Robust Constrained Offline Reinforcement Learning with Linear Function Approximation

Wenbin Wang[*]
School of Information Science and Technology
ShanghaiTech University
Shanghai, China
wangwb2023@shanghaitech.edu.cn

He Wang[*]
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213, USA
wanghe@cmu.edu

Abstract

Bridging the sim-to-real gap requires reinforcement learning policies that achieve not only high rewards but also safety and robustness under distribution shifts. Yet the high-dimensionality of the state-action space makes learning sample-inefficient. To this end, we study *robust constrained linear Markov decision processes (Lin-RCMDPs)* in the offline setting, where an agent seeks to maximize expected return while satisfying *safety constraints* against the *worst-case dynamics* drawn from an ambiguity set defined by a total-variation ball. We propose a sample-efficient, model-based primal-dual algorithm `CROP-VI` that integrates robust planning with rectified Lagrangian updates to ensure constraint feasibility across all transitions in the ambiguity set. Specifically, we introduce pessimism into the reward function to prevent over-estimation, and apply asymmetric optimism to constraint to balance exploration-exploitation trade-off. Under mild data-coverage assumptions, we establish the *first* instance-dependent sub-optimality bound of `CROP-VI` for Lin-RCMDPs, where the learned policy is not only feasible with respect to the worst-case model but achieves near-optimal robust return. We further establish the sample-complexity bound of `CROP-VI` under partial or full feature coverage data, and extend the analysis beyond the linear MDP idealization to a misspecified regime, showing performance degrades gracefully with approximation error.

1 Introduction

Reinforcement learning (RL) seeks policies that maximize the expected cumulative reward through interactions with the environment [1]. In many high-stakes domains—autonomous driving [2–5], healthcare [6–8], wireless security [9–11], and robotics [12–14], agents must also obey resources and safety constraints. Examples include avoiding collision for robots [15, 16], cost budgets in medical decision making [17, 18], and adhering to regulatory limits in finance [19]. Constrained Markov Decision Processes (CMDPs) address such requirements by optimizing utility while ensuring policy-level constraint satisfaction [20–26].

[*] Equal contribution.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

	State Representation	Blanchet et al. [36]	Wang et al. [40]	Ghosh [25]	**This Work**
unconstraint	$S \times A$-rectangular (Tabular)	✓	✓	✓	✓
	d-rectangular (Linear)	✓	✓	✗	✓
constraint	$S \times A$-rectangular (Tabular)	✗	✗	✓!	✓
	d-rectangular (Linear)	✗	✗	✗	✓

Table 1: Comparison with the most relevant works in robust RL. ✓ indicates that the work is capable of addressing the model with robust partial coverage data, ✓! signifies that the work requires full coverage data to solve the model, and ✗ denotes that the work is not applicable to the model. Light green highlights the models that are either introduced or proven to be tractable in this work.

Many CMDP algorithms learn constraints and rewards through online interaction [27–29]. However, online data collection can be expensive, slow, or unsafe in real-world systems. This motivates offline learning, where a policy is learned solely from logged experience. Offline RL without constraints has seen substantial progress in both practice and theory [30–36], yet a standard assumption persists: the deployment environment matches the training one. In practice, even a mild distribution shift can severely degrade performance. Thus, there is an urgent need for robust policies that remain effective and safe under environmental uncertainty.

Distributionally robust constrained RL tackles this challenge by optimizing the worst-case performance over an uncertainty set inferred from historical data [21, 25, 37], Existing results, however, largely restrict attention (i) online interaction [21, 25, 37] or (ii) offline data with tabular settings where sample complexity scales with the size of the state-action space [25], rendering them impractical for high-dimensional problems. This raises a central question:

> *Can we design a provably sample-efficient offline RL method with linear function approximation that is both distributionally robust and safe?*

1.1 Contribution

We answer this question affirmatively by studying finite-horizon distributionally robust constrained linear MDPs (Lin-RCMDPs). Our uncertainty sets are grounded in practice [38] and recent theory [39, 40]. Concretely:

