|\hat{\theta}}(\tau_h), \lambda^*_{\hat{\theta}\|\theta^*}(\tau_h)\}$.

Remark 4.6: The optimization problem mentioned in Assumption 4.5 is the dual problem that need to solve in the Bellman-type equation (Lemma 4.1). In the algorithm implementation, it suffices to estimate $\min_{a_{1:H-1}}\{\eta^*_{\hat{\theta},a_{1:H-1}}\}$ and $\min_{h,\tau_h}\{\lambda^*_{\hat{\theta}\|\theta^*}(\tau_h)\}$, which is not difficult to obtain such value since $\hat{\theta}$ is known. We make these two assumptions for notation simplicity. Such assumptions also appear in [17].

Define the scaling parameters $C^2_{\mathcal{P}} = 3\exp(1/\eta^*)$, and $C^2_{\mathcal{T}} = C_B\max\{\exp(\xi)/\xi, \lambda^*\exp(1/\lambda^*)\}$. We are now ready to present the theorem characterizing the performance of Algorithm 1.

Theorem 4.7: Suppose Assumption 3.2, Assumption 4.4, and Assumption 4.5 hold. Let $Q_A = \max_h|\mathcal{Q}_h^A|$ be the maximum number of core action sequences. Let $\iota = \min_{q^{\mathrm{a}}_{h,\ell},\tau_h}\rho(q^{\mathrm{a}}_{h,\ell}|\tau_h)$, $p_{\min} = \frac{\delta}{N(|\mathcal{O}||\mathcal{A}|)^{2H}}$, $\varepsilon = \frac{p_{\min}}{NH}$, $\beta = O(\log(\mathcal{N}_\varepsilon(\Theta)/\delta))$, $\lambda = H^2Q_A^2$, and $\alpha = O\left(\frac{Q_A\sqrt{dH}}{\gamma^2}\sqrt{\lambda} + \frac{\sqrt{\beta}}{\iota\gamma}\right)$. Then, with probability at least $1-\delta$, we have the following four results for the suboptimal gap of the policy $\hat{\pi}$ output by Algorithm 1 under four settings.

(i) $\mathcal{P}$-type uncertainty set under TV distance: $V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\theta^*),R} \lesssim C_0\sqrt{\frac{C_{\mathrm{p}}(\pi^*|\rho)}{N}}$,

(ii) $\mathcal{P}$-type uncertainty set under KL divergence: $V^{\pi^*}_{B^2_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^2_{\mathcal{P}}(\theta^*),R} \lesssim C_0C^{\mathrm{kl}}_{\mathcal{P}}\sqrt{\frac{C_{\mathrm{p}}(\pi^*|\rho)}{N}}$,

(iii) $\mathcal{T}$-type uncertainty set under TV distance: $V^{\pi^*}_{B^1_{\mathcal{T}}(\theta^*),R} - V^{\hat{\pi}}_{B^1_{\mathcal{T}}(\theta^*),R} \lesssim C_0C_B\sqrt{\frac{C_{\mathrm{p}}(\pi^*|\rho)}{N}}$,

(iv) $\mathcal{T}$-type uncertainty set under KL divergence, $V^{\pi^*}_{B^2_{\mathcal{T}}(\theta^*),R} - V^{\hat{\pi}}_{B^2_{\mathcal{T}}(\theta^*),R} \lesssim C_0C_BC^{\mathrm{kl}}_{\mathcal{T}}\sqrt{\frac{C_{\mathrm{p}}(\pi^*|\rho)}{N}}$,

where π^* is the optimal robust policy under each setting, $C_{\mathrm{p}}(\pi^*|\rho)$ is the type-I coefficient of the behavior policy ρ against π^*, $C_0 = \frac{H^2Q_A\sqrt{rd\beta}}{\iota\gamma^2}\left(\sqrt{r} + \frac{Q_A\sqrt{H}}{\gamma}\right)$, C_B is the wellness condition number of θ^*, $C^{\mathrm{kl}}_{\mathcal{P}} = 3\exp(1/\eta^*)$, $C^{\mathrm{kl}}_{\mathcal{T}} = \max\{\exp(\xi)/\xi, \lambda^*\exp(1/\lambda^*)\}$, and η^*, λ^* are defined in Assumption 4.4 and Assumption 4.5.

The theorem states that as long as the type-I concentrability coefficient of ρ against the optimal robust policy $\pi^* = \arg\max_\pi V^\pi_{B(\theta),R}$ is bounded, and the nominal model θ^* has finite wellness condition number C_B, Algorithm 1 finds

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

an ϵ-optimal robust policy provided that the offline dataset has $\Omega(1/\epsilon^2)$ sample trajectories.

Proof sketch of Theorem 4.7. The proof consists of three main steps. (i) **A novel MLE guarantee.** We show that the estimator $\hat{\theta}$ produced by *vanilla* MLE already performs well on a large portion of offline data samples, namely, $\sum_{\tau_h \in \mathcal{D}^{\mathrm{g}}_h} \|\mathcal{D}^{\rho}_{\hat{\theta}}(\cdot|\tau_h) - \mathcal{D}^{\rho}_{\theta^*}(\cdot|\tau_h)\|_1 \le O(\beta)$. This result is different from the MLE guarantee proved in [21], where a constrained MLE oracle is required. (ii) **Simulation lemma for robust value.** By utilizing the newly proposed dual forms in Equation (6) and Equation (7), we show that for any model θ (not necessarily low-rank), the difference $|V^{\pi}_{B(\theta^*),R} - V^{\pi}_{B(\theta),R}|$ is generally upper bounded by $O(c_B \|\mathbb{D}^{\pi}_{\theta^*} - \mathbb{D}^{\pi}_{\theta}\|_1)$, where c_B depends on the uncertainty type. (iii) **Applying concentrability coefficient.** In low-rank non-Markovian decision processes, ℓ_1 distance $\|\mathbb{D}^{\pi}_{\theta^*} - \mathbb{D}^{\pi}_{\theta}\|_1$ can be upper bounded by $\mathbb{E}^{\pi}_{\theta^*}[\|\bar{\psi}^*(\tau_h)\|_{\Lambda^{-1}}]$ [21], [68]. By Cauchy's inequality, this quantity is determined by the covariance matrix $\mathbb{E}^{\pi}_{\theta^*}[\bar{\psi}(\tau_h)\bar{\psi}^*(\tau_h)^{\top}]$, which can thus be controlled by the type-I concentrability coefficient $C_{\mathrm{p}}(\pi|\rho)$ and the covariance matrix under the behavior policy ρ. It then yields the final result when combined with the MLE guarantee. The full proof can be found in Section C.

Remark 4.8 (Comparison and Computational Considerations): Our work is the first to address robust offline learning in PSRs and general non-Markovian decision processes. Consequently, the assumptions on data coverage and model structure are fundamentally different and not directly comparable to those in prior MDP-based settings or robust POMDPs with known uncertainty sets. In terms of computational efficiency, we develop a tractable algorithm under the assumption that the nominal model possesses a low-rank structure, even when the model itself is unknown. In contrast, existing approaches typically alleviate computational difficulties by imposing relaxed conditions on either the model or the uncertainty set.

Regarding practical deployment of our algorithm, the model can be estimated using standard stochastic gradient descent (SGD) by minimizing the loss in Equation (3), followed by an additional training epoch for dataset distillation. For robust value estimation, we can apply another gradient-based optimization procedure over the dual variables η, γ, and λ. Thus, all major components of our approach can be efficiently implemented with standard optimization techniques, making it well-suited for practical scenarios.

V. Robust Offline RL Under General Non-Markovian Decision Processes

In this section, we first introduce a more general algorithm (see Algorithm 2) for robust offline learning in non-Markovian decision processes, and then present its theoretical guarantee.

A. Algorithm Design

The general algorithm utilizes double pessimism for robust offline RL. The algorithm consists of two parts: *model estimation* and *robust policy planning*. We note that similar double pessimism has been adopted in [17]. However, due to the non-Markovian nature, our robust value and its dual form as well as the corresponding analysis are different from theirs.

Model estimation. We construct the confidence set $\mathcal{C}$ for nominal model θ^* using offline dataset $\mathcal{D}$.

$$\mathcal{C} = \left\{\theta : \textstyle\sum_{n=1}^{N} \mathcal{P}^{\theta}(\tau_H^n) \ge \max_{\theta} \sum_{n=1}^{N} \mathcal{P}^{\theta}(\tau_H^n) - \beta\right\}, \quad (8)$$

where β is a pre-defined parameter. This confidence set ensures that $\theta^* \in \mathcal{C}$ and any model $\theta \in \mathcal{C}$ is "close" to the nominal model θ^* under the behavior policy ρ.

Robust policy planning. Then, we output a policy under the double pessimism principle. Specifically, for each model $\hat{\theta} \in \mathcal{C}$, since it is close to the nominal model θ^*, we examine its robust value function $V^{\pi}_{B(\hat{\theta}),R}$ under a policy π. By minimizing the robust value $V^{\pi}_{B(\hat{\theta}),R}$ over the confidence set $\mathcal{C}$, we find a *lower confidence bound* for the true robust value $V^{\pi}_{B(\theta^*),R}$ due to $\theta^* \in \mathcal{C}$. Finally, we output an optimal policy $\hat{\pi}$ with respect to the LCB of $\theta^* \in \mathcal{C}$. Mathematically, $\hat{\pi} = \arg\max_{\pi \in \Pi} \min_{\hat{\theta} \in \mathcal{C}} V^{\pi}_{B(\hat{\theta}),R}$

B. Theoretical Guarantee

In this section, we present the theoretical guarantees of Algorithm 2. Since the nominal model does not have specific structure or representations, type-I concentrability coefficient is not suitable for this case. Therefore, we introduce the **Type-II concentrability coefficient** $C(\pi|\rho)$ of a behavior policy ρ against a robust policy π defined as $C(\pi|\rho) = \sum_{h=1}^{H} \mathbb{E}^{\rho}_{\theta^*}\left[\left(\frac{\mathbb{D}^{\pi}_{\theta^*}(\tau_h)}{\mathbb{D}^{\rho}_{\theta^*}(\tau_h)}\right)^2\right]$.

Remark 5.1: In general, requiring a finite type-II concentrability coefficient is stronger than requiring a finite type-I concentrability coefficient defined in Section IV-B. However, it is still a relaxed coverage assumption than that in [21]. This is because in the analysis, we rescale the estimation error across the distribution induced by the behavior policy instead of using a point-wise bound.

We are now ready to present the main result for Algorithm 2.

Theorem 5.2: Let $\varepsilon = \frac{1}{NH}$, $\beta = O(\log(\mathcal{N}_{\varepsilon}(\Theta)/\delta))$. Then, with probability at least $1-\delta$, we have the following four results for the suboptimal gap of the policy $\hat{\pi}$ output by Algorithm 2.

(i) $\mathcal{P}$-type uncertainty set under TV distance $V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\theta^*),R} \lesssim H\sqrt{\frac{C(\pi^*|\rho)\beta}{N}}$,

(ii) $\mathcal{P}$-type uncertainty set under KL divergence: $V^{\pi^*}_{B^2_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^2_{\mathcal{P}}(\theta^*),R} \lesssim HC^{\mathrm{kl}}_{\mathcal{P}}\sqrt{\frac{C(\pi^*|\rho)\beta}{N}}$,

(iii) $\mathcal{T}$-type uncertainty set under TV distance $V^{\pi^*}_{B^1_{\mathcal{T}}(\theta^*),R} - V^{\hat{\pi}}_{B^1_{\mathcal{T}}(\theta^*),R} \lesssim HC_B\sqrt{\frac{C(\pi^*|\rho)\beta}{N}}$,

(iv) $\mathcal{T}$-type uncertainty set under KL divergence $V^{\pi^*}_{B^2_{\mathcal{T}}(\theta^*),R} - V^{\hat{\pi}}_{B^2_{\mathcal{T}}(\theta^*),R} \lesssim HC_BC^{\mathrm{kl}}_{\mathcal{P}}\sqrt{\frac{C(\pi^*|\rho)\beta}{N}}$,

where π^* is the optimal robust policy under each setting, $C(\pi^*|\rho)$ is the type-II coefficient of the behavior policy ρ against π^*, C_B is the wellness condition number defined in Section IV-B, and $C^{\mathrm{kl}}_{\mathcal{P}}, C^{\mathrm{kl}}_{\mathcal{T}}$ are defined in Theorem 4.7.

The theorem suggests that Algorithm 2 is more sample efficient than Algorithm 1, due to the general double pessimism design and the type-II concentrability coefficient, while Algorithm 1 is more amenable to practical implementation due to low-rank structure and computationally tractable bonus function design and dual forms development.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

VI. Conclusion

In this work, we studied robust offline non-Markovian RL with unkown transitions. We developed the first sample efficient algorithm when the nominal model admits a low-rank structure. To characterize the theoretical guarantee of our algorithm, we proposed a novel type-I concentrability coefficient for PSRs, which can be of independent interest in the study of offline low-rank non-Markovian RL. In addition, we extended our algorithm to a more general case when the nominal model does not have any specific structure. With the notion of type-II concentrability coefficient, we proved that the extended algorithm also enjoys sample efficiency under all different types of the uncertainty set.

Appendix A
Technical Lemmas

In this section, we present the technical lemmas for proving our main results. It includes three subsections. First, Section A-A primarily describes the dual forms for robust values under $\mathcal{P}$-type uncertainty sets. Second, Section A-B verifies that robust value functions satisfies Bellman-type equations. Finally, Section A-C provides various simulation lemmas for robust values under different uncertainty sets.

A. Dual Form of Robust Values

We first present two classical lemmas for total variation distance and KL divergence, which can be found in [16] and [17].

Proposition 1.1: Let D_{f_1} be defined with the convex function $f_1(t) = |t-1|/2$ corresponding to the total variation distance. Then,

$$\inf_{D_{f_1}(P\|P_o)\le\xi} \mathbb{E}_P[\ell(X)] = \max_{\lambda\in\mathbb{R}} \Big\{\lambda - \mathbb{E}_{P_o}[(\lambda-\ell(X))_+] - \xi \max_x (\lambda-\ell(x))_+\Big\}.$$

Proposition 1.2: Let D_{f_2} be defined with the convex function $f_2(t) = t\log t$ corresponding to the KL-divergence. Then,

$$\inf_{D_{f_2}(P\|P_o)\le\xi} \mathbb{E}_P[l(X)] = \sup_{\lambda\ge 0} \left\{-\lambda \log \mathbb{E}_{P_o}\left[\exp(-\ell(X)/\lambda)\right] - \lambda\xi\right\}.$$

Then, we show the dual form for the newly proposed $\mathcal{P}$-type uncertainty sets.

Proposition 1.3: If $B(\theta^*) = B^1_{\mathcal{P}}(\theta^*)$, then, we have

$$V^\pi_{B(\theta^*),R} = \max_{\gamma\ge 0,\lambda} \Big\{ \sum_{x_H} (f(x_H) - \gamma_{x_H})\mathcal{P}^*(x_H) - \sum_{a_{1:H-1}} \max_{o_{1:H}} \left|\gamma_{x_H} - f(x_H) - \lambda_{(a_{H-1},x_{H-1})}\right| \xi\Big\},$$

where $f(x_H) = R(x_H)\pi(\tau_{H-1})$, and the maximum must be achieved when

$$-f(x_H) \le \lambda_{(a_{H-1},x_{H-1})} \le 0,$$
$$0 \le \gamma_{x_H} \le \max\left\{f(x_H) + \lambda_{a_{H-1},x_{H-1}}, 0\right\} \le f(x_H).$$

Proof: We recall the definition of $V^\pi_{B^1_{\mathcal{P}}(\theta^*),R}$:

$$V^\pi_{B^1_{\mathcal{P}}(\theta^*),R} = \min_{\theta\in B^1_{\mathcal{P}}(\theta^*)} \sum_{x_H} \mathcal{P}^\theta(x_H)\pi(\tau_{H-1})R(x_H)$$

To simplify the notation, we denote $\pi(\tau_{H-1})R(x_H)$ by $f(x_H)$. Then, we need to specify the equations that form $B^1_{\mathcal{P}}(\theta^*)$.