- **Safe and robust algorithm design.** We propose `CROP-VI`, a safe and distributionally robust variant of offline least-squares value iteration for MDPs that admit linear representation and satisfy safety constraints. We carefully design a data-driven robust penalty to mitigate the scarcity and covariate shift of offline data, and we optimistically incorporate this penalty into the constraint to balance the exploration-exploitation trade-off.
- **Theoretical guarantees of sample efficiency.** Under the partial feature coverage assumption of offline data, `CROP-VI` returns an ϵ-optimal robust policy whose constraint violation is at most ϵ provided the sample size satisfies $\widetilde{\mathcal{O}}(C^{\star}_{\mathrm{rob}} d^2 H^4/\epsilon^2)$ (cf. Corollary 1), where $C^{\star}_{\mathrm{rob}}$ is a clipped concentrability coefficient that captures the degree of partial coverage, d is the feature dimension, and H the horizon. Under full feature coverage, the sample complexity improves to $\widetilde{\mathcal{O}}(dH^4\kappa^{-1}\epsilon^{-2})$ (cf. Corollary 2), where $\kappa > 0$ quantifies coverage quality, matching the best known for the unconstrained robust counterpart ([40], Corollary 2).
- **Robustness to model misspecification.** We extend the analysis beyond exact linearity and prove that `CROP-VI` retains robust performance and safety guarantees when transition is only approximately linear (e.g., from soft state aggregation), quantifying the degradation due to misspecification.

A comprehensive comparison with relevant works is presented in Table 1.

1.2 Related works

Constrained tabular and linear MDPs. Primal-dual algorithms are a central tool for CMDPs, with tight performance guarantees established [41, 21, 22, 37]. These methods leverage strong duality to prove convergence to the optimality. In the tabular settings, early analyses provided rigorous convergence and sample-complexity bounds [42, 20, 43–47]. Subsequent work extended these ideas to structured models such as linear-kernel MDPs [48] (see [21] for distinctions from linear

MDPs). Recent studies have begun to address linear MDPs directly [49, 37, 21], but typically *without* robustness: policies that are safe in expectation can violate constraints under misspecified dynamics. Robust CMDPs are fundamentally more challenging because the worst-case transition kernel is policy dependent, making the feasible occupancy-measure set nonconvex and thwarting standard analyses.

Distributionally robust RL with linear MDPs. Robust linear MDPs (Lin-RMDPs) have recently drawn significant attention [36, 50, 40]. Early progress in the offline learning under total variation (TV) uncertainty balls established the first sample complexity guarantees [36], with [40] sharpening the dependence on the feature dimension d to match standard offline linear MDPs [33, 51]. Beyond TV distance, robustness has also been studied under Kullback-Leibler (KL) divergence [50, 36]. In the online setting, [52] achieved near-optimal regret under additional structure on the uncertainty set, while [53] analyzed offline learning with the well-explored data coverage assumption. Complementary lines of work consider richer function classes beyond linear models [54, 55].

Distributionally robust RL with constraints. Early algorithms for RCMDPs adopted a primal-dual framework [56, 24, 23]. The schemes in [56, 24] alternate between computing the worst-case return over a prescribed uncertainty set and improving the policy via standard updates (e.g., policy gradients). [23] proposed a robust primal-dual method and showed that strong duality can fail even under Slater's condition, precluding traditional sample-complexity guarantees for these heuristics. More recently, [26] introduced a trust-region update that preserves feasibility and performance across iterations, though convergence to a near-optimal policy remains open. In the tabular case, [25] gave the first finite-sample bounds, and [57] developed a policy-gradient approach based on an epigraph reformulation of the RCMDP objective.

Notation. For any finite set $\mathcal{S}$, let $\Delta(\mathcal{S})$ denote its probability simplex. For integers $H, d \geq 1$, write $[H] \triangleq \{1, \ldots, H\}$ and $[d] \triangleq \{1, \ldots, d\}$. Let $\mathbf{e}_i$ be the i-th standard basis vector (of appropriate dimension), and let I_d denote the $d \times d$ identity matrix. For $x, y \in \mathbb{R}^d$, $\|x\|_2$ and $\|x\|_1$ denote the ℓ_2- and ℓ_1-norms, and $\langle x, y \rangle$ is the Euclidean inner product. For any positive semidefinite matrix $A \in \mathbb{R}^{d \times d}$, define $\|x\|_A \triangleq \sqrt{x^\top A x}$. For a set $\mathcal{D}$, $|\mathcal{D}|$ denotes its cardinality. We write $\min\{a, b\}_+ \triangleq \max\{\min\{a, b\},\, 0\}$. We use $\mathcal{O}(\cdot)$ for orderwise scaling and $\widetilde{\mathcal{O}}(\cdot)$ to suppress logarithmic factors.

2 Problem Setup

In this section, we introduce the formulation of *distributionally robust constrained Markov Decision Processes (RCMDPs)*, along with the corresponding learning objective and assumptions regarding the model and batch data. To begin with, we consider standard constrained MDPs (CMPDs) as follows.

CMDPs: standard constrained MDPs. Consider a standard constrained MDP, denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P = \{P_h\}_{h=1}^H, r = \{r_h\}_{h=1}^H, g = \{g_h\}_{h=1}^H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and H is the horizon length. At each step $h \in [H]$, $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function and $g_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the utility function. In addition, we denote $\pi = \{\pi_h\}_{h=1}^H$ as the policy of the agent, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ is the action selection probability over the action space $\mathcal{A}$ at the step h. Given any policy π in the CMDP, the learning objective is to solve the following optimization problem:

$$\max_{\pi} \mathbb{E}_{s_0 \sim \zeta}\left[V_{r,1}^{\pi,P}(s_0)\right] \qquad \text{subjection to } \mathbb{E}_{s_0 \sim \zeta}\left[V_{g,1}^{\pi,P}(s_0)\right] \geq b,$$

where b is some positive threshold, ζ is initial state distribution, and for any state $s \in \mathcal{S}$ and $h \in [H]$, $V_{r,h}^{\pi,P}(s) = \mathbb{E}_{\pi,P}\left[\sum_{t=h}^H r_t(s_t, a_t) \mid s_h = s\right]$ and $V_{g,h}^{\pi,P}(s) = \mathbb{E}_{\pi,P}\left[\sum_{t=h}^H g_t(s_t, a_t) \mid s_h = s\right]$ are value functions for the reward function r and utility function g, respectively.

Linear function approximation. To handle the potentially massive state space, we impose the following linear model assumption, which is commonly used in the previous literature [49, 21, 37, 40].

Assumption 1 (Linear CMDPs)**.** *A finite-horizon CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r, g)$ is a linear CMDP if given a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, there exist d unknown measures $\mu_h^P = (\mu_{h,1}^P, \cdots, \mu_{h,d}^P)$ over the state space $\mathcal{S}$ and two unknown vectors $\theta_{r,h}, \theta_{g,h} \in \mathbb{R}^d$ at each step h such that for $\forall\,(h, s, a, s') \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$r_h(s, a) = \langle \phi(s, a), \theta_{r,h} \rangle,\ g_h(s, a) = \langle \phi(s, a), \theta_{g,h} \rangle,\ P_h(s' \mid s, a) = \langle \phi(s, a), \mu_h^P(s') \rangle.$$

Without loss of generality, we assume that $\|\phi(s,a)\|_2 \leq 1$ *and* $\phi_i(s,a) \geq 0$ *for any* $(s,a,i) \in \mathcal{S} \times \mathcal{A} \times [d]$*, and* $\max\left\{\int_{\mathcal{S}} \left\|\mu_h^P(s)\right\|_2 \mathrm{d}s, \|\theta_{r,h}\|_2, \|\theta_{g,h}\|_2\right\} \leq \sqrt{d}$ *for all* $h \in [H]$.

Lin-RCMDPs: distributionally robust linear CMDPs. To handle the uncertainty of environments, we consider *distributionally robust* linear CMDPs (Lin-RCMDPs), where the transition kernel can be an arbitrary one within an uncertainty set around the nominal kernel [21]. Formally, we denote it by $\mathcal{M}_{\mathrm{rob}} = \left(\mathcal{S}, \mathcal{A}, H, \mathcal{P}^\rho\left(P^0\right), r, g\right)$, where P^0 represents a nominal transition kernel and then $\mathcal{P}^\rho\left(P^0\right)$ represents the uncertainty set (a ball) around the nominal P^0 with some uncertainty level $\rho \geq 0$. We assume that the uncertainty set satisfies the following d-rectangularity assumption [58, 40].