The most important one is that for any $a_{1:H-1}$ $\sum_{o_2,\ldots,o_H} |\mathcal{P}(o_{2:H}|a_{1:H-1}) - \mathcal{P}^*(o_{2:H}|a_{1:H-1})| \le \xi$. Another implicit constraint is that if we sum over all o_H, the value $\sum_{o_H}\mathcal{P}(o_H,\ldots,o_2|a_1,\ldots,a_{H-1})$ is constant with respect to a_{H-1}. Therefore, we construct some additional variables $\mathcal{P}(x_h)$ and $s(x_H) \ge |\mathcal{P}(x_H) - \mathcal{P}^*(x_H)|$ and formulate the following constrained optimization problem.

$$\text{P1:}\begin{cases} \min \sum_{x_H} \mathcal{P}(x_H)f(x_H) \\ \text{s.t.} \sum_{o_2} \mathcal{P}(x_2) = 1, \quad \forall(a_1,x_1) \\ \sum_{o_3} \mathcal{P}(x_3) = \mathcal{P}(x_2), \quad \forall(a_2,x_2) \\ \cdots \\ \sum_{o_h} \mathcal{P}(x_h) = \mathcal{P}(x_{h-1}), \quad \forall(a_{h-1},x_{h-1}) \\ \cdots \\ \sum_{o_H} \mathcal{P}(x_H) = \mathcal{P}(x_{H-1}), \quad \forall(a_{H-1},x_{H-1}) \\ -s(x_H) \le \mathcal{P}(x_H) - \mathcal{P}^*(x_H) \le s(x_H), \quad \forall x_H \\ \sum_{o_{2:H}} s(x_H) \le \xi, \quad \forall a_{1:H-1} \\ \mathcal{P}(x_H) \ge 0, s(x_H) \ge 0, \quad \forall x_H \end{cases}$$

Since P1 is a linear programming, we can formulate the Lagrangian function $L_1(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta})$ as follows:

$$\begin{aligned} &L_1(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\eta}) \\ &= \sum_{x_H}\mathcal{P}(x_H)f(x_H) + \sum_{(a_1,x_1)}\lambda_{(a_1,x_1)}\left(\sum_{o_2}\mathcal{P}(x_2) - 1\right) \\ &\quad + \sum_{h=2}^{H-1}\sum_{(a_h,x_h)}\lambda_{(a_h,x_h)}\left(\sum_{o_{h+1}}\mathcal{P}(x_{h+1}) - \mathcal{P}(x_h)\right) \\ &\quad + \sum_{x_H}\alpha_{x_H}(\mathcal{P}(x_H) - \mathcal{P}^*(x_H) - s(x_H)) \\ &\quad + \sum_{x_H}\beta_{x_H}(-\mathcal{P}(x_H) + \mathcal{P}^*(x_H) - s(x_H)) \\ &\quad - \sum_{x_H}(\gamma_{x_H}\mathcal{P}(x_H) + \delta_{x_H}s(x_H)) + \sum_{a_{1:H-1}}\eta_{a_{1:H-1}} \\ &\quad \left(\sum_{o_{2:H}} s(x_H) - \xi\right), \end{aligned}$$

where $\lambda_{(a_h,x_h)} \in \mathbb{R}$, and $\alpha_{x_H},\beta_{x_H},\gamma_{x_H},\delta_{x_H},\eta_{a_{1:H-1}} \ge 0, \quad \forall x_H$.

Thus, the dual problem is

$$\text{D1:}\begin{cases} \max_{\lambda,\alpha,\beta,\gamma,\delta,\eta} \inf_{\mathcal{P},s} L_1(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\eta) \\ \text{s.t. } f(x_H) + \lambda_{(a_{H-1},x_{H-1})} + \alpha_{x_H} - \beta_{x_H} - \gamma_{x_H} = 0, \\ \qquad \forall x_H \\ \eta_{a_{1:H-1}} - \delta_{x_H} - \alpha_{x_H} - \beta_{x_H} = 0, \quad \forall x_H \\ \lambda_{(a_{h-1},x_{h-1})} = \sum_{a_h}\lambda_{(a_h,x_h)}, \quad \forall h, x_h \\ \alpha_{x_H},\beta_{x_H},\gamma_{x_H},\delta_{x_H},\eta_{a_{1:H-1}} \ge 0, \quad \forall x_H. \end{cases}$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

Denote $\inf_{\mathcal{P},s} L_1(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\eta)$ by $\bar{L}_1$. Then, we have

$$\bar{L}_1 = \sum_{x_H}(\beta_{x_H} - \alpha_{x_H})\mathcal{P}^*(x_H) - \lambda_{(a_1,x_1)} - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}}\xi.$$

In addition, by the third constraint in the dual problem D1, we have

$$\begin{aligned}\lambda_{(a_1,x_1)} &= \sum_{a_2}\lambda_{(a_2,x_2)}\\ &= \sum_{o_2}\mathcal{P}^*(x_2)\sum_{a_2}\lambda_{(a_2,x_2)}\\ &\cdots\\ &= \sum_{x_H}\mathcal{P}^*(x_H)\lambda_{(a_{H-1},x_{H-1})}.\end{aligned}$$

Thus, we can simplify $\bar{L}_1$ to

$$\begin{aligned}&\bar{L}_1(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\eta})\\ &= \sum_{x_H}(\beta_{x_H} - \alpha_{x_H})\mathcal{P}^*(x_H) - \lambda_{(a_1,x_1)} - \sum_{a_{1:H-1}}\eta_{a_{1:H-1}}\xi\\ &= \sum_{x_H}(\beta_{x_H} - \alpha_{x_H} - \lambda_{(a_{H-1},x_{H-1})})\mathcal{P}^*(x_H) - \sum_{a_{1:H-1}}\eta_{a_{1:H-1}}\xi\\ &= \sum_{x_H}(f(x_H) - \gamma_{x_H})\mathcal{P}^*(x_H) - \sum_{a_{1:H-1}}\eta_{a_{1:H-1}}\xi.\end{aligned}$$

Hence, by the Slater's condition, the primal value equals to the dual value. Mathematically, we have

$$\begin{aligned}&V^{\pi}_{B(\theta^*),R}\\ &= \max_{\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\eta}} \bar{L}_1(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\eta})\\ &= \max_{\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\eta}} \left\{\sum_{x_H}(f(x_H) - \gamma_{x_H})\mathcal{P}^*(x_H) - \sum_{a_{1:H-1}}\eta_{a_{1:H-1}}\xi\right\}\\ &\quad \text{s.t.} \begin{cases} \eta_{a_{1:H-1}} = \alpha_{x_H} + \beta_{x_H} + \delta_{x_H}\\ \gamma_{x_H} = f(x_H) + \lambda_{(a_{H-1},x_{H-1})} + \alpha_{x_H} - \beta_{x_H}\\ \eta_{a_{1:H-1}}, \gamma_{x_H}, \alpha_{x_H}, \beta_{x_H}, \delta_{x_H} \geq 0.\end{cases}\end{aligned}$$

If we fix $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$, then the maximum is achieved when

$$\begin{cases} \eta_{a_{1:H-1}} = \max_{o_{1:H}} \left|\gamma_{x_H} - f(x_H) - \lambda_{(a_{H-1},x_{H-1})}\right|\\ \alpha_{x_H} = \max\left\{\gamma_{x_H} - f(x_H) - \lambda_{(a_{H-1},x_{H-1})}, 0\right\}\\ \beta_{x_H} = \max\left\{-\gamma_{x_H} + f(x_H) + \lambda_{(a_{H-1},x_{H-1})}, 0\right\}\end{cases}$$

Therefore,

$$\begin{aligned}&V^{\pi}_{B(\theta^*),R}\\ &= \max_{\boldsymbol{\gamma}\geq 0,\boldsymbol{\lambda}} \Bigg\{\sum_{x_H}(f(x_H) - \gamma_{x_H})\mathcal{P}^*(x_H) - \sum_{a_{1:H-1}}\max_{o_{1:H}}\left|\gamma_{x_H}\right.\\ &\qquad \left. - f(x_H) - \lambda_{(a_{H-1},x_{H-1})}\right|\xi\Bigg\}\end{aligned}$$

Moreover, since $1 \geq f(x_H) \geq 0$, the maximum must be achieved when

$$\begin{aligned}&-f(x_H) \leq \lambda_{(a_{H-1},x_{H-1})} \leq 0,\\ &\quad 0 \leq \gamma_{x_H} \leq \max\left\{f(x_H) + \lambda_{a_{H-1},x_{H-1}}, 0\right\} \leq f(x_H).\end{aligned}$$

□

Proposition 1.4: If $B(\theta^*) = B^2_{\mathcal{P}}(\theta^*)$, then, we have

$$\begin{aligned}V^{\pi}_{B(\theta^*),R} = \max_{\boldsymbol{\eta}\geq 0}\Bigg\{&-\sum_{a_{H-1},x_{H-1}}\eta_{a_{1:H-1}}\mathcal{P}^{\theta^*}(x_{H-1})\\ &\log\mathbb{E}_{o\sim\mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta_{a_{1:H-1}}}\right)\right]\\ &-\sum_{a_{1:H-1}}\eta_{a_{1:H-1}}\xi\Bigg\}.\end{aligned}$$

Proof: Similar to Proposition 1.3, we denote $\pi(x_H)R(x_H)$ by $f(x_H)$. The robust value equals to the following optimization problem.

$$\text{P2:}\begin{cases}\min\sum_{x_H}\mathcal{P}(x_H)f(x_H)\\ \text{s.t.}\sum_{o_2}\mathcal{P}(x_2) = 1, \quad \forall(a_1,x_1)\\ \quad\sum_{o_3}\mathcal{P}(x_3) = \mathcal{P}(x_2), \quad \forall(a_2,x_2)\\ \quad\cdots\\ \quad\sum_{o_h}\mathcal{P}(x_h) = \mathcal{P}(x_{h-1}), \quad \forall(a_{h-1},x_{h-1})\\ \quad\cdots\\ \quad\sum_{o_H}\mathcal{P}(x_H) = \mathcal{P}(x_{H-1}), \quad \forall(a_{H-1},x_{H-1})\\ \quad\sum_{o_{2:H}}\mathcal{P}(x_H)\log\frac{\mathcal{P}(x_H)}{\mathcal{P}^*(x_H)} \leq \xi, \quad \forall a_{1:H-1}\\ \quad\mathcal{P}(x_H) \geq 0, \quad \forall x_H\end{cases}$$

Since P2 is a convex optimization problem, we can formulate the Lagrangian function $L_2(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\eta})$ as follows

$$\begin{aligned}&L_2(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\eta})\\ &= \sum_{x_H}\mathcal{P}(x_H)f(x_H) + \sum_{(a_1,x_1)}\lambda_{(a_1,x_1)}\left(\sum_{o_2}\mathcal{P}(x_2) - 1\right)\\ &\quad + \sum_{h=2}^{H-1}\sum_{(a_h,x_h)}\lambda_{(a_h,x_h)}\left(\sum_{o_{h+1}}\mathcal{P}(x_{h+1}) - \mathcal{P}(x_h)\right)\\ &\quad - \sum_{x_H}\gamma_{x_H}\mathcal{P}(x_H) + \sum_{a_{1:H-1}}\eta_{a_{1:H-1}}\\ &\quad \left(\sum_{o_{2:H}}\mathcal{P}(x_H)\log\frac{\mathcal{P}(x_H)}{\mathcal{P}^*(x_H)} - \xi\right),\end{aligned}$$

where $\lambda_{(a_h,x_h)} \in \mathbb{R}$, and $\gamma_{x_H}, \eta_{a_{1:H-1}} \geq 0, \quad \forall x_H$.

Thus, the dual problem is

$$\text{D1:}\begin{cases}\max_{\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\eta}}\inf_{\mathcal{P},s} L_2(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\eta})\\ \text{s.t. } f(x_H) + \lambda_{(a_{H-1},x_{H-1})} - \gamma_{x_H} + \eta_{a_{1:H-1}}\\ \quad\left(\log\frac{\mathcal{P}(x_H)}{\mathcal{P}^*(x_H)} + 1\right) = 0, \quad \forall x_H\\ \quad\lambda_{(a_{h-1},x_{h-1})} = \sum_{a_h}\lambda_{(a_h,x_h)}, \quad \forall h, x_h\\ \quad\gamma_{x_H}, \eta_{a_{1:H-1}} \geq 0, \quad \forall x_H.\end{cases}$$

Denote $\inf_{\mathcal{P},s} L_2(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\eta})$ by $\bar{L}_2$. Then, we have

$$\begin{aligned}\bar{L}_2 &= -\sum_{x_H}\eta_{a_{1:H-1}}\mathcal{P}(x_H) - \lambda_{(a_1,x_1)} - \sum_{a_{1:H-1}}\eta_{a_{1:H-1}}\xi\\ &= -\sum_{x_H}\eta_{a_{1:H-1}}\mathcal{P}(x_H) - \sum_{x_H}\mathcal{P}^*(x_H)\lambda_{(a_{H-1},x_{H-1})}\end{aligned}$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

$$
\begin{aligned}
&- \sum_{a_{1:H-1}} \eta_{a_{1:H-1}} \xi \\
&= - \sum_{x_H} \eta_{a_{1:H-1}} \mathcal{P}^*(x_H) \\
&\quad \exp\left(\frac{\gamma_{x_H} - f(x_H) - \lambda_{(a_{H-1}, x_{H-1})}}{\eta_{a_{1:H-1}}} - 1 \right) \\
&\quad - \sum_{x_H} \mathcal{P}^*(x_H) \lambda_{(a_{H-1}, x_{H-1})} - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}} \xi.
\end{aligned}
$$

Hence, by the Slater's condition, the primal value equals to the dual value. Mathematically, we have, shown in the equation at the bottom of the page,where we choose $\gamma_{x_H} = 0$ and

$$
\lambda_{(a_{H-1}, x_{H-1})} = \eta_{a_{1:H-1}} \log \left(\sum_o \mathcal{T}^*_{H-1}(o|\tau_{H-1}) \exp\left(\frac{-f(x_H)}{\eta_{a_{1:H-1}}} - 1 \right) \right).
$$

□

B. Bellman-Type Equations Under $\mathcal{T}$-Type Uncertainty Sets

Recall that robust value function is defined as follows.

$$
V^\pi_{B(\theta^*),R,h}(\tau_{h-1}, o_h) = \inf_{\theta \in B(\theta^*)} \mathbb{E}^\pi_\theta \left[\sum_{h'=h}^{H} R(\boldsymbol{\tau}_{h'}) \middle| \tau_{h-1}, o_h \right],
$$

$$
Q^\pi_{B(\theta^*),R,h}(\tau_h) = \inf_{\theta \in B(\theta^*)} \mathbb{E}^\pi_\theta \left[\sum_{h'=h}^{H} R(\boldsymbol{\tau}_{h'}) \middle| \tau_h \right].
$$

Then, we have the following Bellman-type equations for robust value functions.

Lemma 1.5: For any θ, we have, shown in the equation at the bottom of the page.

Proof: The first equation is straightforward.