Assumption 2 (d-rectangularity of the uncertainty set)**.** *In Lin-RCMDPs, the uncertainty set* $\mathcal{P}^\rho\left(P^0\right)$ *is* d*-rectangular, i.e.,* $\mu_{h,i}^0 \in \Delta(\mathcal{S})$ *for any* $(h,i) \in [H] \times [d]$ *and*

$$\mathcal{P}^\rho\left(P^0\right) \coloneqq \otimes_{[H],\mathcal{S},\mathcal{A}} \mathcal{P}^\rho\left(P_{h,s,a}^0\right), \quad \textit{with } \mathcal{P}^\rho\left(P_{h,s,a}^0\right) \coloneqq \left\{\langle \phi(s,a), \mu_h(\cdot)\rangle : \mu_h \in \mathcal{U}^\rho\left(\mu_h^0\right)\right\},$$

where $\mu_h^0 \coloneqq \mu_h^{P^0}$ *for simplicity,* $\mathcal{U}^\rho(\mu_h^0) \coloneqq \otimes_{[d]} \mathcal{U}^\rho(\mu_{h,i}^0)$*,* $\mathcal{U}^\rho(\mu_{h,i}^0) \coloneqq \{\mu_{h,i} : D_{\mathrm{TV}}(\mu, \mu_{h,i}^0) \leq \rho$ *and* $\mu_{h,i}^0 \in \Delta(\mathcal{S})\}$*,* $\otimes_{[d]}$ *(resp.* $\otimes_{[H],\mathcal{S},\mathcal{A}}$*) denotes Cartesian products over* $[d]$ *(resp.* $[H]$*,* $\mathcal{S}$*, and* $\mathcal{A}$*) and* ρ *is the uncertainty level. Here,* $D_{\mathrm{TV}}(\mu_{h,i}, \mu_{h,i}^0) = \frac{1}{2}\|\mu_{h,i} - \mu_{h,i}^0\|_1$ *represent the total variation (TV) distance between two probability measures.*

Note that when the uncertainty radius $\rho = 0$, Lin-RCMDPs reduce to the standard linear CMDPs satisfying Assumption 1. To further robustify the learned policy when $\rho > 0$, we are interested in the *worst-case* performance induced by all possible transition kernels over the uncertainty set, characterized by *robust value function* (resp. *robust Q function*) for the reward r and utility g, respectively. For simplicity, we interchangeably denote that for $j = r, g$,

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \; V_{j,h}^{\pi,\rho}(s) \coloneqq \inf_{P \in \mathcal{P}^\rho(P^0)} V_{j,h}^{\pi,P}(s), \; Q_{j,h}^{\pi,\rho}(s,a) = \inf_{P \in \mathcal{P}^\rho(P^0)} Q_{j,h}^{\pi,P}(s,a).$$

Additionally, we denote $\pi^\star = \{\pi_h^\star\}_{h=1}^H$ as a deterministic *optimal robust policy* [59] that maximizes the robust value function for reward r while satisfying the constraint for utility g simultaneously, i.e., $\pi_h^\star(s) = \arg\max V_{r,h}^{\pi,\rho}(s)$ and $V_{g,h}^{\pi,\rho}(s) \geq b$ for any $(s,h) \in \mathcal{S} \times [H]$. The resulting *optimal robust value/utility function* and *optimal robust Q-function* are defined as:

$$V_{r,h}^{\star,\rho}(s) \coloneqq V_{r,h}^{\pi^\star,\rho}(s) = \max_\pi V_{r,h}^{\pi,\rho}(s), \qquad \forall \in [H] \times \mathcal{S},$$

$$V_{g,h}^{\star,\rho}(s) \coloneqq V_{g,h}^{\pi^\star,\rho}(s) = \inf_{P \in \mathcal{P}^\rho(P^0)} \mathbb{E}_{\pi^\star,P}\left[\sum_{t=h}^H g_t(s_t,a_t) \mid s_h = s\right], \qquad \forall \in [H] \times \mathcal{S},$$

and for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$Q_{r,h}^{\star,\rho}(s,a) \coloneqq Q_{r,h}^{\pi^\star,\rho}(s,a) = \max_\pi Q_{r,h}^{\pi,\rho}(s,a), \qquad Q_{g,h}^{\star,\rho}(s,a) \coloneqq Q_{g,h}^{\pi^\star,\rho}(s,a) = Q_{g,h}^{\pi^\star,\rho}(s,\pi^\star(s)).$$

In addition, conditioned on some initial state distribution ζ, we define the induced *occupancy distribution* w.r.t. any policy π and transition kernel P as follows: for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$d_h^{\pi,P}(s) \coloneqq d_h^{\pi,P}(s;\zeta) = \mathbb{P}(s_h = s \mid \zeta, \pi, P), \quad d_h^{\pi,P}(s,a) = d_h^{\pi,P}(s)\pi_h(a \mid s). \tag{1}$$

Similar to (1), we also denote the occupancy distribution associated with the optimal robust policy $\pi^\star$ and some transition kernel P by

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad d_h^{\star,P}(s) \coloneqq d_h^{\pi^\star,P}(s;\zeta), \qquad d_h^{\star,P}(s,a) \coloneqq d_h^{\pi^\star,P}(s)\pi_h^\star(a \mid s).$$

Robust linear Bellman operator and optimality equations. A key component of Lin-RCMDPs is the extension of the Bellman optimality principle for both reward and utility functions, captured by the following *robust Bellman consistency equation* (or equivalently, the *robust Bellman optimality equation* for π^*): for $j = r, g$,

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad Q_{j,h}^{\pi,\rho}(s,a) = \mathbb{B}_{j,h}^\rho V_{j,h+1}^{\pi,\rho}(s,a), \; V_{j,h}^{\pi,\rho}(s) = \mathbb{E}_{a \sim \pi_h(\cdot|s)}[Q_{j,h}^{\pi,\rho}(s,a)].$$

where the robust Bellman operator and transition operator for any function $f : \mathcal{S} \to \mathbb{R}$ are defined by

$$[\mathbb{B}^{\rho}_{j,h} f](s,a) := j_h(s,a) + [\mathbb{P}^{\rho}_h f](s,a), \tag{2}$$
$$[\mathbb{P}^{\rho}_h f](s,a) := \inf_{\mu_h \in \mathcal{U}^{\rho}(\mu^0_h)} \int_{\mathcal{S}} \langle \phi(s,a), \mu_h(s') \rangle f(s') \mathrm{d}s'.$$

Note that under Assumption 2, the robust Bellman operator inherits the linearity of the Bellman operator in standard linear MDPs [42, Proposition 2.3], as shown in the following lemma.

Lemma 1 (Linearity of robust Bellman operators)**.** *[Lemma 1 in [40]] Suppose that the finite-horizon Lin-RCMDPs satisfies Assumption 1 and 2. There exist weights $w^{\rho}_j = \{w^{\rho}_{j,h}\}_{h=1}^H$, where $w^{\rho}_{j,h} := \theta_{j,h} + \inf_{\mu_h \in \mathcal{U}^{\rho}(\mu^0_h)} \int_{\mathcal{S}} \mu_h(s') f(s') \mathrm{d}s'$ for any $h \in [H]$, such that $\mathbb{B}^{\rho}_{j,h} f(s,a)$ is linear with respect to the feature map ϕ, i.e., $\mathbb{B}^{\rho}_{j,h} f(s,a) = \langle \phi(s,a), w^{\rho}_{j,h} \rangle$ for $j = r, g$.*

Offline data. Consider a batch dataset $\mathcal{D} = \left\{ \left(s^{\tau}_h, a^{\tau}_h, r^{\tau}_h, g^{\tau}_h, s^{\tau}_{h+1} \right) \right\}_{h \in [H], \tau \in [K]}$ consisting of K i.i.d. trajectories generated by executing some (mixed) behavior policy $\pi^{\mathrm{b}} = \left\{ \pi^{\mathrm{b}}_h \right\}_{h=1}^H$ in interaction with a nominal linear MDP $\mathcal{M}^0 = \left(\mathcal{S}, \mathcal{A}, H, P^0, r, g \right)$. Note that $\mathcal{D}$ contains KH transition-reward sample tuples, where each sample tuple $\left(s^{\tau}_h, a^{\tau}_h, r^{\tau}_h, g^{\tau}_h, s^{\tau}_{h+1} \right)$ represents that the agent took the action a^{τ}_h in state s^{τ}_h, received the reward $r^{\tau}_h = r_h\left(s^{\tau}_h, a^{\tau}_h\right)$ and the utility $g^{\tau}_h = g_h(s^{\tau}_h, a^{\tau}_h)$, and then transitioned to the next state $s^{\tau}_{h+1} \sim P^0_h\left(\cdot \mid s_h = s^{\tau}_h, a_h = a^{\tau}_h\right)$. We further define the slice of data at step h as

$$\mathcal{D}^0_h = \left\{ \left(s^{\tau}_h, a^{\tau}_h, r^{\tau}_h, g^{\tau}_h, s^{\tau}_{h+1} \right) \right\}_{\tau \in [K]}.$$