For the second equation, by the definition of robust Q-value function, for any $\iota > 0$, there exists $\theta_\iota \in B(\theta^*)$ such that

$$
Q^\pi_{B(\theta^*),R,h}(\tau_h) \geq \mathbb{E}^\pi_{\theta_\iota} \left[\sum_{h'=h}^{H} R(\boldsymbol{\tau}_{h'}) \middle| \tau_h \right] - \iota.
$$

Then, we have

$$
\begin{aligned}
&Q^\pi_{B(\theta^*),R,h}(\tau_h) \\
&\geq \mathbb{E}^\pi_{\theta_\iota} \left[\sum_{h'=h}^{H} R(\boldsymbol{\tau}_{h'}) \middle| \tau_h \right] - \iota \\
&= R(\tau_h) + \sum_{o_{h+1}} \mathcal{T}^{\theta_\iota}_h(o_{h+1}|\tau_h) V^\pi_{\theta_\iota,R,h+1}(\tau_h, o_{h+1}) - \iota \\
&\geq R(\tau_h) + \inf_{\theta \in B(\theta^*)} \sum_{o_{h+1}} \mathcal{T}^{\theta}_h(o_{h+1}|\tau_h) \\
&\quad V^\pi_{B(\theta^*),R,h+1}(\tau_h, o_{h+1}) - \iota.
\end{aligned}
$$

Let $\iota \to 0$, we conclude that

$$
\begin{aligned}
&Q^\pi_{B(\theta^*),R,h}(\tau_h) \geq R(\tau_h) \\
&\quad + \inf_{\theta \in B(\theta^*)} \sum_{o_{h+1}} \mathcal{T}^{\theta}_h(o_{h+1}|\tau_h) V^\pi_{B(\theta^*),R,h+1}(\tau_h, o_{h+1}).
\end{aligned}
$$

On the other hand, for any $\iota_1, \iota_2 > 0$, there exist $\theta_{\iota_1}, \theta_{\iota_2} \in B(\theta^*)$ such that

$$
V^\pi_{B(\theta^*),R,h+1}(\tau_h, o_{h+1}) \geq V^\pi_{\theta_{\iota_1},R,h+1}(\tau_h, o_{h+1}) - \iota_1,
$$

and

$$
\begin{aligned}
&\inf_{\theta \in B(\theta^*)} \sum_{o_{h+1}} \mathcal{T}^{\theta}_h(o_{h+1}|\tau_h) V^\pi_{B(\theta^*),R,h+1}(\tau_h, o_{h+1}) \\
&\qquad \geq \sum_{o_{h+1}} \mathcal{T}^{\theta_{\iota_2}}_h(o_{h+1}|\tau_h) V^\pi_{B(\theta^*),R,h+1}(\tau_h, o_{h+1}) - \iota_2.
\end{aligned}
$$

In addition, there exists $\theta_3 \in B(\theta^*)$ such that $\mathcal{T}_{\theta_3}(o_{h+1}|\tau_h) = \mathcal{T}_{\theta_{\iota_2}}(o_{h+1}|\tau_h)$ and $V^\pi_{\theta_{\iota_1},R,h+1} = V^\pi_{\theta_3,R,h+1}$. Then, we have

$$
\begin{aligned}
&R(\tau_h) + \inf_{\theta \in B(\theta^*)} \sum_{o_{h+1}} \mathcal{T}^{\theta}_h(o_{h+1}|\tau_h) V^\pi_{B(\theta^*),R,h+1}(\tau_h, o_{h+1}) \\
&\quad \geq R(\tau_h) + \sum_{o_{h+1}} \mathcal{T}^{\theta_3}_h(o_{h+1}|\tau_h) V^\pi_{B(\theta^*),R,h+1}(\tau_h, o_{h+1}) - \iota_2 \\
&\quad \geq R(\tau_h) + \sum_{o_{h+1}} \mathcal{T}^{\theta_3}_h(o_{h+1}|\tau_h) V^\pi_{\theta_3,R,h+1}(\tau_h, o_{h+1}) - \iota_1 - \iota_2 \\
&\quad = \mathbb{E}^\pi_{\theta_3} \left[\sum_{h'=h}^{H} R(\boldsymbol{\tau}_h) \middle| \tau_h \right] - \iota_1 - \iota_2 \\
&\quad \geq Q^\pi_{B(\theta^*),R,h}(\tau_h) - \iota_1 - \iota_2.
\end{aligned}
$$

$$
\begin{aligned}
&V^\pi_{B(\theta^*),R} \\
&= \max_{\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\eta}} \bar{L}_2(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\eta}) \\
&= \max_{\boldsymbol{\eta} \geq 0} \max_{\boldsymbol{\gamma} \geq 0, \boldsymbol{\lambda}} \left\{ - \sum_{x_H} \eta_{a_{1:H-1}} \mathcal{P}^*(x_H) \exp\left(\frac{\gamma_{x_H} - f(x_H) - \lambda_{(a_{H-1}, x_{H-1})}}{\eta_{a_{1:H-1}}} - 1 \right) \right. \\
&\qquad \left. - \sum_{x_H} \mathcal{P}^*(x_H) \lambda_{(a_{H-1}, x_{H-1})} - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}} \xi \right\} \\
&= \max_{\boldsymbol{\eta} \geq 0} \left\{ - \sum_{a_{H-1}, x_{H-1}} \eta_{a_{H-1}, x_{H-1}} \mathcal{P}^*(x_{H-1}) \log \mathbb{E}_{o \sim \mathcal{T}^*_{H-1}(\cdot|\tau_{H-1})} \left[\exp\left(- \frac{f(x_H)}{\eta_{a_{1:H-1}}} \right) \right] - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}} \xi \right\},
\end{aligned}
$$

$$
\begin{cases}
V^\pi_{B(\theta^*),R,h}(\tau_{h-1}, o_h) = \sum_{a_h} \pi(a_h|\tau_{h-1}, o_h) Q^\pi_{B(\theta^*),R,h}(\tau_h) \\
Q^\pi_{B(\theta^*),R,h}(\tau_h) = R(\tau_h) + \inf_{\theta \in B(\theta^*)} \sum_{o_{h+1}} \mathcal{T}^{\theta}_h(o_{h+1}|\tau_h) V^\pi_{B(\theta^*),R,h+1}(\tau_h, o_{h+1}).
\end{cases}
$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

Let $\iota_1, \iota_2 \to 0$, we can conclude that

$$R(\tau_h) + \inf_{\theta \in B(\theta^*)} \sum_{o_{h+1}} \mathcal{T}_h^{\theta}(o_{h+1}|\tau_h) V^{\pi}_{B(\theta^*),R,h+1}(\tau_h, o_{h+1}) \geq Q^{\pi}_{B(\theta^*),R,h}(\tau_h),$$

which finishes the proof.□

C. Simulation Lemmas of Robust Values

Lemma 1.6: (Simulation Lemma of Robust Value Under $B^1_{\mathcal{P}}$):

$$\left| V^{\pi}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\pi}_{B^1_{\mathcal{P}}(\theta),R} \right| \leq \left\| \mathbb{D}^{\pi}_{\theta} - \mathbb{D}^{\pi}_{\theta^*} \right\|_1$$

Proof: Let $f(x_H) = \pi(\tau_{H-1})R(x_H)$.

We first prove that $V^{\pi}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\pi}_{B^1_{\mathcal{P}}(\theta),R} \leq \left\| \mathbb{D}^{\pi}_{\theta} - \mathbb{D}^{\pi}_{\theta^*} \right\|_1$. By Proposition 1.3, we have, shown in the equation at the bottom of the page.

Let $\boldsymbol{\gamma}^*, \boldsymbol{\lambda}^*$ be the solution to

$$\max_{\boldsymbol{\gamma} \geq 0, \boldsymbol{\lambda}} \left\{ \sum_{x_H} (f(x_H) - \gamma_{x_H}) \mathcal{P}^{\theta^*}(x_H) - \sum_{a_{1:H-1}} \max_{o_{1:H}} \left| \gamma_{x_H} - f(x_H) - \lambda_{(a_{H-1}, x_{H-1})} \right| \xi \right\}.$$

Again, from Proposition 1.3, we know that

$$-f(x_H) \leq \lambda^*_{(a_{H-1}, x_{H-1})} \leq 0,$$
$$0 \leq \gamma^*_{x_H} \leq \max\left\{ f(x_H) + \lambda^*_{a_{H-1}, x_{H-1}}, 0 \right\} \leq f(x_H). \quad (9)$$

Thus, the difference can be upper bounded as follows.

$$\begin{aligned} & V^{\pi}_{B(\theta^*),R} - V^{\pi}_{B(\theta),R} \\ & \leq \sum_{x_H} (f(x_H) - \gamma^*_{x_H}) \mathcal{P}^{\theta^*}(x_H) - \sum_{a_{1:H-1}} \max_{o_{1:H}} \left| \gamma^*_{x_H} - f(x_H) - \lambda^*_{(a_{H-1}, x_{H-1})} \right| \xi \\ & \quad - \sum_{x_H} (f(x_H) - \gamma^*_{x_H}) \mathcal{P}^{\theta}(x_H) + \sum_{a_{1:H-1}} \max_{o_{1:H}} \left| \gamma^*_{x_H} - f(x_H) - \lambda^*_{(a_{H-1}, x_{H-1})} \right| \xi \\ & = \sum_{x_H} (f(x_H) - \gamma^*_{x_H}) (\mathcal{P}^{\theta^*}(x_H) - \mathcal{P}^{\theta}(x_H)) \\ & \leq \sum_{x_H} |f(x_H) - \gamma^*_{x_H}| |\mathcal{P}^{\theta^*}(x_H) - \mathcal{P}^{\theta}(x_H)| \\ & \overset{(a)}{\leq} \sum_{x_H} f(x_H) |\mathcal{P}^{\theta^*}(x_H) - \mathcal{P}^{\theta}(x_H)| \\ & = \sum_{x_H} \pi(\tau_{H-1}) R(x_H) |\mathcal{P}^{\theta^*}(x_H) - \mathcal{P}^{\theta}(x_H)| \\ & \leq \left\| \mathbb{D}^{\pi}_{\theta} - \mathbb{D}^{\pi}_{\theta^*} \right\|_1, \end{aligned}$$

where (a) is due to Equation (9).

The opposite side $V^{\pi}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\pi}_{B^1_{\mathcal{P}}(\theta),R} \geq -\left\| \mathbb{D}^{\pi}_{\theta} - \mathbb{D}^{\pi}_{\theta^*} \right\|_1$ follows exactly the same proof.□

Lemma 1.7: (Simulation Lemma of Robust Value Under $B^2_{\mathcal{P}}$):

$$\left| V^{\pi}_{B^2_{\mathcal{P}}(\theta^*),R} - V^{\pi}_{B^2_{\mathcal{P}}(\theta),R} \right| \leq 3 \exp(1/\eta^*) \left\| \mathbb{D}^{\pi}_{\theta} - \mathbb{D}^{\pi}_{\theta^*} \right\|_1,$$

where η^* is a lower bound of the optimal solution, shown in the equation at the bottom of the page.

Here, if $x \leq v$, where x is a scalar and $\mathbf{v}$ is a vector, then x is less than or equal to each coordinate of the vector.

Proof: Let $f(x_H) = \pi(\tau_{H-1})R(x_H)$. By Proposition 1.4, we have, shown in the equation at the bottom of the next page.

Let $\boldsymbol{\eta}^*$ be the solution to, shown in the equation at the bottom of the next page.

Then, the difference can be upper bounded as follows, shown in the equation at the bottom of the next page.

For term I_0, we fix τ_{H-1} and write $f(x_H)$ as $\pi(\tau_{H-1})R(o_H)$ for convenience.

Since τ_{H-1} is fixed and we aim to upper bound the term

$$I_0 = \log \frac{\mathbb{E}_{o_H \sim \mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1})} \left[\exp\left(-\frac{\pi(\tau_{H-1})R(o_H)}{\eta^*_{a_{1:H-1}}} \right) \right]}{\mathbb{E}_{o_H \sim \mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1})} \left[\exp\left(-\frac{\pi(\tau_{H-1})R(o_H)}{\eta^*_{a_{1:H-1}}} \right) \right]},$$

$$\begin{aligned} & V^{\pi}_{B(\theta^*),R} - V^{\pi}_{B(\theta),R} \\ & = \max_{\boldsymbol{\gamma} \geq 0, \boldsymbol{\lambda}} \left\{ \sum_{x_H} (f(x_H) - \gamma_{x_H}) \mathcal{P}^{\theta^*}(x_H) - \sum_{a_{1:H-1}} \max_{o_{1:H}} \left| \gamma_{x_H} - f(x_H) - \lambda_{(a_{H-1}, x_{H-1})} \right| \xi \right\} \\ & \quad - \max_{\boldsymbol{\gamma} \geq 0, \boldsymbol{\lambda}} \left\{ \sum_{x_H} (f(x_H) - \gamma_{x_H}) \mathcal{P}^{\theta}(x_H) - \sum_{a_{1:H-1}} \max_{o_{1:H}} \left| \gamma_{x_H} - f(x_H) - \lambda_{(a_{H-1}, x_{H-1})} \right| \xi \right\}, \end{aligned}$$

$$\eta^* \leq \min \left\{ \begin{array}{l} \arg\max_{\boldsymbol{\eta} \geq 0} \left\{ - \sum_{a_{H-1}, x_{H-1}} \eta_{a_{H-1}, x_{H-1}} \mathcal{P}^{\theta^*}(x_{H-1}) \log \mathbb{E}_{o \sim \mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1})} \left[\exp\left(-\frac{f(x_H)}{\eta_{a_{1:H-1}}} \right) \right] \right. \\ \qquad \left. - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}} \xi \right\}, \\ \arg\max_{\boldsymbol{\eta} \geq 0} \left\{ - \sum_{a_{H-1}, x_{H-1}} \eta_{a_{H-1}, x_{H-1}} \mathcal{P}^{\theta}(x_{H-1}) \log \mathbb{E}_{o \sim \mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1})} \left[\exp\left(-\frac{f(x_H)}{\eta_{a_{1:H-1}}} \right) \right] \right. \\ \qquad \left. - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}} \xi \right\}. \end{array} \right.$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

$\pi(\tau_{H-1})/\eta^*_{a_{1:H-1}} \in [0, 1/\eta^*]$ can be treated as a variable. We define a function $g(t)$ for $1/\eta^* \geq t \geq 0$ as follows

$$g(t) = \log \frac{\mathbb{E}_{o\sim P}\left[\exp\left(-tR(o)\right)\right]}{\mathbb{E}_{o\sim Q}\left[\exp\left(-tR(o)\right)\right]},$$

where $P, Q \in \Delta(\mathcal{O})$ are two distributions over the observation space.

Taking the derivative with respect to t, we have

$$\begin{aligned}\frac{dg}{dt} &= \frac{\mathbb{E}_{o\sim P}\left[-R(o)\exp\left(-tR(o)\right)\right]}{\mathbb{E}_{o\sim P}\left[\exp\left(-tR(o)\right)\right]} \\ &\quad - \frac{\mathbb{E}_{o\sim Q}\left[-R(o)\exp\left(-tR(o)\right)\right]}{\mathbb{E}_{o\sim Q}\left[\exp\left(-tR(o)\right)\right]} \\ &= \mathbb{E}_{o\sim Q}\left[R(o)\frac{\exp\left(-tR(o)\right)}{\mathbb{E}_{o\sim Q}\left[\exp\left(-tR(o)\right)\right]}\right] \\ &\quad - \mathbb{E}_{o\sim P}\left[R(o)\frac{\exp\left(-tR(o)\right)}{\mathbb{E}_{o\sim P}\left[\exp\left(-tR(o)\right)\right]}\right].\end{aligned}$$

Due to $R(o) \in [0, 1]$ and $1/\eta^* \geq t \geq 0$, we have

$$\left|\frac{dg}{dt}\right| \leq \|P - Q\|_1 \frac{1}{\exp(-1/\eta^*)}.$$

Thus, for $t \in [0, 1/\eta^*]$, we have

$$g(t) = \int_0^t g'(t)dt \leq t\|P - Q\|_1 \exp(1/\eta^*).$$

$$\begin{aligned}&V^{\pi}_{B(\theta^*),R} - V^{\pi}_{B(\theta),R} \\ &= \max_{\boldsymbol{\eta}\geq 0}\left\{-\sum_{a_{H-1},x_{H-1}} \eta_{a_{H-1},x_{H-1}} \mathcal{P}^{\theta^*}(x_{H-1}) \log \mathbb{E}_{o\sim\mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta_{a_{1:H-1}}}\right)\right] - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}}\xi\right\} \\ &\quad - \max_{\boldsymbol{\eta}\geq 0}\left\{-\sum_{a_{H-1},x_{H-1}} \eta_{a_{H-1},x_{H-1}} \mathcal{P}^{\theta'}(x_{H-1}) \log \mathbb{E}_{o\sim\mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta_{a_{1:H-1}}}\right)\right] - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}}\xi\right\}.\end{aligned}$$

$$\max_{\boldsymbol{\eta}\geq 0}\left\{-\sum_{a_{H-1},x_{H-1}} \eta_{a_{H-1},x_{H-1}} \mathcal{P}^{\theta^*}(x_{H-1}) \log \mathbb{E}_{o\sim\mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta_{a_{1:H-1}}}\right)\right] - \sum_{a_{1:H-1}} \eta_{a_{1:H-1}}\xi\right\}.$$

$$\begin{aligned}&V^{\pi}_{B(\theta^*),R} - V^{\pi}_{B(\theta),R} \\ &\leq \sum_{a_{H-1},x_{H-1}} \eta^*_{a_{H-1},x_{H-1}} \mathcal{P}^{\theta} \log \mathbb{E}_{o_H\sim\mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right] + \sum_{a_{1:H-1}} \eta^*_{a_{1:H-1}}\xi \\ &\quad - \sum_{a_{H-1},x_{H-1}} \eta^*_{a_{H-1},x_{H-1}} \mathcal{P}^{\theta^*}(x_{H-1}) \log \mathbb{E}_{o_H\sim\mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right] - \sum_{a_{1:H-1}} \eta^*_{a_{1:H-1}}\xi \\ &= \sum_{a_{H-1},x_{H-1}} \eta^*_{a_{H-1},x_{H-1}} \left(\mathcal{P}^{\theta}(x_{H-1}) - \mathcal{P}^{\theta^*}(x_{H-1})\right) \log \mathbb{E}_{o_H\sim\mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right] \\ &\quad + \sum_{a_{H-1},x_{H-1}} \eta^*_{a_{H-1},x_{H-1}} \mathcal{P}^{\theta^*}(x_{H-1}) \left(\log \mathbb{E}_{o_H\sim\mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right]\right. \\ &\quad \left. - \log \mathbb{E}_{o_H\sim\mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right]\right) \\ &\leq \sum_{a_{H-1},x_{H-1}} -\eta^*_{a_{H-1},x_{H-1}} \left|\mathcal{P}^{\theta}(x_{H-1}) - \mathcal{P}^{\theta^*}(x_{H-1})\right| \log \mathbb{E}_{o_H\sim\mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right] \\ &\quad + \sum_{a_{H-1},x_{H-1}} \eta^*_{a_{H-1},x_{H-1}} \mathcal{P}^{\theta^*}(x_{H-1}) \log \frac{\mathbb{E}_{o_H\sim\mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right]}{\mathbb{E}_{o_H\sim\mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right]} \\ &\leq \sum_{a_{H-1},x_{H-1}} \left|\mathcal{P}^{\theta}(x_{H-1}) - \mathcal{P}^{\theta^*}(x_{H-1})\right| \max_{o_H} f(x_H) \\ &\quad + \sum_{a_{H-1},x_{H-1}} \eta^*_{a_{H-1},x_{H-1}} \mathcal{P}^{\theta^*}(x_{H-1}) \underbrace{\log \frac{\mathbb{E}_{o_H\sim\mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right]}{\mathbb{E}_{o_H\sim\mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1})}\left[\exp\left(-\frac{f(x_H)}{\eta^*_{a_{1:H-1}}}\right)\right]}}_{I_0}\end{aligned}$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