For simplicity, we abuse the notation $\tau \in \mathcal{D}^0_h$ to denote $\left(s^{\tau}_h, a^{\tau}_h, r^{\tau}_h, g^{\tau}_h, s^{\tau}_{h+1} \right) \in \mathcal{D}^0_h$. In addition, we define the occupancy distribution induced by the behavior policy π^{b} and the nominal transition kernel P^0 at each step h, conditioned on the initial state distribution ζ, as:

$$d^{\mathrm{b}}_h(s) \coloneqq d^{\pi^{\mathrm{b}}, P^0}_h(s; \zeta) \quad \text{and} \quad d^{\mathrm{b}}_h(s,a) \coloneqq d^{\pi^{\mathrm{b}}, P^0}_h(s,a; \zeta), \quad \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}.$$

Learning goal. Given the batch dataset $\mathcal{D}$ and an initial state distribution ζ, the Lin-RCMDP $\mathcal{M}_{\mathrm{rob}}$ seeks to solve the following problem:

$$\max_{\pi} \mathbb{E}_{s_0 \sim \zeta}\left[V^{\pi,\rho}_{r,1}(s_0) \right] \qquad \text{subjection to } \mathbb{E}_{s_0 \sim \zeta}\left[V^{\pi,\rho}_{g,1}(s_0) \right] \geq b. \tag{3}$$

Our goal is to learn an ϵ-optimal and ϵ-level constrained robust policy $\widehat{\pi}$ satisfies

$$\mathrm{Sub-Opt}\left(\widehat{\pi}; \zeta, \mathcal{P}^{\rho}\right) \coloneqq V^{\star,\rho}_{r,1}(\zeta) - V^{\widehat{\pi},\rho}_{r,1}(\zeta) \leq \epsilon, \ \mathrm{Violation}\left(\widehat{\pi}; \zeta, \mathcal{P}^{\rho}\right) \coloneqq b - V^{\widehat{\pi},\rho}_{g,1}(\zeta) \leq \epsilon, \tag{4}$$

with as few samples as possible, where ϵ represents the targeted accuracy, and

$$V^{\star,\rho}_{r,1}(\zeta) = \mathbb{E}_{s_1 \sim \zeta}[V^{\star,\rho}_{r,1}(s_1)], \quad V^{\widehat{\pi},\rho}_{j,1}(\zeta) = \mathbb{E}_{s_1 \sim \zeta}[V^{\widehat{\pi},\rho}_{j,1}(s_1)], \quad j \in \{r, g\}.$$

3 CROP-VI: Constrained Robust Optimistic-Pessimistic Value Iteration

In this section, we introduce a novel model-based offline algorithm—*Constrained Robust Optimistic-Pessimistic Value Iteration* (CROP-VI) for Lin-RCMDPs. We first consider a rectified form of problem (3) using the dual variable. Then, we construct an empirical Bellman operator for Lin-RCMDPs.

3.1 Rectified dual form

As highlighted by [25], the traditional primal-dual method developed for non-robust CMDPs does not directly extend to the robust CMDPs, since the robust state-action occupancy measure is not convex and strong duality does not hold. A feasible solution is to transform the original problem (3) into the following rectified dual form:

$$\max_{\pi} \ V^{\pi,\rho}_{r,1}(\zeta) - \beta \left(b - V^{\pi,\rho}_{g,1}(\zeta) \right)_+, \tag{5}$$

where $\beta > 0$ is the penalty coefficient and $(x)_+ = \max\{x, 0\}$. The equivalence between the rectified formulation (5) and the original problem (3) is established in the following lemma.

Algorithm 1: `CROP-VI`: Constrained Robust Optimistic-Pessimistic Value Iteration

Input: Dataset $\mathcal{D}$, feature map $\phi(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, performance tolerance ϵ, parameters $\beta = H/\epsilon$, $\lambda_0, \gamma_0 > 0$;

Construct a temporally independent dataset $\mathcal{D}^0 = \texttt{Two-fold-subsampling}(\mathcal{D})$ (Algorithm 2).

Initialization: $\widehat{V}_{r,H+1}(\cdot) = 0$, $\widehat{V}_{g,H+1}(\cdot) = 0$

for *steps* $h = H, H-1, \ldots, 1$ **do**

- $\Lambda_h = \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda_0 I_d$; `// sample covariance matrix`
- $\widehat{\theta}_{r,h} = \Lambda_h^{-1}\left(\sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) r_h^\tau\right)$;
- $\widehat{\theta}_{g,h} = \Lambda_h^{-1}\left(\sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) g_h^\tau\right)$;
- **for** *feature* $i = 1, \cdots, d$ **do**
  - Update $\widehat{\nu}_{h,i}^{\rho,\widehat{V}_r}$ and $\widehat{\nu}_{h,i}^{\rho,\widehat{V}_g}$ via (8).
- **end**
- $\widehat{w}_{r,h}^{\rho,\widehat{V}_r} = \widehat{\theta}_{r,h} + \widehat{\nu}_h^{\rho,\widehat{V}_r}$, $\widehat{w}_{g,h}^{\rho,\widehat{V}_g} = \widehat{\theta}_{g,h} + \widehat{\nu}_h^{\rho,\widehat{V}_g}$;
- $\bar{Q}_{r,h}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{w}_{r,h}^{\rho,\widehat{V}_r} - \Gamma_h(s,a)$; `// Pessimistic reward Q-function`
- $\widehat{Q}_{r,h}(\cdot,\cdot) = \min\left\{\bar{Q}_{r,h}(\cdot,\cdot), H-h+1\right\}$
- $\bar{Q}_{g,h}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{w}_{g,h}^{\rho,\widehat{V}_g} + \Gamma_h(s,a)$; `// Optimistic constraint Q-function`
- $\widehat{Q}_{g,h}(\cdot,\cdot) = \min\left\{\bar{Q}_{g,h}(\cdot,\cdot), H-h+1\right\}$
- $\widehat{\pi}_h(\cdot) = \arg\max_{a \in \mathcal{A}} \widehat{Q}_{r,h}(\cdot,a) - \beta(b - \widehat{Q}_{g,h}(\cdot,a))_+$; `// Rectified optimization`
- $\widehat{V}_{r,h}^{\widehat{\pi}}(\cdot) = \widehat{Q}_{r,h}(\cdot, \widehat{\pi}(\cdot))$, $\widehat{V}_{g,h}^{\widehat{\pi}}(\cdot) = \widehat{Q}_{g,h}(\cdot, \widehat{\pi}(\cdot))$

end

Output: $\widehat{V}_r \coloneqq \{\widehat{V}_{r,h}^{\widehat{\pi}}\}_{h=1}^{H+1}$, $\widehat{V}_g \coloneqq \{\widehat{V}_{g,h}^{\widehat{\pi}}\}_{h=1}^{H+1}$, $\widehat{\pi} \coloneqq \{\widehat{\pi}_h\}_{h=1}^{H}$

Lemma 2 (Lemma 1 in [25])**.** *For any $\epsilon > 0$, choosing $\beta = H/\epsilon$ ensures that the optimal solution $\widehat{\pi}$ of* (5) *has constraint shortfall at most ϵ, i.e., $b - V_{g,1}^{\widehat{\pi},\rho}(\zeta) \le \epsilon$. Moreover, if there is an ϵ-separation between feasible and infeasible policies, i.e., every infeasible π satisfies $V_{g,1}^{\pi,\rho}(\zeta) - b < \epsilon$, then the optimal solution $\pi^\star$ of* (3) *and* (5) *coincide.*

It is readily seen that the cumulative penalty paid for violating the constraint at any step is upper-bounded by H. According to Lemma 2, scaling by $\beta = H/\epsilon$ makes any violation greater than ϵ strictly suboptimal relative to a feasible alternative. As $\beta \to \infty$, the optimal solution of (5) converges to that of the original problem (3).