Plugging the upper bound of $g(t)$ back to the robust value difference, we have

$$\begin{aligned}
&V^{\pi}_{B(\theta^*),R} - V^{\pi}_{B(\theta),R} \\
&\quad\le \sum_{a_{H-1},x_{H-1}} \left|\mathcal{P}^{\theta}(x_{H-1}) - \mathcal{P}^{\theta^*}(x_{H-1})\right| \max_{o_H} f(x_H) \\
&\qquad + \sum_{a_{H-1},x_{H-1}} \eta^*_{a_{H-1},x_{H-1}} \mathcal{P}^{\theta^*}(x_{H-1}) \left\| \mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1}) \right. \\
&\qquad \left. - \mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1}) \right\|_1 \frac{\pi(\tau_{H-1})}{\eta_{a_{1:H-1}}} \exp(1/\eta^*) \\
&\quad\le \sum_{\tau_{H-1}} \left|\mathbb{D}^{\pi}_{\theta}(\tau_{H-1}) - \mathbb{D}^{\pi}_{\theta^*}(\tau_{H-1})\right| + \exp(1/\eta^*)\mathbb{E}^{\pi}_{\theta^*} \\
&\qquad \left[\left\| \mathcal{T}^{\theta}_{H-1}(\cdot|\tau_{H-1}) - \mathcal{T}^{\theta^*}_{H-1}(\cdot|\tau_{H-1}) \right\|_1\right] \\
&\quad\overset{(a)}{\le} 3\exp(1/\eta^*) \left\| \mathbb{D}^{\pi}_{\theta} - \mathbb{D}^{\pi}_{\theta^*} \right\|_1,
\end{aligned}$$

where (a) is due to Lemma 5.1.

For the second inequality, we follow the same argument and the proof is complete.□

Lemma 1.8: (Simulation Lemma for Robust Value Under $B^i_{\mathcal{T}}$):

$$\begin{aligned}
\left|V^{\pi}_{B^1_{\mathcal{T}}(\theta^*),R} - V^{\pi}_{B^1_{\mathcal{T}}(\theta),R}\right| &\le 2C_B \|\mathbb{D}^{\pi}_{\theta^*} - \mathbb{D}^{\pi}_{\theta}\|_1, \\
\left|V^{\pi}_{B^2_{\mathcal{T}}(\theta^*),R} - V^{\pi}_{B^2_{\mathcal{T}}(\theta),R}\right| &\le 2C_B C^{\mathrm{kl}}_{\mathcal{T}} \|\mathbb{D}^{\pi}_{\theta^*} - \mathbb{D}^{\pi}_{\theta}\|_1
\end{aligned}$$

where $C^{\mathrm{kl}}_{\mathcal{T}} = \max\left\{\exp(\xi)/\xi, \lambda^* \exp(1/\lambda^*)\right\}$, and λ^* is a lower bound for the optimal solutions of the following problems, shown in the equation at the bottom of the page.

Proof: First, we have, shown in the equation at the bottom of the page.

When $i = 1$, we have, shown in the equation at the bottom of the next page, where (a) is due to Proposition 1.1.

When $i = 2$, we have, shown in the equation at the bottom of the page, where (a) is due to Proposition 1.2.

Let $\lambda^*(\tau_h) = \arg\sup_{\lambda\ge 0}\left\{ -\lambda \log \mathbb{E}_{\mathcal{T}^{\theta^*}_h}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda)] - \lambda\xi \right\}$.

Then, we further have, shown in the equation at the bottom of the next page, where (a) follows from $\log(x) \le x - 1$ for $x > 0$, and $0 < \lambda^* \le \min_{h,\tau_h} \lambda^*(\tau_h)$.

In summary, the difference between robust values can be upper bounded by, shown in the equation at the bottom of page 7222, where

$$C_i = \begin{cases} 1, & i = 1, \\ \max\left\{\frac{\exp(\xi)}{\xi}, \lambda^* \exp(1/\lambda^*)\right\}, & i = 2. \end{cases}$$

The opposite side of the inequalities of the lemma follows the same argument.□

Appendix B
MLE Analysis

In this section, we provide the estimation guarantee for MLE oracle in Lemma 2.5.

We first present a sequence of supporting lemmas in the following. Notably, the first lemma states that after dataset distillation, we still have sufficient offline samples.

Lemma 2.1 (Sufficient Good Samples): Let $p_{\min} < \frac{\delta}{|\mathcal{O}|^{2H}|\mathcal{A}|^{2H}}$, then, with probability at least $1 - \delta$, we have $|\mathcal{D}^{\mathrm{g}}| \ge N/2$.

Proof: Let $L^* = H\log(|\mathcal{O}||\mathcal{A}|)$. Then, we have

$$\mathbb{E}\left[\exp\left(\sum_{n=1}^{N}(-\log \mathbb{D}^{o}_{\theta^*}(\tau^n_H) - L^*)\right)\right]$$

$$\lambda^* = \min_{h,\tau_h} \begin{cases} \arg\sup_{\lambda\ge 0}\left\{-\lambda \log \mathbb{E}_{\mathcal{T}^{\theta^*}_h}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda)] - \lambda\xi\right\}, \\ \arg\sup_{\lambda\ge 0}\left\{-\lambda \log \mathbb{E}_{\mathcal{T}^{\theta}_h}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta^*),R,h+1}(\tau_h, o)/\lambda)] - \lambda\xi\right\}. \end{cases}$$

$$\begin{aligned}
&V^{\pi}_{B^i_{\mathcal{T}}(\theta^*),R} - V^{\pi}_{B^i_{\mathcal{T}}(\theta),R} \\
&\quad= \mathbb{E}^{\pi}\left[\left. Q^{\pi}_{B^1_{\mathcal{T}}(\theta^*),R,1}(\tau_1) - Q^{\pi}_{B^i_{\mathcal{T}}(\theta),R,1}(\tau_1)\right| o_1\right] \\
&\quad= \mathbb{E}^{\pi}\left[\inf_{\theta'\in B^i_{\mathcal{T}}(\theta^*)} \sum_{o_2} \mathcal{T}^{\theta'}_1(o_2|\tau_1) V^{\pi}_{B^i_{\mathcal{T}}(\theta^*),R,2}(\tau_1, o_2) - \inf_{\theta'\in B^i_{\mathcal{T}}(\theta)} \sum_{o_2} \mathcal{T}^{\theta'}_1(o_2|\tau_1) V^{\pi}_{B^i_{\mathcal{T}}(\theta),R,2}(\tau_1, o_2)\right] \\
&\quad= \mathbb{E}^{\pi}\left[\inf_{\theta'\in B^i_{\mathcal{T}}(\theta^*)} \sum_{o_2} \mathcal{T}^{\theta}_1(o_2|\tau_1) V^{\pi}_{B^i_{\mathcal{T}}(\theta^*),R,2}(\tau_1, o_2) - \inf_{\theta'\in B^i_{\mathcal{T}}(\theta^*)} \sum_{o_2} \mathcal{T}^{\theta}_1(o_2|\tau_1) V^{\pi}_{B^i_{\mathcal{T}}(\theta),R,2}(\tau_1, o_2)\right. \\
&\qquad\left. + \inf_{\theta'\in B^i_{\mathcal{T}}(\theta^*)} \sum_{o_2} \mathcal{T}^{\theta'}_1(o_2|\tau_1) V^{\pi}_{B^i_{\mathcal{T}}(\theta),R,2}(\tau_1, o_2) - \inf_{\theta'\in B^i_{\mathcal{T}}(\theta)} \sum_{o_2} \mathcal{T}^{\theta'}_1(o_2|\tau_1) V^{\pi}_{B^i_{\mathcal{T}}(\theta),R,2}(\tau_1, o_2)\right] \\
&\quad\le \max_{\theta'\in B^1_{\mathcal{T}}(\theta^*)} \mathbb{E}^{\pi}_{\theta'}\left[V^{\pi}_{B^i_{\mathcal{T}}(\theta^*),R,2}(\tau_1, o_2) - V^{\pi}_{B^i_{\mathcal{T}}(\theta),R,2}(\tau_1, o_2)\right] + \mathbb{E}^{\pi}\left[\mathcal{E}_{B^i_{\mathcal{T}}}(\theta, \tau_1)\right] \\
&\quad\le \dots \\
&\quad\le \max_{\theta'\in B^i_{\mathcal{T}}(\theta^*)} \sum_{h=1}^{H} \mathbb{E}^{\pi}_{\theta'}\left[\mathcal{E}_{B^i_{\mathcal{T}}}(\theta, \tau_h)\right] \le C_B \sum_{h=1}^{H} \mathbb{E}^{\pi}_{\theta^*}\left[\mathcal{E}_{B^i_{\mathcal{T}}}(\theta, \tau_h)\right]
\end{aligned}$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

$$= \mathbb{E}\left[\prod_{n=1}^{N} \frac{1}{\mathbb{D}^{\rho}_{\theta^*}(\tau_H^n) e^{L^*}}\right]$$
$$= \prod_{n=1}^{N} \frac{|\mathcal{O}|^H |\mathcal{A}|^H}{e^{L^*}}$$
$$= 1.$$

Thus, by Chernoff bound, we have that with probability $1-\delta$,

$$\min_{\theta} \mathcal{L}(\theta|\mathcal{D}) \le \frac{1}{N}\sum_{n=1}^{N} -\log \mathbb{D}^{\rho}_{\theta^*}(\tau_H^n) \le L^* + \frac{1}{N}\log\frac{1}{\delta}.$$

Recall that $\hat{\theta} = \min_\theta \mathcal{L}(\theta|\mathcal{D})$. Let $\mathcal{D}^{\mathrm{b}} = \mathcal{D}/\mathcal{D}^{\mathrm{g}} = \{\tau_H \in \mathcal{D}, \mathbb{D}^{\rho}_{\hat{\theta}}(\tau_H) < p_{\min}\}$. Thus, we further have

$$\begin{aligned} NL^* + \log\frac{1}{\delta} &\ge \sum_{n=1}^{N} -\log \mathbb{D}^{\rho}_{\hat{\theta}}(\tau_H^n) \\ &\ge \sum_{\tau_h \in \mathcal{D}^{\mathrm{b}}} -\log \mathbb{D}^{\rho}_{\hat{\theta}}(\tau_H) \\ &\ge -|\mathcal{D}^{\mathrm{b}}| \log p_{\min}, \end{aligned}$$

which implies

$$|\mathcal{D}^{\mathrm{b}}| \le \frac{NL^* + \log(1/\delta)}{\log\frac{1}{p_{\min}}} < \frac{N}{2}.$$

□

Lemma 2.2: Fix $\varepsilon < \frac{1}{N}$. Let $\bar{\Theta}_\varepsilon$ with size $\mathcal{N}_\varepsilon(\Theta)$ consist of all ε-brackets who can cover Θ. With probability at least $1-\delta$, for any $[\underline{\theta}, \bar{\theta}] \in \bar{\Theta}_\varepsilon$, the following two inequalities hold: shown in the equation at the bottom of the next page.

Proof: We start with the first inequality. Suppose the data in $\mathcal{D}_h^{\mathrm{g}}$ is indexed by t. Then,

$$\begin{aligned} &\mathbb{E}\left[\exp\left(\sum_h \sum_{\tau_h \in \mathcal{D}_h^{\mathrm{g}}} \log \frac{\mathbb{D}^{\rho}_{\bar{\theta}}(\tau_h)}{\mathbb{D}^{\rho}_{\theta^*}(\tau_h)}\right)\right] \\ &= \mathbb{E}\left[\prod_{t \le |\mathcal{D}_h^{\mathrm{g}}|} \prod_h \frac{\mathbb{D}^{\rho}_{\bar{\theta}}(\tau_h^t)}{\mathbb{D}^{\rho}_{\theta^*}(\tau_h^t)}\right] \\ &= \mathbb{E}\left[\prod_{t \le |\mathcal{D}_h^{\mathrm{g}}|-1} \prod_h \frac{\mathbb{D}^{\rho}_{\bar{\theta}}(\tau_h^t)}{\mathbb{D}^{\rho}_{\theta^*}(\tau_h^t)} \mathbb{E}\left[\frac{\mathbb{D}^{\rho}_{\bar{\theta}}(\tau_h^{|\mathcal{D}_h^{\mathrm{g}}|})}{\mathbb{D}^{\rho}_{\theta^*}(\tau_h^{|\mathcal{D}_h^{\mathrm{g}}|})}\right]\right] \end{aligned}$$

$$\begin{aligned} \mathcal{E}_{B^1_{\mathcal{T}}}(\theta, \tau_h) &= \min_{\theta' \in B^1_{\mathcal{T}}(\theta^*)} \mathop{\mathbb{E}}_{o \sim \mathcal{T}_h^{\theta'}(\cdot|\tau_h)} \left[V^{\pi}_{B^1_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)\right] - \min_{\theta' \in B^1_{\mathcal{T}}(\theta)} \mathop{\mathbb{E}}_{o \sim \mathcal{T}_h^{\theta'}(\cdot|\tau_h)} \left[V^{\pi}_{B^1_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)\right] \\ &\overset{(a)}{=} \sup_{\lambda \in [0,1]} \left\{-\mathbb{E}_{\mathcal{T}_h^{\theta^*}}[(\lambda - V^{\pi}_{B^1_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o))_+] - \xi \max_o (\lambda - V^{\pi}_{B^1_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o))_+ + \lambda\right\} \\ &\quad - \sup_{\lambda \in [0,1]} \left\{-\mathbb{E}_{\mathcal{T}_h^{\theta}}[(\lambda - V^{\pi}_{B^1_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o))_+] - \xi \max_o (\lambda - V^{\pi}_{B^1_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o))_+ + \lambda\right\} \\ &\le \sup_{\lambda \in [0,1]} \left\{\mathbb{E}_{\mathcal{T}_h^{\theta}}[(\lambda - V^{\pi}_{B^1_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o))_+] - \mathbb{E}_{\mathcal{T}_h^{\theta^*}}[(\lambda - V^{\pi}_{B^1_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o))_+]\right\} \\ &\le \left\|\mathcal{T}_h^{\theta}(\cdot|\tau_h) - \mathcal{T}_h^{\theta^*}(\cdot|\tau_h)\right\|_1, \end{aligned}$$

$$\begin{aligned} \mathcal{E}_{B^2_{\mathcal{T}}}(\theta, \tau_h) &= \min_{\theta' \in B^2_{\mathcal{T}}(\theta^*)} \mathop{\mathbb{E}}_{o \sim \mathcal{T}_h^{\theta'}(\cdot|\tau_h)} \left[V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)\right] - \min_{\theta' \in B^2_{\mathcal{T}}(\theta)} \mathop{\mathbb{E}}_{o \sim \mathcal{T}_h^{\theta'}(\cdot|\tau_h)} \left[V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)\right] \\ &\overset{(a)}{=} \sup_{\lambda \ge 0} \left\{-\lambda \log \mathbb{E}_{\mathcal{T}_h^{\theta^*}}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda)] - \lambda\xi\right\} \\ &\quad - \sup_{\lambda \ge 0} \left\{-\lambda \log \mathbb{E}_{\mathcal{T}_h^{\theta}}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda)] - \lambda\xi\right\}, \end{aligned}$$

$$\begin{aligned} \mathcal{E}_{B^2_{\mathcal{T}}}(\theta, \tau_h) &\le \lambda^*(\tau_h) \log \mathbb{E}_{\mathcal{T}_h^{\theta}}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda^*(\tau_h))] \\ &\quad - \lambda^*(\tau_h) \log \mathbb{E}_{\mathcal{T}_h^{\theta^*}}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda^*(\tau_h))] \\ &= \lambda^*(\tau_h) \log \frac{\mathbb{E}_{\mathcal{T}_h^{\theta}}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda^*(\tau_h))]}{\mathbb{E}_{\mathcal{T}_h^{\theta^*}}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda^*(\tau_h))]} \\ &\overset{(a)}{\le} \lambda^*(\tau_h) \frac{\mathbb{E}_{\mathcal{T}_h^{\theta}}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda^*(\tau_h))] - \mathbb{E}_{\mathcal{T}_h^{\theta^*}}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda^*(\tau_h))]}{\mathbb{E}_{\mathcal{T}_h^{\theta^*}}[\exp(-V^{\pi}_{B^2_{\mathcal{T}}(\theta),R,h+1}(\tau_h, o)/\lambda^*(\tau_h))]} \\ &\le \frac{\lambda^*(\tau_h)}{\exp(-1/\lambda^*(\tau_h))} \left\|\mathcal{T}_h^{\theta^*}(\cdot|\tau_h) - \mathcal{T}_h^{\theta}(\cdot|\tau_h)\right\|_1 \\ &\le \max\left\{\frac{\exp(\xi)}{\xi}, \lambda^* \exp(1/\lambda^*)\right\} \left\|\mathcal{T}_h^{\theta^*}(\cdot|\tau_h) - \mathcal{T}_h^{\theta}(\cdot|\tau_h)\right\|_1, \end{aligned}$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

$$
\begin{aligned}
&= \mathbb{E}\left[\prod_{t\le|\mathcal{D}_h^{\mathrm{g}}|-1}\prod_h \frac{\mathbb{D}_{\bar\theta}^{\rho}(\tau_h^t)}{\mathbb{D}_{\theta^*}^{\rho}(\tau_h^t)}\prod_h\sum_{\tau_h}\mathbb{D}_{\bar\theta}^{\rho}(\tau_h)\right]\\
&\overset{(a)}{\le}(1+\varepsilon)^H\,\mathbb{E}\left[\prod_{t\le|\mathcal{D}_h^{\mathrm{g}}|-1}\prod_h \frac{\mathbb{D}_{\bar\theta}^{\rho}(\tau_h^t)}{\mathbb{D}_{\theta^*}^{\rho}(\tau_h^t)}\right]\\
&\le(1+\varepsilon)^{|\mathcal{D}^{\mathrm{g}}|}\\
&\overset{(b)}{\le} e,
\end{aligned}
$$

where (a) follows because $\sum_{\tau_h}|\mathbb{D}_{\bar\theta}^{\rho}(\tau_h)-\mathbb{D}_{\theta}^{\rho}(\tau_h)|\le\|\mathbb{D}_{\bar\theta}^{\rho}-\mathbb{D}_{\theta}^{\rho}\|_1\le\varepsilon$, and (b) follows because $\varepsilon\le\frac{1}{N}\le\frac{1}{|\mathcal{D}^{\mathrm{g}}|}$.

By the Chernoff bound and the union bound over $\bar\Theta_\epsilon$, with probability at least $1-\delta$, we have

$$
\forall\bar\theta\in\bar\Theta_\epsilon,\quad \sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\log\frac{\mathbb{D}_{\bar\theta}^{\rho}(\tau_h)}{\mathbb{D}_{\theta^*}^{\rho}(\tau_h)}\le 3\log\frac{|\bar\Theta_\epsilon|}{\delta},
$$

which yields the first result of this proposition.

To show the second inequality, we follow an argument similar to that for the first inequality. We have

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(\sum_{\tau_H\in\mathcal{D}^{\mathrm{g}}}\log\frac{\mathbb{D}_{\bar\theta}^{\rho}(\tau_H)}{\mathbb{D}_{\theta^*}^{\rho}(\tau_H)}\right)\right]\\
&\le\mathbb{E}\left[\exp\left(\sum_{\tau_H\in\mathcal{D}^{\mathrm{g}}}\log\frac{\mathbb{D}_{\bar\theta}^{\rho}(\tau_H)}{\mathbb{D}_{\theta^*}^{\rho}(\tau_H)}\right)\right]\\
&\overset{(a)}{\le}(1+\varepsilon)^N\le e,
\end{aligned}
$$

where (a) follows from the tower rule of the expectation and because $\sum_{\tau_H}\mathbb{D}_{\bar\theta}^{\rho}(\tau_H)\le 1+\varepsilon$.

Thus, with probability at least $1-\delta$, for any $\bar\theta\in\Theta$, the following inequality holds

$$
\sum_{\tau_H\in\mathcal{D}^{\mathrm{g}}}\log\frac{\mathbb{D}_{\theta}^{\rho}(\tau_H)}{\mathbb{D}_{\theta^*}^{\rho}(\tau_H)}\le 3\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta},
$$

which completes the proof.□

Proposition 2.3: Fix $p_{\min}$ and $\varepsilon\le\frac{p_{\min}}{N}$. Let $\Theta_{\min}=\{\theta:\forall h,\tau_h\in\mathcal{D}_h^{\mathrm{g}},\ \ \mathbb{D}_{\theta}^{\rho}(\tau_h)\ge p_{\min}\}$. Consider the following event, shown in the equation at the bottom of the page.

Then, $\mathbb{P}(\mathcal{E}_\omega)\ge 1-\delta$.

Proof: We start with a general upper bound on the total variation distance between two conditional distributions. Note that for any $\theta,\theta'\in\Theta\cup\bar\Theta_\epsilon$ and fixed (τ_h,π), we have, shown in the equation at the bottom of the next page.

By symmetry, we also have

$$
\begin{aligned}
&\left\|\mathbb{D}_{\theta}^{\rho}(\omega_h|\tau_h)-\mathbb{D}_{\theta'}^{\rho}(\omega_h|\tau_h)\right\|_1\\
&\qquad\le\frac{2}{\max\left\{\mathbb{D}_{\theta}^{\rho}(\tau_h),\mathbb{D}_{\theta'}^{\rho}(\tau_h)\right\}}\left\|\mathbb{D}_{\theta}^{\rho}-\mathbb{D}_{\theta'}^{\rho}\right\|_1.
\end{aligned}
$$

We replace θ' by a $\bar\theta$ where $[\underline\theta,\bar\theta]\in\bar\Theta_\varepsilon$ is an ε-bracket of θ (recall Definition 4.2), i.e. $\left\|\mathbb{D}_{\theta}^{\rho}-\mathbb{D}_{\bar\theta}^{\rho}\right\|_1\le\varepsilon$, and $\mathbb{D}_{\bar\theta}^{\rho}(\tau_h)\ge\mathbb{D}_{\theta}^{\rho}(\tau_h)$, $\forall\tau_h$. Then, due the construction of $\Theta_{\min}^k$, we have

$$
\forall\tau_h\in\mathcal{D}_h^{\mathrm{g}},\qquad\left\|\mathbb{D}_{\theta}^{\rho}(\omega_h|\tau_h)-\mathbb{D}_{\theta'}^{\rho}(\omega_h|\tau_h)\right\|_1\le\frac{2\varepsilon}{p_{\min}}\le\frac{2}{N},
$$

which implies, shown in the equation at the bottom of the next page.

Here (a) follows because the total variation distance satisfies the triangle inequality and $(a+b)^2\le 2a^2+2b^2$.

Moreover, note that

$$
\begin{aligned}
&\left\|\mathbb{D}_{\bar\theta}^{\rho}(\omega_h|\tau_h)-\mathbb{D}_{\theta^*}^{\rho}(\omega_h|\tau_h)\right\|_1^2\\
&\overset{(a)}{\le}4(2+2/(N))D_H^2\left(\mathbb{D}_{\bar\theta}^{\rho}(\omega_h|\tau_h),\mathbb{D}_{\theta^*}^{\rho}(\omega_h|\tau_h)\right)\\
&\le 6\left(1+\frac{1}{N}-\mathop{\mathbb{E}}_{\omega_h\sim\mathbb{D}_{\theta^*}^{\pi}}\sqrt{\frac{\mathbb{D}_{\bar\theta}^{\rho}(\omega_h|\tau_h)}{\mathbb{D}_{\theta^*}^{\rho}(\omega_h|\tau_h)}}\right)\\
&\overset{(b)}{\le}-6\log\mathop{\mathbb{E}}_{\omega_h\sim\mathbb{D}_{\theta^*}^{\rho}(\cdot|\tau_h)}\sqrt{\frac{\mathbb{D}_{\bar\theta}^{\rho}(\omega_h|\tau_h)}{\mathbb{D}_{\theta^*}^{\rho}(\omega_h|\tau_h)}}+\frac{6}{N},
\end{aligned}
$$

where (a) is due to Lemma 5.2 and (b) follows because $1-x\le-\log x$ for any $x>0$.

Thus, the summation of the total variation distance between conditional distributions conditioned on $\tau_h\in\mathcal{D}_h^{\mathrm{g}}$ can be upper bounded by

$$
\begin{aligned}
&\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\left\|\mathbb{D}_{\bar\theta}^{\rho}(\omega_h|\tau_h)-\mathbb{D}_{\theta^*}^{\rho}(\omega_h|\tau_h)\right\|_1^2\\
&\qquad\le\frac{18}{N}-12\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\log\mathop{\mathbb{E}}_{\omega_h\sim\mathbb{D}_{\theta^*}^{\rho}(\cdot|\tau_h)}\sqrt{\frac{\mathbb{D}_{\bar\theta}^{\rho}(\omega_h|\tau_h)}{\mathbb{D}_{\theta^*}^{\rho}(\omega_h|\tau_h)}}.
\end{aligned}
$$

$$
V_{B^i_{\mathcal{T}}(\theta^*),R}^{\pi}-V_{B^i_{\mathcal{T}}(\theta),R}^{\pi}\le C_BC_i\sum_{h=1}^{H}\mathbb{E}_{\theta^*}^{\pi}\left[\left\|\mathcal{T}_h^{\theta^*}(\cdot|\tau_h)-\mathcal{T}_h^{\theta}(\cdot|\tau_h)\right\|_1\right]\le 2C_BC_i\|\mathbb{D}_{\theta^*}^{\pi}-\mathbb{D}_{\theta}^{\pi}\|_1,
$$

$$
\mathcal{E}_{\log}=\left\{\begin{array}{ll}\forall\bar\theta\in\bar\Theta_\varepsilon, & \sum_h\sum\limits_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\log\mathbb{D}_{\bar\theta}^{\rho}(\tau_h)-3\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta}\le\sum\limits_h\sum\limits_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\log\mathbb{D}_{\theta^*}^{\pi}(\tau_h),\\ \forall\bar\theta\in\bar\Theta_\varepsilon, & \sum\limits_{\tau_H\in\mathcal{D}^{\mathrm{g}}}\log\mathbb{D}_{\bar\theta}^{\rho}(\tau_H)-3\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta}\le\sum\limits_{\tau_H\in\mathcal{D}^{\mathrm{g}}}\log\mathbb{D}_{\theta^*}^{\rho}(\tau_H).\end{array}\right.
$$

$$
\mathcal{E}_\omega=\left\{\forall\theta\in\Theta_{\min},\quad\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\left\|\mathbb{D}_{\hat\theta}^{\rho}(\cdot|\tau_h)-\mathbb{D}_{\theta^*}^{\rho}(\cdot|\tau_h)\right\|_1^2\le 6\sum_h\sum_{\tau_H\in\mathcal{D}_h^{\mathrm{g}}}\log\frac{\mathbb{D}_{\theta^*}^{\rho}(\tau_H)}{\mathbb{D}_{\theta}^{\rho}(\tau_H)}+31\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta}\right\}.
$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

In addition, by only taking expectation over ω_h, we have

$$\mathop{\mathbb{E}}_{\substack{\forall h,\tau_h\in\mathcal{D}_h^{\mathrm{g}},\\ \omega_h\sim\mathbb{D}^{\rho}_{\theta^*}(\cdot|\tau_h)}}\left[\exp\left(\frac{1}{2}\sum_h\sum_{(\omega_h,\tau_h)\in\mathcal{D}_h^{\mathrm{g}}}\log\frac{\mathbb{D}^{\rho}_{\bar\theta}(\omega_h|\tau_h)}{\mathbb{D}^{\rho}_{\theta^*}(\omega_h|\tau_h)}-\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\log\mathop{\mathbb{E}}_{\omega_h\sim\mathbb{D}^{\rho}_{\theta^*}(\cdot|\tau_h)}\sqrt{\frac{\mathbb{D}^{\rho}_{\bar\theta}(\omega_h|\tau_h)}{\mathbb{D}^{\rho}_{\theta^*}(\omega_h|\tau_h)}}\right)\right]$$

$$=\frac{\mathop{\mathbb{E}}_{\substack{\forall h,\tau_h\in\mathcal{D}_h^{\mathrm{g}},\\ \omega_h\sim\mathbb{D}^{\rho}_{\theta^*}(\cdot|\tau_h)}}\left[\prod_h\prod_{(\omega_h,\tau_h)\in\mathcal{D}_h^{\mathrm{g}}}\sqrt{\frac{\mathbb{D}^{\rho}_{\bar\theta}(\omega_h|\tau_h)}{\mathbb{D}^{\rho}_{\theta^*}(\omega_h|\tau_h)}}\right]}{\prod_h\prod_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\mathbb{E}_{\omega_h\sim\mathbb{D}^{\rho}_{\theta^*}(\cdot|\tau_h)}\left[\sqrt{\frac{\mathbb{D}^{\rho}_{\bar\theta}(\omega_h|\tau_h)}{\mathbb{D}^{\rho}_{\theta^*}(\omega_h|\tau_h)}}\right]}=1,$$

where the last equality is due to the conditional independence of $\omega_h\in\mathcal{D}_h^{\mathrm{g}}$ given $\tau_h\in\mathcal{D}_h^{\mathrm{g}}$.

Therefore, by the Chernoff bound, with probability $1-\delta$, we have

$$-\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\log\mathop{\mathbb{E}}_{\omega_h\sim\mathbb{D}^{\rho}_{\theta^*}(\cdot|\tau_h)}\sqrt{\frac{\mathbb{D}^{\rho}_{\bar\theta}(\omega_h|\tau_h)}{\mathbb{P}^{\rho}_{\theta^*}(\omega_h|\tau_h)}}\le\frac{1}{2}\sum_h\sum_{(\omega_h,\tau_h)\in\mathcal{D}_h^{\mathrm{g}}}\log\frac{\mathbb{D}^{\rho}_{\theta^*}(\omega_h|\tau_h)}{\mathbb{D}^{\rho}_{\bar\theta}(\omega_h|\tau_h)}+\log\frac{1}{\delta}.$$

Taking the union bound over $\bar\Theta_\epsilon$, and rescaling δ, we have, with probability at least $1-\delta$, the following inequality holds:

$$\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\left\|\mathbb{D}^{\rho}_{\theta}(\omega_h|\tau_h)-\mathbb{D}^{\rho}_{\theta^*}(\omega_h|\tau_h)\right\|_1^2\le\frac{18}{N}+6\sum_h\sum_{(\omega_h,\tau_h)\in\mathcal{D}_h^{\mathrm{g}}}\log\frac{\mathbb{D}^{\rho}_{\theta^*}(\omega_h|\tau_h)}{\mathbb{D}^{\rho}_{\bar\theta}(\omega_h|\tau_h)}+12\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta}$$

$$\le 6\sum_h\sum_{(\omega_h,\tau_h)\in\mathcal{D}_h^{\mathrm{g}}}\log\frac{\mathbb{D}^{\rho}_{\theta^*}(\omega_h,\tau_h)}{\mathbb{D}^{\rho}_{\bar\theta}(\omega_h,\tau_h)}+6\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\log\frac{\mathbb{D}^{\rho}_{\bar\theta}(\tau_h)}{\mathbb{D}^{\rho}_{\theta^*}(\tau_h)}+13\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta}.$$

Note that, following from Lemma 2.2, with probability at least $1-\delta$, we have for any $k\in[K]$,

$$\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\log\frac{\mathbb{D}^{\rho}_{\bar\theta}(\tau_h)}{\mathbb{D}^{\rho}_{\theta^*}(\tau_h)}\le 3\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta}.$$

Hence, combining with the optimistic property of $\bar\theta$ and rescaling δ, we have that the following inequality holds with probability at least $1-\delta$:

$$\forall\theta\in\Theta_{\min},\quad\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\left\|\mathbb{D}^{\rho}_{\theta}(\omega_h|\tau_h)-\mathbb{D}^{\rho}_{\theta^*}(\omega_h|\tau_h)\right\|_1^2\le 6\sum_h\sum_{\tau_H\in\mathcal{D}_h^{k}}\log\frac{\mathbb{P}^{\rho}_{\theta^*}(\tau_H)}{\mathbb{P}^{\rho}_{\theta}(\tau_H)}+31\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta},$$

which yields the final result.□

Proposition 2.4: Fix $\varepsilon<\frac{1}{N^2}$. Define the following event:

$$\mathcal{E}_{\mathrm{H}}=\left\{\forall\theta\in\Theta,\quad|\mathcal{D}^{\mathrm{g}}|D_H^2(\mathbb{D}^{\rho}_{\theta}(\tau_H),\mathbb{D}^{\rho}_{\theta^*}(\tau_H))\le\frac{1}{2}\sum_{\tau_H\in\mathcal{D}^{\mathrm{g}}}\log\frac{\mathbb{D}^{\pi}_{\theta^*}(\tau_H)}{\mathbb{D}^{\rho}_{\theta}(\tau_H)}+2\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta}\right\}.$$

We have $\mathbb{P}(\mathcal{E}_\pi)\ge 1-\delta$.

Proof: First, by the construction of $\bar\Theta_\varepsilon$, for any θ, let $\bar\theta$ satisfy $\sum_{\tau_H}\left|\mathbb{D}^{\rho}_{\theta}(\tau_H)-\mathbb{D}^{\rho}_{\bar\theta}(\tau_H)\right|\le\varepsilon$. We translate the distance

$$\begin{aligned}&\left\|\mathbb{D}^{\rho}_{\theta}(\omega_h|\tau_h)-\mathbb{D}^{\rho}_{\theta'}(\omega_h|\tau_h)\right\|_1\\&=\sum_{\omega_h}\left|\frac{\mathbb{D}^{\pi}_{\theta'}(\omega_h,\tau_h)\mathbb{D}^{\rho}_{\theta}(\tau_h)-\mathbb{D}^{\rho}_{\theta}(\omega_h,\tau_h)\mathbb{D}^{\rho}_{\theta'}(\tau_h)}{\mathbb{D}^{\pi}_{\theta}(\tau_h)\mathbb{D}^{\rho}_{\theta'}(\tau_h)}\right|\\&=\sum_{\omega_h}\left|\frac{\left(\mathbb{D}^{\rho}_{\theta'}(\omega_h,\tau_h)-\mathbb{D}^{\rho}_{\theta}(\omega_h,\tau_h)\right)\mathbb{P}^{\rho}_{\theta}(\tau_h)+\mathbb{D}^{\rho}_{\theta}(\omega_h,\tau_h)\left(\mathbb{D}^{\pi}_{\theta}(\tau_h)-\mathbb{D}^{\rho}_{\theta'}(\tau_h)\right)}{\mathbb{D}^{\rho}_{\theta}(\tau_h)\mathbb{D}^{\rho}_{\theta'}(\tau_h)}\right|\\&\le\frac{|\mathbb{D}^{\rho}_{\theta}(\tau_h)-\mathbb{D}^{\pi}_{\theta'}(\tau_h)|}{\mathbb{D}^{\pi}_{\theta'}(\tau_h)}+\frac{1}{\mathbb{D}^{\rho}_{\theta'}(\tau_h)}\sum_{\omega_h}\left|\left(\mathbb{D}^{\rho}_{\theta'}(\omega_h,\tau_h)-\mathbb{D}^{\rho}_{\theta}(\omega_h,\tau_h)\right)\right|\\&\le\frac{2}{\mathbb{D}^{\rho}_{\theta'}(\tau_h)}\left\|\mathbb{D}^{\rho}_{\theta}-\mathbb{D}^{\rho}_{\theta'}\right\|_1.\end{aligned}$$

$$\begin{aligned}&\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\left\|\mathbb{D}^{\rho}_{\theta}(\omega_h|\tau_h)-\mathbb{D}^{\rho}_{\theta'}(\omega_h|\tau_h)\right\|_1^2\\&\overset{(a)}{\le}\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}2\left\|\mathbb{D}^{\rho}_{\theta}(\omega_h|\tau_h)-\mathbb{P}^{\rho}_{\bar\theta}(\omega_h|\tau_h)\right\|_1^2+2\left\|\mathbb{D}^{\rho}_{\bar\theta}(\omega_h|\tau_h)-\mathbb{D}^{\rho}_{\theta^*}(\omega_h|\tau_h)\right)\|_1^2\\&\le\frac{8}{N}+2\sum_h\sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}D_{TV}^2\left(\mathbb{P}^{\pi}_{\bar\theta}(\omega_h|\tau_h),\mathbb{P}^{\pi}_{\theta^*}(\omega_h|\tau_h)\right).\end{aligned}$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

between θ and θ^* to the distance between $\bar{\theta}$ and θ^* as follows, shown in the equation at the bottom of the page, where (a) follows because $\sqrt{a+b} \geq \sqrt{a} - \sqrt{|b|}$ if $a > 0$ and $a + b > 0$, and (b) follows from the Cauchy's inequality and the fact that $1 - x \leq -\log x$.

Hence, in order to upper bound $|\mathcal{D}^{\mathrm{g}}| D_H^2(\mathbb{D}_\theta^\rho, \mathbb{D}_{\theta^*}^\rho)$, it suffices to upper bound $-\log \mathbb{E}_{\tau_H \sim \mathbb{D}_{\theta^*}^\rho(\cdot)} \sqrt{\frac{\mathbb{D}_{\bar{\theta}}^\rho(\tau_H)}{\mathbb{D}_{\theta^*}^\rho(\tau_H)}}$. To this end, we observe that,

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\sum_{\tau_H \in \mathcal{D}^{\mathrm{g}}} \log \frac{\mathbb{D}_{\bar{\theta}}^\rho(\tau_H)}{\mathbb{D}_{\theta^*}^\pi(\tau_H)} - |\mathcal{D}^{\mathrm{g}}| \log \mathop{\mathbb{E}}_{\tau_H \sim \mathbb{D}_{\theta^*}^\rho(\cdot)} \sqrt{\frac{\mathbb{D}_{\theta}^\rho(\tau_H)}{\mathbb{D}_{\theta^*}^\rho(\tau_H)}}\right)\right] \overset{(a)}{=} \frac{\mathbb{E}\left[\prod_{\tau_H \in \mathcal{D}^{\mathrm{g}}} \sqrt{\frac{\mathbb{D}_{\bar{\theta}}^\rho(\tau_H)}{\mathbb{D}_{\theta^*}^\rho(\tau_H)}}\right]}{\mathbb{E}\left[\prod_{\tau_H \in \mathcal{D}^{\mathrm{g}}} \sqrt{\frac{\mathbb{D}_{\theta}^\rho(\tau_H)}{\mathbb{D}_{\theta^*}^\rho(\tau_H)}}\right]} = 1.$$

Then, by the Chernoff bound, we have

$$\mathbb{P}\left(\frac{1}{2}\sum_{\tau_H \in \mathcal{D}^{\mathrm{g}}} \log \frac{\mathbb{D}_{\bar{\theta}}^\rho(\tau_H)}{\mathbb{D}_{\theta^*}^\rho(\tau_H)} - |\mathcal{D}^{\mathrm{g}}| \log \mathop{\mathbb{E}}_{\tau_H \sim \mathbb{D}_{\theta^*}^\rho(\cdot)} \sqrt{\frac{\mathbb{D}_{\theta}^\rho(\tau_H)}{\mathbb{D}_{\theta^*}^\rho(\tau_H)}} \geq \log \frac{1}{\delta}\right) \leq \delta.$$

Finally, rescaling δ to $\delta/(\mathcal{N}_\varepsilon(\Theta))$ and taking the union bound over $\bar{\Theta}_\epsilon$, we conclude that, with probability at least $1 - \delta$,,

$$\begin{aligned}\forall \theta \in \Theta, \quad & |\mathcal{D}^{\mathrm{g}}| D_H^2(\mathbb{D}_\theta^\rho(\tau_H), \mathbb{D}_{\theta^*}^\rho(\tau_H)) \\ & \leq N\sqrt{\varepsilon} + \frac{1}{2}\sum_{\tau_H \in \mathcal{D}^{\mathrm{g}}} \log \frac{\mathbb{D}_{\theta^*}^\rho(\tau_H)}{\mathbb{D}_{\bar{\theta}}^\rho(\tau_H)} + \log \frac{\mathcal{N}_\varepsilon(\Theta)}{\delta} \\ & \overset{(a)}{\leq} \frac{1}{2}\sum_{\tau_H \in \mathcal{D}^{\mathrm{g}}} \log \frac{\mathbb{D}_{\theta^*}^\rho(\tau_H)}{\mathbb{D}_{\theta}^\rho(\tau_H)} + 2\log \frac{\mathcal{N}_\varepsilon(\Theta)}{\delta},\end{aligned}$$

where (a) follows because $\varepsilon \leq \frac{1}{N^2}$.□

Now, we are ready to provide the MLE guarantee of under the behavior policy ρ.

Lemma 2.5 (MLE guarantee): Let $p_{\min} < \delta/(|\mathcal{O}||\mathcal{A}|)^{2H}$. With probability at least $1 - \delta$, for any parameter θ satisfying $\mathcal{L}(\theta|\mathcal{D}) \leq \min_{\theta'} \mathcal{L}(\theta'|\mathcal{D}) + \beta/N$, we have the following inequalities.

$$\sum_h \sum_{\tau_h \in \mathcal{D}_h^{\mathrm{g}}} \left\| \mathbb{D}_\theta^\rho(\cdot|\tau_h) - \mathbb{D}_{\theta^*}^\rho(\cdot|\tau_h) \right\|_1^2 \leq 13\beta,$$

$$D_H^2\left(\mathbb{D}_\theta^\rho, \mathbb{D}_{\theta^*}^\rho\right) \leq \frac{4\beta}{N},$$

where $\beta_1 = 31 \log \frac{3\mathcal{N}_\varepsilon(\Theta)}{\delta}$.

Proof: Consider the three events $\mathcal{E}_{\log}$, $\mathcal{E}_\omega$, and $\mathcal{E}_{\mathrm{H}}$, defined in Lemma 2.2, Proposition 2.3, Proposition 2.4, respectively. Let $\mathcal{E}_o = \mathcal{E}_{\log} \cap \mathcal{E}_\omega \cap \mathcal{E}_{\mathrm{H}}$. Then, we have $\mathbb{P}(\mathcal{E}_o) \geq 1 - \delta$. The following proof is under the case when $\mathcal{E}_o$ happens.

First, due to the condition that $\mathcal{L}(\theta|\mathcal{D}) \leq \min_{\theta'} \mathcal{L}(\theta'|\mathcal{D}) + \beta/N$, we have

$$\begin{aligned}\sum_h \sum_{\tau_H \in \mathcal{D}_h^{\mathrm{g}}} & \log \frac{\mathbb{D}_{\theta^*}^\rho(\tau_H)}{\mathbb{D}_{\theta}^\rho(\tau_H)} \\ & = \sum_{\tau_H \in \mathcal{D}} \log \mathbb{D}_{\theta^*}^\rho(\tau_H) - \sum_{\tau_H \in \mathcal{D}} \log \mathbb{D}_{\theta}^\rho(\tau_H) \\ & \quad + \sum_{\tau_H \notin \mathcal{D}^{\mathrm{g}}} \log \mathbb{D}_{\theta}^\rho(\tau_H) - \sum_{\tau_H \notin \mathcal{D}^{\mathrm{g}}} \log \mathbb{D}_{\theta^*}^\rho(\tau_H) \\ & \leq \sum_{\tau_H \notin \mathcal{D}^{\mathrm{g}}} \log \mathbb{D}_{\theta}^\rho(\tau_H) - \sum_{\tau_H \notin \mathcal{D}^{\mathrm{g}}} \log \mathbb{D}_{\theta^*}^\rho(\tau_H) + \beta \\ & \overset{(a)}{\leq} \sum_{\tau_H \notin \mathcal{D}^{\mathrm{g}}} \log \mathbb{D}_{\bar{\theta}}^\rho(\tau_H) - \sum_{\tau_H \notin \mathcal{D}^{\mathrm{g}}} \log \mathbb{D}_{\theta^*}^\rho(\tau_H) + \beta \\ & \overset{(b)}{\leq} 2\beta,\end{aligned}$$

where (a) is due to the definition of $\bar{\theta}$, and (b) follows from Lemma 2.2.

Then, due to Proposition 2.3, we have

$$\begin{aligned}\sum_h \sum_{\tau_h \in \mathcal{D}_h^{\mathrm{g}}} & \left\| \mathbb{D}_\theta^\rho(\cdot|\tau_h) - \mathbb{D}_{\theta^*}^\rho(\cdot|\tau_h) \right\|_1^2 \\ & \leq 6\sum_h \sum_{\tau_H \in \mathcal{D}_h^{\mathrm{g}}} \log \frac{\mathbb{D}_{\theta^*}^\rho(\tau_H)}{\mathbb{D}_{\theta}^\rho(\tau_H)} + 31 \log \frac{\mathcal{N}_\varepsilon(\Theta)}{\delta} \\ & \leq 12\beta + \beta \\ & \leq 13\beta.\end{aligned}$$

$$\begin{aligned}D_H^2(\mathbb{D}_\theta^\rho, \mathbb{D}_{\theta^*}^\rho) & = 1 - \sum_{\tau_H} \sqrt{\mathbb{D}_\theta^\rho(\tau_H)\mathbb{D}_{\theta^*}^\rho(\tau_H)} \\ & = 1 - \sum_{\tau_H} \sqrt{\mathbb{D}_{\bar{\theta}}^\rho(\tau_H)\mathbb{D}_{\theta^*}^\rho(\tau_H) + \left(\mathbb{D}_\theta^\rho(\tau_H) - \mathbb{D}_{\bar{\theta}}^\rho(\tau_H)\right)\mathbb{D}_{\theta^*}^\rho(\tau_H)} \\ & \overset{(a)}{\leq} 1 - \sum_{\tau_H} \sqrt{\mathbb{D}_{\bar{\theta}}^\rho(\tau_H)\mathbb{D}_{\theta^*}^\rho(\tau_H)} + \sum_{\tau_H} \sqrt{\left|\mathbb{D}_\theta^\rho(\tau_H) - \mathbb{D}_{\bar{\theta}}^\rho(\tau_H)\right|\mathbb{D}_{\theta^*}^\rho(\tau_H)} \\ & \overset{(b)}{\leq} -\log \mathop{\mathbb{E}}_{\tau_H \sim \mathbb{D}_{\theta^*}^\rho(\cdot)} \sqrt{\frac{\mathbb{D}_{\bar{\theta}}^\rho(\tau_H)}{\mathbb{D}_{\theta^*}^\rho(\tau_H)}} + \sqrt{\sum_{\tau_H} \left|\mathbb{D}_\theta^\rho(\tau_H) - \mathbb{D}_{\bar{\theta}}^\rho(\tau_H)\right|} \\ & \leq -\log \mathop{\mathbb{E}}_{\tau_H \sim \mathbb{D}_{\theta^*}^\rho(\cdot)} \sqrt{\frac{\mathbb{D}_{\bar{\theta}}^\rho(\tau_H)}{\mathbb{D}_{\theta^*}^\rho(\tau_H)}} + \sqrt{\varepsilon},\end{aligned}$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

Similarly, due to Proposition 2.4

$$\begin{aligned} &D_H^2\left(\mathbb{D}_\theta^\rho, \mathbb{D}_{\theta^*}^\rho\right) \\ &\le \frac{1}{|\mathcal{D}^{\mathrm{g}}|}\left(\sum_h \sum_{\tau_H \in \mathcal{D}_h^{\mathrm{g}}} \frac{1}{2}\log\frac{\mathbb{D}_{\theta^*}^\rho(\tau_H)}{\mathbb{D}_\theta^\rho(\tau_H)} + 2\log\frac{\mathcal{N}_\varepsilon(\Theta)}{\delta}\right) \\ &\le \frac{1}{|\mathcal{D}^{\mathrm{g}}|}(\beta+\beta) \\ &\le 4\beta/N, \end{aligned}$$

where the last inequality follows from Lemma 2.1.□

Appendix C
Proof of Theorem 4.7

In this section, we provide the full proof for Theorem 4.7. Recall that $\mathcal{E}_o = \mathcal{E}_{\log} \cap \mathcal{E}_\omega \cap \mathcal{E}_{\mathrm{H}}$, where $\mathcal{E}_{\log}$, $\mathcal{E}_\omega$, and $\mathcal{E}_{\mathrm{H}}$, defined in Lemma 2.2, Proposition 2.3, Proposition 2.4, respectively.

Recall that we make the following assumption.

Assumption 3.1: The nominal model θ^* has rank r.

Thus, we have the following lemma.

Lemma 3.2 (Theorem C.1 in [68]) Any low-rank decision-making problem θ with core tests $\{\mathcal{Q}_h\}_{h=1}^H$ admits a (self-consistent) predictive state representation $\theta = \{\phi_h, M_h\}_{h=1}^H$ given core tests $\{\mathcal{Q}_h\}_{h=0}^{H-1}$, such that for any $\tau_h \in \mathcal{H}_h, \omega_h \in \Omega_h$, we have

$$\begin{cases} \psi(\tau_h) = M_h(o_h, a_h)\cdots M_1(o_1, a_1)\psi_0, \\ [\psi(\tau_h)]_\ell = \mathcal{P}^\theta(q_{h,\ell}^{\mathrm{o}}, \tau_h^{\mathrm{o}} | \tau_h^{\mathrm{a}}, q_{h,\ell}^{\mathrm{a}}), \\ m(\omega_h)^\top = \phi_H^\top M_H(o_H, a_H)\cdots M_{h+1}(o_{h+1}, a_{h+1}), \\ \sum\limits_{o_{h+1}} \phi_{h+1}^\top M_{h+1}(o_{h+1}, a_{h+1}) = \phi_h^\top, \\ \mathcal{P}^\theta(o_h, \ldots, o_1 | a_1, \ldots, a_h) = \phi_h^\top \psi(\tau_h), \end{cases}$$

where $M_h : \mathcal{O}\times\mathcal{A} \to \mathbb{R}^{d\times d}$, $\phi_h \in \mathbb{R}^d$, and $\psi_0 \in \mathbb{R}^d$. In particular, Θ is the set of all low-rank problems with the same core tests $\{\mathcal{Q}_h\}_{h=1}^H$.

First, we derive the following guarantee for the difference of robust values.

Corollary 3.3: Under the event $\mathcal{E}_o$, we have that

$$\begin{aligned} \left|V^\pi_{B^1_{\mathcal{P}}(\theta^*),R} - V^\pi_{B^1_{\mathcal{P}}(\hat\theta),R}\right| &\le V^\pi_{\hat\theta,\hat b}, \\ \left|V^\pi_{B^2_{\mathcal{P}}(\theta^*),R} - V^\pi_{B^2_{\mathcal{P}}(\hat\theta),R}\right| &\le C_{\mathcal{P}}^{\mathrm{kl}} V^\pi_{\hat\theta,\hat b}, \\ \left|V^\pi_{B^1_{\mathcal{T}}(\theta^*),R} - V^\pi_{B^1_{\mathcal{T}}(\hat\theta),R}\right| &\le 2C_B V^\pi_{\hat\theta,\hat b}, \\ \left|V^\pi_{B^2_{\mathcal{T}}(\theta^*),R} - V^\pi_{B^2_{\mathcal{T}}(\hat\theta),R}\right| &\le 2C_B C_{\mathcal{T}}^{\mathrm{kl}} V^\pi_{\hat\theta,\hat b}. \end{aligned}$$

Proof: The proof is finished by directly combining Lemma 1.6, Lemma 1.7, Lemma 1.8 and Lemma 4.8 and Corollary 3.2 in [21].□

Theorem 4.4 (Restatement of Theorem 4.7) Suppose Assumption 3.1 and Assumption 3.2 hold. The type-I coefficient of the behavior policy ρ against the robust policy π^* is finite. Let $Q_A = \max_h |\mathcal{Q}_h^A|$ be the maximum number of core action sequences. Let $\iota = \min_{q_{h,\ell}^{\mathrm{a}}} \rho(q_{h,\ell}^{\mathrm{a}})$, $p_{\min} = \frac{\delta}{N(|\mathcal{O}||\mathcal{A}|)^{2H}}$, $\varepsilon = \frac{p_{\min}}{NH}$, $\beta = O(\log(\mathcal{N}_\varepsilon(\Theta)/\delta))$, $\lambda = H^2Q_A^2$, and $\alpha = O\left(\frac{Q_A\sqrt{dH}}{\gamma^2}\sqrt{\lambda} + \frac{\sqrt{\beta}}{\iota\gamma}\right)$. Then, with probability at least $1-\delta$, the output $\hat\pi$ under different settings of Algorithm 1 satisfies that

$$\begin{aligned} V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\hat\pi}_{B^1_{\mathcal{P}}(\theta^*),R} &\lesssim \frac{H^2Q_A\sqrt{d\beta}}{\iota\gamma^2}\left(\sqrt{r} + \frac{Q_A\sqrt{H}}{\gamma}\right)\sqrt{\frac{C_{\mathrm{p}}(\pi^*|\rho)}{N}}, \\ V^{\pi^*}_{B^2_{\mathcal{P}}(\theta^*),R} - V^{\hat\pi}_{B^2_{\mathcal{P}}(\theta^*),R} &\lesssim \frac{H^2Q_A\sqrt{d\beta}}{\iota\gamma^2}\left(\sqrt{r} + \frac{Q_A\sqrt{H}}{\gamma}\right) C_{\mathcal{P}}^{\mathrm{kl}}\sqrt{\frac{C_{\mathrm{p}}(\pi^*|\rho)}{N}}, \\ V^{\pi^*}_{B^1_{\mathcal{T}}(\theta^*),R} - V^{\hat\pi}_{B^1_{\mathcal{T}}(\theta^*),R} &\lesssim \frac{H^2Q_A\sqrt{d\beta}}{\iota\gamma^2}\left(\sqrt{r} + \frac{Q_A\sqrt{H}}{\gamma}\right) C_B\sqrt{\frac{C_{\mathrm{p}}(\pi^*|\rho)}{N}}, \\ V^{\pi^*}_{B^2_{\mathcal{T}}(\theta^*),R} - V^{\hat\pi}_{B^2_{\mathcal{T}}(\theta^*),R} &\lesssim \frac{H^2Q_A\sqrt{d\beta}}{\iota\gamma^2}\left(\sqrt{r} + \frac{Q_A\sqrt{H}}{\gamma}\right) C_B C_{\mathcal{T}}^{\mathrm{kl}}\sqrt{\frac{C_{\mathrm{p}}(\pi^*|\rho)}{N}}. \end{aligned}$$

Proof: We first prove the result for $\mathcal{P}$-type uncertainty set with ℓ_1 distance.

Recall that $\hat\pi = \arg\max_\pi \left\{V^\pi_{B^1_{\mathcal{P}}(\hat\theta),R} - V^\pi_{\hat\theta,\hat b}\right\}$. Let $\pi^* = \arg\max_\pi V^\pi_{B^1_{\mathcal{P}}(\theta^*),R}$. Then, by utilizing Corollary 3.3, the sub-optimal gap of $\hat\pi$ can be upper bounded as follows.

$$\begin{aligned} &V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\hat\pi}_{B^1_{\mathcal{P}}(\theta^*),R} \\ &\quad\le V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - \left(V^{\pi^*}_{B^1_{\mathcal{P}}(\hat\theta),R} - V^{\pi^*}_{\hat\theta,\hat b}\right) \\ &\qquad + \left(V^{\pi^*}_{B^1_{\mathcal{P}}(\hat\theta),R} - V^{\pi^*}_{\hat\theta,\hat b}\right) - \left(V^{\hat\pi}_{B^1_{\mathcal{P}}(\hat\theta),R} - V^{\hat\pi}_{\hat\theta,\hat b}\right) \\ &\quad\le 2V^{\pi^*}_{\hat\theta,\hat b}. \end{aligned}$$

By Lemma 4.6 and Lemma 5.1 in [21] we have

$$\begin{aligned} V^{\pi^*}_{\hat\theta,\hat b} &\le V^{\pi^*}_{\theta^*,\hat b} + \left\|\mathbb{D}_{\theta^*}^{\pi^*} - \mathbb{D}_{\hat\theta}^{\pi^*}\right\|_1 \\ &\overset{(a)}{\lesssim} \alpha\frac{\sqrt{r\beta}}{\iota\sqrt{\lambda}}\sum_{h=0}^{H-1}\mathbb{E}_{\theta^*}^\pi\left[\left\|\bar\psi_h^*(\tau_h)\right\|_{U_h^{-1}}\right] \\ &\quad + \frac{\alpha HQ_A}{\sqrt{\lambda}}\left\|\mathbb{D}_{\theta^*}^\pi - \mathbb{D}_{\hat\theta}^\pi\right\|_1 \\ &\lesssim \alpha\frac{\sqrt{r\beta}}{\iota\sqrt{\lambda}}\sum_{h=0}^{H}\mathbb{E}_{\theta^*}^\pi\left[\left\|\bar\psi^*(\tau_h)\right\|_{\bar U_h^{-1}}\right] \\ &\quad + \frac{\alpha HQ_A}{\sqrt{\lambda}}\sqrt{\frac{\beta}{H\iota^2\gamma^2}}\sum_{h=0}^{H}\mathbb{E}_{\theta^*}^\pi\left[\left\|\bar\psi^*(\tau_h)\right\|_{\Lambda_h^{-1}}\right], \end{aligned}$$

where (a) is due to $\lambda < H^2Q_A^2$, $\alpha = \sqrt{\frac{4\lambda HQ_A^2 d}{\gamma^4} + \frac{7\beta}{\iota^2\gamma^2}}$, and

$$\begin{aligned} \hat U_h &= \lambda I + \sum_{\tau_h\in\mathcal{D}_h^{\mathrm{g}}}\bar\psi^*(\tau_h)\bar\psi^*(\tau_h)^\top, \\ \bar U_h &= \lambda I + |\mathcal{D}_h^{\mathrm{g}}|\mathbb{E}_{\theta^*}^\rho\left[\bar\psi^*(\tau_h)\bar\psi^*(\tau_h)^\top\right], \\ \Lambda_h &= \lambda_0 I + \frac{N}{2H}\mathbb{E}_{\theta^*}^\rho\left[\bar\psi^*(\tau_h)\bar\psi^*(\tau_h)^\top\right], \\ \lambda_0 &= \frac{\gamma^4}{4Q_A^2 d}. \end{aligned}$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

Recall that type-I concentrability coefficient implies that

$$\mathbb{E}^{\pi^*}_{\theta^*}\left[\bar{\psi}^*(\tau_h)\bar{\psi}^*(\tau_h)^\top\right] \le C_{\mathrm{p}}(\pi^*|\rho)\mathbb{E}^{\rho}_{\theta^*}\left[\bar{\psi}^*(\tau_h)\bar{\psi}^*(\tau_h)^\top\right].$$

Thus, due to Cauchy's inequality, we have

$$\begin{aligned}
&V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\theta^*),R} \\
&\lesssim \alpha\frac{\sqrt{\beta}}{\iota\sqrt{\lambda}}\left(\sqrt{r}+\frac{Q_A\sqrt{H}}{\gamma}\right) \\
&\quad \sum_h \sqrt{C_{\mathrm{p}}(\pi^*|\rho)\mathbb{E}^{\rho}_{\theta^*}\left[\left\|\bar{\psi}^*(\tau_h)\right\|^2_{\Lambda_h^{-1}}\right]} \\
&\le \alpha H\frac{\sqrt{\beta}}{\iota\sqrt{\lambda}}\left(\sqrt{r}+\frac{Q_A\sqrt{H}}{\gamma}\right)\sqrt{\frac{HC_{\mathrm{p}}(\pi^*|\rho)}{N}} \\
&\lesssim \frac{H^2Q_A\sqrt{d\beta}}{\iota\gamma^2}\left(\sqrt{r}+\frac{Q_A\sqrt{H}}{\gamma}\right)\sqrt{\frac{C_{\mathrm{p}}(\pi^*|\rho)}{N}}.
\end{aligned}$$

For the other three inequalities, we follow the same argument as above and the proof is finished.□

Appendix D
Proof of Theorem 5.2

We first provide the pseudo code for our proposed algorithm.

Algorithm 2 Robust Offline RL

1: **Input:** offline dataset $\mathcal{D}$
2: Construct confidence set $\mathcal{C}$ according to Equation (8).
3: Planning $\hat{\pi}$ such that $\hat{\pi} = \arg\max_{\pi\in\Pi}\min_{\hat{\theta}\in\mathcal{C}} V^{\pi}_{B(\hat{\theta}),R}$.

Theorem 4.1: Suppose the type-II coefficient of the behavior policy ρ against the robust policy π^* is finite. Let $\varepsilon = \frac{1}{NH}$, $\beta = O(\log(\mathcal{N}_\varepsilon(\Theta)/\delta))$. Then, with probability at least $1-\delta$, the output $\hat{\pi}$ under different settings of Algorithm 2 satisfies that

$$\begin{aligned}
V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\theta^*),R} &\lesssim H\sqrt{\frac{C(\pi^*|\rho)\beta}{N}}, \\
V^{\pi^*}_{B^2_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^2_{\mathcal{P}}(\theta^*),R} &\lesssim HC^{\mathrm{kl}}_{\mathcal{P}}\sqrt{\frac{C(\pi^*|\rho)\beta}{N}}, \\
V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\theta^*),R} &\lesssim HC_B\sqrt{\frac{C(\pi^*|\rho)\beta}{N}}, \\
V^{\pi^*}_{B^2_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^2_{\mathcal{P}}(\theta^*),R} &\lesssim HC_BC^{\mathrm{kl}}_{\mathcal{P}}\sqrt{\frac{C(\pi^*|\rho)\beta}{N}}.
\end{aligned}$$

Proof: We first prove the result for the setting when the uncertainty set is $\mathcal{P}$-type with ℓ_1 distance.

Let $\hat{\theta} = \arg\min_{\theta\in\mathcal{C}} V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\theta),R}$. Then, we perform the decomposition as follows.

$$\begin{aligned}
&V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\theta^*),R} \\
&= V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\pi^*}_{B^1_{\mathcal{P}}(\hat{\theta}),R} + \underbrace{V^{\pi^*}_{B^1_{\mathcal{P}}(\hat{\theta}),R} - V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\hat{\theta}),R}}_{I_1} \\
&\quad + \underbrace{V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\hat{\theta}),R} - V^{\hat{\pi}}_{B^1_{\mathcal{P}}(\theta^*),R}}_{I_2} \\
&\le V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\pi^*}_{B^1_{\mathcal{P}}(\hat{\theta}),R},
\end{aligned}$$

where I_1 is due to the optimality $\hat{\pi}$, and I_2 follows from the property of $\hat{\theta}$.

By Lemma 1.6, we have

$$\begin{aligned}
&V^{\pi^*}_{B^1_{\mathcal{P}}(\theta^*),R} - V^{\pi^*}_{B^1_{\mathcal{P}}(\hat{\theta}),R} \\
&\le \left\|\mathbb{D}^{\pi^*}_{\theta^*} - \mathbb{D}^{\pi^*}_{\hat{\theta}}\right\|_1 \\
&\overset{(a)}{\le} \sum_{h=1}^H \mathbb{E}^{\pi^*}_{\theta^*}\left[\left\|\mathcal{T}^{\theta^*}(\cdot|\tau_h) - \mathcal{T}^{\hat{\theta}}(\cdot|\tau_h)\right\|_1\right] \\
&\overset{(b)}{\le} \sum_{h=1}^H \sqrt{C(\pi^*|\rho)\mathbb{E}^{\rho}_{\theta^*}\left[\left\|\mathcal{T}^{\theta^*}(\cdot|\tau_h) - \mathcal{T}^{\hat{\theta}}(\cdot|\tau_h)\right\|_1^2\right]} \\
&\overset{(c)}{\lesssim} \sum_{h=1}^H \sqrt{C(\pi^*|\rho)\mathbb{E}^{\rho}_{\theta^*}\left[D^2_{\mathrm{H}}\left(\mathcal{T}^{\theta^*}(\cdot|\tau_h), \mathcal{T}^{\hat{\theta}}(\cdot|\tau_h)\right)\right]} \\
&\overset{(d)}{\lesssim} H\sqrt{C(\pi^*|\rho)D^2_{\mathrm{H}}\left(\mathbb{D}^{\rho}_{\theta^*}, \mathbb{D}^{\rho}_{\hat{\theta}}\right)} \\
&\le H\sqrt{\frac{\beta C(\pi^*|\rho)}{N}},
\end{aligned}$$

where (a) follows from Lemma 5.1, (b) is due to the definition of type-II concentrability coefficient, and (c), (d) follows from Lemma 5.2.

The other three results follows the same argument by utilizing Lemma 1.7 and Lemma 1.8.□

Appendix E
Auxiliary Lemmas

Lemma 5.1: For two distributions P and Q, we have

$$\mathbb{E}_{X\sim P_X}\left[\|P_Y(\cdot|X) - Q_Y(\cdot|X)\|_1\right] \le 2\|P_{X,Y} - Q_{X,Y}\|_1.$$

In addition, the following inequality holds.

$$\begin{aligned}
\|P_{X,Y} - Q_{X,Y}\|_1 \le \mathbb{E}_P[\|P_{Y|X}(\cdot|X) - Q_{Y|X}(\cdot|X)\|_1] \\
+ \|P_X - Q_X\|_1.
\end{aligned}$$

Proof: We have

$$\begin{aligned}
&\mathbb{E}_{X\sim P_X}\left[\|P_Y(\cdot|X) - Q_Y(\cdot|X)\|_1\right] \\
&= \sum_x P_X(x)\sum_y |P_{Y|X}(y|x) - Q_{Y|X}(y|x)| \\
&= \sum_x\sum_y |P_{Y|X}(y|x)P_X(x) - Q_{Y|X}(y|x)P_X(x)| \\
&= \sum_x\sum_y |P_{Y|X}(y|x)P_X(x) - Q_{Y|X}(y|x)Q_X(x) \\
&\quad + Q_{Y|X}(y|x)Q_X(x) - Q_{Y|X}(y|x)P_X(x)| \\
&\le \sum_x\sum_y |P_{Y|X}(y|x)P_X(x) - Q_{Y|X}(y|x)Q_X(x)| \\
&\quad + \sum_x\sum_y |Q_{Y|X}(y|x)Q_X(x) - Q_{Y|X}(y|x)P_X(x)| \\
&= \|P_{X,Y} - Q_{X,Y}\|_1 + \sum_x |P_X(x) - Q_X(x)| \\
&= \|P_{X,Y} - Q_{X,Y}\|_1 + \sum_x |\sum_y P_{Y|X}(y|x)P_X(x) \\
&\quad - \sum_y Q_{Y|X}(y|x)Q_X(x)|
\end{aligned}$$

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

$$\leq 2\|P_{X,Y} - Q_{X,Y}\|_1.$$

On the other hand,

$$\begin{aligned}
&\|P_{X,Y} - Q_{X,Y}\|_1 \\
&\quad = \sum_x \sum_y |P_{Y|X}(y|x)P_X(x) - Q_{Y|X}(y|x)Q_X(x)| \\
&\quad = \sum_x \sum_y |P_{Y|X}(y|x)P_X(x) - Q_{Y|X}(y|x)P_X(x) \\
&\qquad + Q_{Y|X}(y|x)P_X(x) - Q_{Y|X}(y|x)Q_X(x)| \\
&\quad \leq \mathbb{E}_P[\|P_{Y|X}(\cdot|X) - Q_{Y|X}(\cdot|X)\|_1] + \|P_X - Q_X\|_1.
\end{aligned}$$

□

Lemma 5.2: Given two bounded measures P and Q defined on the set $\mathcal{X}$. Let $|P| = \sum_{x\in\mathcal{X}} P(x)$ and $|Q| = \sum_{x\in\mathcal{X}} Q(x)$. We have

$$\|P - Q\|_1^2 \leq 4(|P| + |Q|)D_H^2(P, Q)$$

In addition, if $P_{Y|X}, Q_{Y|X}$ are two conditional distributions over a random variable Y, and $P_{X,Y} = P_{Y|X}P$, $Q_{X,Y} = Q_{Y|X}Q$ are the joint distributions when X follows the distributions P and Q, respectively, we have

$$\mathop{\mathbb{E}}_{X\sim P}\left[D_H^2(P_{Y|X}, Q_{Y|X})\right] \leq 8D_H^2(P_{X,Y}, Q_{X,Y}).$$

Proof: We first prove the first inequality. By the definition of total variation distance, we have

$$\begin{aligned}
\|P - Q\|_1^2 &= \left(\sum_x |P(x) - Q(x)|\right)^2 \\
&= \left(\sum_x \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)\left(\sqrt{P(x)} + \sqrt{Q(x)}\right)\right)^2 \\
&\overset{(a)}{\leq} \left(\sum_x \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2\right) \\
&\quad \left(2\sum_x (P(x) + Q(x))\right) \\
&\leq 4(|P| + |Q|)D_H^2(P, Q),
\end{aligned}$$

where (a) follows from the Cauchy's inequality and because $(a + b)^2 \leq 2a^2 + 2b^2$.

For the second inequality, we have,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim P}\left[\mathrm{D}_{\mathrm{H}}^2\left(P_{Y|X}, Q_{Y|X}\right)\right] \\
&= \sum_x P(x)\left(\sum_y \left(\sqrt{P_{Y|X}(y)} - \sqrt{Q_{Y|X}(y)}\right)^2\right) \\
&= \sum_{x,y} \Big(\sqrt{P_{X,Y}(x,y)} - \sqrt{Q_{X,Y}(x,y)} \\
&\quad + \sqrt{Q_{Y|X}(y)Q(x)} - \sqrt{Q_{Y|X}(y)P(x)}\Big)^2 \\
&\leq 2\sum_{x,y} \left(\sqrt{P_{X,Y}(x,y)} - \sqrt{Q_{X,Y}(x,y)}\right)^2 \\
&\quad + 2\sum_{x,y} Q_{Y|X}(y)\left(\sqrt{Q(x)} - \sqrt{P(x)}\right)^2 \\
&= 4\mathrm{D}_{\mathrm{H}}^2\left(P_{X,Y}, Q_{X,Y}\right) + 2\left(|P| + |Q| - 2\sum_x \sqrt{P(x)Q(x)}\right) \\
&\overset{(a)}{=} 4\mathrm{D}_{\mathrm{H}}^2\left(P_{X,Y}, Q_{X,Y}\right) + 2(|P| + |Q| \\
&\quad - 2\sum_x \sum_y \sqrt{P_{Y|X}(y)P(x)Q_{Y|X}(y)Q(x)}\Big) \\
&= 8\mathrm{D}_{\mathrm{H}}^2\left(P_{X,Y}, Q_{X,Y}\right),
\end{aligned}$$

where (a) follows from the Cauchy's inequality that applies on $\sum_y \sqrt{P_{Y|X}(y)Q_{Y|X}(y)}$.□

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[2] D. Ye et al., "Mastering complex control in MOBA games with deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 6672–6679.

[3] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning: Lessons we have learned," *Int. J. Robot. Res.*, vol. 40, nos. 4–5, pp. 698–721, Apr. 2021.

[4] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 27730–27744.

[5] S. Lange, T. Gabel, and M. Riedmiller, "Batch reinforcement learning," in *Reinforcement Learning: State-of-the-Art*, vol. 12. Berlin, Germany: Springer, 2012, pp. 45–73.

[6] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020, *arXiv:2005.01643*.

[7] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Jun. 2018, pp. 3803–3810.

[8] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: A survey," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2020, pp. 737–744.

[9] G. Iyengar, "Robust dynamic programming," *Math. Oper. Res.*, vol. 30, no. 2, pp. 257–280, May 2005.

[10] H. Xu and S. Mannor, "Distributionally robust Markov decision processes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, Apr. 2012, pp. 288–300.

[11] W. Wiesemann, D. Kühn, and B. Rüstem, "Robust Markov decision processes," *Math. Oper. Res.*, vol. 38, no. 1, pp. 153–183, Nov. 2012.

[12] P. Yu and H. Xu, "Distributionally robust counterpart in Markov decision processes," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2538–2543, Sep. 2016.

[13] K. Panaganti and D. Kalathil, "Sample complexity of robust reinforcement learning with a generative model," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2022, pp. 9582–9602.

[14] Y. Wang and S. Zou, "Online robust reinforcement learning with model uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 7193–7206.

[15] J. Dong, J. Li, B. Wang, and J. Zhang, "Online policy optimization for robust MDP," 2022, *arXiv:2209.13841*.

[16] K. Panaganti, Z. Xu, D. Kalathil, and M. Ghavamzadeh, "Robust reinforcement learning using offline data," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 32211–32224.

[17] J. Blanchet, M. Lu, T. Zhang, and H. Zhong, "Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage," 2023, *arXiv:2305.09659*.

[18] T. Osogami, "Robust partially observable Markov decision process," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 106–115.

[19] M. Rasouli and S. Saghafian, "Robust partially observable Markov decision processes," John F. Lemmedy School Government, Harvard Univ., Work. Paper Ser., 2018.

[20] H. Nakao, R. Jiang, and S. Shen, "Distributionally robust partially observable Markov decision process with moment-based ambiguity," *SIAM J. Optim.*, vol. 31, no. 1, pp. 461–488, Jan. 2021.

[21] R. Huang, Y. Liang, and J. Yang, "Provably efficient UCB-type algorithms for learning predictive state representations," 2023, *arXiv:2307.00405*.

[22] A. Antos, C. Szepesvári, and R. Munos, "Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path," *Mach. Learn.*, vol. 71, no. 1, pp. 89–129, Apr. 2008.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.

[23] D. P. Bertsekas, "Approximate policy iteration: A survey and some new methods," *J. Control Theory Appl.*, vol. 9, no. 3, pp. 310–335, Aug. 2011.

[24] J. Chen and N. Jiang, "Information-theoretic considerations in batch reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1042–1051.

[25] T. Xie and N. Jiang, "Q* approximation schemes for batch reinforcement learning: A theoretical comparison," in *Proc. Conf. Uncertainty Artif. Intell.*, Jan. 2020, pp. 550–559.

[26] T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal, "Bellman-consistent pessimism for offline reinforcement learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 6683–6694.

[27] Y. Jin, Z. Yang, and Z. Wang, "Is pessimism provably efficient for offline RL," in *Proc. Int. Conf. Mach. Learn.*, Dec. 2020, pp. 5084–5096.

[28] G. J. Gordon, "Stable function approximation in dynamic programming," in *Machine Learning Proceedings 1995*. Amsterdam, The Netherlands: Elsevier, 1995, pp. 261–268.

[29] D. Ernst, P. Geurts, and L. Wehenkel, "Tree-based batch mode reinforcement learning," *J. Mach. Learn. Res.*, vol. 6, no. 18, pp. 503–556, Dec. 2005.

[30] R. Munos and C. Szepesvári, "Finite-time bounds for fitted value iteration," *J. Mach. Learn. Res.*, vol. 9, no. 27, pp. 815–857, Jun. 2008.

[31] A. Farahmand, C. Szepesvári, and R. Munos, "Error propagation for approximate policy and value iteration," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23, Jan. 2010.

[32] A. Lazaric, M. Ghavamzadeh, and R. Munos, "Finite-sample analysis of least-squares policy iteration," *J. Mach. Learn. Res.*, vol. 13, pp. 3041–3074, Oct. 2012.

[33] Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill, "Provably good batch off-policy reinforcement learning without great exploration," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1264–1274.

[34] T. Nguyen-Tang, M. Yin, S. Gupta, S. Venkatesh, and R. Arora, "On instance-dependent bounds for offline reinforcement learning with linear function approximation," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 8, pp. 9310–9318.

[35] W. Xiong, H. Zhong, C. Shi, C. Shen, L. Wang, and T. Zhang, "Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent MDP and Markov game," 2022, *arXiv:2205.15512*.

[36] L. Zhou, K. Small, O. Rokhlenko, and C. Elkan, "End-to-end offline goal-oriented dialog policy learning via policy gradient," 2017, *arXiv:1712.02838*.

[37] M. Uehara and W. Sun, "Pessimistic model-based offline reinforcement learning under partial coverage," 2021, *arXiv:2107.06226*.

[38] M. Uehara, X. Zhang, and W. Sun, "Representation learning for online and offline RL in low-rank MDPs," 2021, *arXiv:2110.04652*.

[39] R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims, "MOReL: Model-based offline reinforcement learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21810–21823.

[40] T. Yu et al., "MOPO: Model-based offline policy optimization," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 14129–14142.

[41] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1179–1191.

[42] J. Chang, M. Uehara, D. Sreenivas, R. Kidambi, and W. Sun, "Mitigating covariate shift in imitation learning via offline data with partial coverage," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 965–979.

[43] M. Yin, Y. Bai, and Y.-X. Wang, "Near-optimal offline reinforcement learning via double variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 7677–7688.

[44] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell, "Bridging offline reinforcement learning and imitation learning: A tale of pessimism," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 11702–11716.

[45] A. Zanette, M. J. Wainwright, and E. Brunskill, "Provable benefits of actor-critic methods for offline reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 13626–13640.

[46] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Oper. Res.*, vol. 53, no. 5, pp. 780–798, Oct. 2005.

[47] A. Tamar, S. Mannor, and H. Xu, "Scaling up robust MDPs using function approximation," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 181–189.

[48] S. Mannor, O. Mebel, and H. Xu, "Robust MDPs with k-rectangular uncertainty," *Math. Oper. Res.*, vol. 41, no. 4, pp. 1484–1509, Nov. 2016.

[49] M. Petrik and R. H. Russell, "Beyond confidence regions: Tight Bayesian ambiguity sets for robust MDPs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2019.

[50] Y. Wang, A. Velasquez, G. K. Atia, A. Prater-Bennette, and S. Zou, "Robust average-reward Markov decision processes," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 12, pp. 15215–15223.

[51] Y. Wang, A. Velasquez, G. K. Atia, A. Prater-Bennette, and S. Zou, "Model-free robust average-reward reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2023, pp. 36431–36469.

[52] Z. Zhou, Z. Zhou, Q. Bai, L. Qiu, J. Blanchet, and P. W. Glynn, "Finite-sample regret bound for distributionally robust offline tabular reinforcement learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, Mar. 2021, pp. 3331–3339.

[53] W. Yang, L. Zhang, and Z. Zhang, "Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics," *Ann. Statist.*, vol. 50, no. 6, pp. 3223–3248, Dec. 2022.

[54] A. Roy, H. Xu, and S. Pokutta, "Reinforcement learning under model mismatch," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jan. 2017.

[55] S. H. Lim, H. Xu, and S. Mannor, "Reinforcement learning in robust Markov decision processes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, Dec. 2013, pp. 701–709.

[56] X. Ni and L. Lai, "Robust risk-sensitive reinforcement learning with conditional value-at-risk," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2024, pp. 520–525.

[57] R. Cipollone, A. Jonsson, A. Ronca, and M. S. Talebi, "Provably efficient offline reinforcement learning in regular decision processes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 39395–39428.

[58] E. Mossel and S. Roch, "Learning nonsingular phylogenies and hidden Markov models," in *Proc. 37th Annu. ACM Symp. Theory Comput.*, May 2005, pp. 366–375.

[59] M. Mundhenk, J. Goldsmith, C. Lusena, and E. Allender, "Complexity of finite-horizon Markov decision process problems," *J. ACM*, vol. 47, no. 4, pp. 681–720, Jul. 2000.

[60] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of Markov decision processes," *Math. Oper. Res.*, vol. 12, no. 3, pp. 441–450, Aug. 1987.

[61] N. Vlassis, M. L. Littman, and D. Barber, "On the computational complexity of stochastic controller optimization in POMDPs," *ACM Trans. Comput. Theory*, vol. 4, no. 4, pp. 1–8, Nov. 2012.

[62] A. Krishnamurthy, A. Agarwal, and J. Langford, "PAC reinforcement learning with rich observations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, Dec. 2016, pp. 1848–1856.

[63] B. Boots, S. M. Siddiqi, and G. J. Gordon, "Closing the learning-planning loop with predictive state representations," *Int. J. Robot. Res.*, vol. 30, no. 7, pp. 954–966, Jun. 2011.

[64] N. Jiang, A. Kulesza, and S. Singh, "Completing state representations using spectral learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Jan. 2018, pp. 4328–4337.

[65] Z. Zhang, Z. Yang, H. Liu, P. Tokekar, and F. Huang, "Reinforcement learning under a multi-agent predictive state representation model: Method and theory," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, 2022.

[66] W. Zhan, M. Uehara, W. Sun, and J. D. Lee, "PAC reinforcement learning for predictive state representations," 2022, *arXiv:2207.05738*.

[67] M. Uehara, A. Sekhari, J. D. Lee, N. Kallus, and W. Sun, "Provably efficient reinforcement learning in partially observable dynamical systems," 2022, *arXiv:2206.12020*.

[68] Q. Liu, P. Netrapalli, C. Szepesvári, and C. Jin, "Optimistic MLE—A generic model-based algorithm for partially observable sequential decision making," 2022, *arXiv:2209.14997*.

[69] F. Chen, Y. Bai, and S. Mei, "Partially observable RL with B-stability: Unified structural condition and sharp sample-efficient algorithms," 2022, *arXiv:2209.14990*.

[70] H. Zhong et al., "GEC: A unified framework for interactive decision making in MDP, POMDP, and beyond," 2022, *arXiv:2211.01962*.

[71] M. L. Littman and R. S. Sutton, "Predictive representations of state," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, Jan. 2001, pp. 1555–1561.

[72] S. Singh, M. James, and M. Rudary, "Predictive state representations: A new theory for modeling dynamical systems," 2012, *arXiv:1207.4167*.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 16,2026 at 02:27:32 UTC from IEEE Xplore. Restrictions apply.