3.2 Empirical robust Bellman operator and strong duality

For $j \in \{r, g\}$ and step $h \in [H]$, the robust Bellman operator given in (2) admits the dual form (by linearity and strong duality; cf. [59, 39, 40]):

$$(\mathbb{B}_{j,h}^\rho V_j)(s,a) = \langle \phi(s,a), \theta_{j,h} + \nu_h^{\rho,V_j} \rangle,$$

where $\nu_h^{\rho,V_j} = [\nu_{h,1}^{\rho,V_j}, \nu_{h,2}^{\rho,V_j} \ldots, \nu_{h,d}^{\rho,V_j}]^\top \in \mathbb{R}^d$ and its i-th coordinate is defined by

$$\nu_{h,i}^{\rho,V_j} \coloneqq \max_{\alpha \in [\min_s V_j(s), \max_s V_j(s)]} \left\{\mathbb{E}_{s \sim \mu_{h,i}^0}[V_j]_\alpha(s) - \rho(\alpha - \min_{s'}[V_j]_\alpha(s'))\right\}, \tag{6}$$

with $[V_j]_\alpha(s) = \min\{V_j(s), \alpha\}$. Since we do not have direct access to the nominal linear MDP $\mathcal{M}^0$ (i.e., $\theta_{j,h}$ and μ_h^0), we cannot perform value iteration directly with the above Bellman operator. To address this, we estimate $\widehat{\theta}_{j,h} \in \mathbb{R}^d$ and $\widehat{\nu}_h^{\rho,V_j} \in \mathbb{R}^d$ using the batch dataset $\mathcal{D}_h^0$ containing all the samples at h-th step in $\mathcal{D}^0$. To ensure numerical stability, we use ridge regression instead of directly minimizing empirical mean squared errors. Specifically, for any value function $V_j : \mathcal{S} \to [0, H]$ and

time step $h \in [H]$, the estimator $\widehat{\theta}_{j,h}$ and the i-th coordinate of $\widehat{\nu}_h^{\rho,V_j}$ are defined as:

$$\widehat{\theta}_{j,h} = \arg\min_{\theta\in\mathbb{R}^d} \sum_{\tau\in\mathcal{D}_h^0} \big(\phi(s_h^\tau, a_h^\tau)^\top\theta_j - j_h^\tau\big)^2 + \lambda_0\|\theta_j\|_2^2 = \Lambda_h^{-1}\Big(\sum_{\tau\in\mathcal{D}_h^0}\phi(s_h^\tau, a_h^\tau) j_h^\tau\Big), \tag{7}$$

$$\widehat{\nu}_{h,i}^{\rho,V_j} = \max_{\alpha\in[\min_s V_j(s), \max_s V_j(s)]} \big\{\bar{\nu}_{h,i}^{V_j}(\alpha) - \rho\big(\alpha - \min_{s'}[V_j]_\alpha(s')\big)\big\}, \quad \forall i\in[d],\ j\in\{r,g\}, \tag{8}$$

where $\lambda_0 > 0$ is the regularization coefficient,

$$\Lambda_h = \sum_{\tau\in\mathcal{D}_h^0}\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda_0 I_d \tag{9}$$

is the cumulative sample covariance matrix, and $\bar{\nu}_{h,i}^{V_j}(\alpha)$ is the i-th coordinate of $\bar{\nu}_h^{V_j}(\alpha)$ defined by

$$\bar{\nu}_h^{V_j}(\alpha) = \arg\min_{\nu\in\mathbb{R}^d} \sum_{\tau\in\mathcal{D}_h^0} \big(\phi(s_h^\tau, a_h^\tau)^\top\nu - [V_j]_\alpha(s_{h+1}^\tau)\big)^2 + \lambda_0\|\nu\|_2^2 = \Lambda_h^{-1}\Big(\sum_{\tau\in\mathcal{D}_h^0}\phi(s_h^\tau, a_h^\tau)[V_j]_\alpha(s_{h+1}^\tau)\Big).$$

Leveraging the linearity of the robust Bellman operator shown in Lemma 1, we construct the empirical robust Bellman operator $\widehat{\mathbb{B}}_h^\rho$ to approximate $\mathbb{B}_h^\rho$: for any function $V_j : \mathcal{S}\to[0,H]$,

$$(\widehat{\mathbb{B}}_{j,h}^\rho V_j)(s,a) = \phi(s,a)^\top(\widehat{\theta}_{j,h} + \widehat{\nu}_h^{\rho,V_j}), \qquad \forall (s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H],\ \ j\in\{r,g\}. \tag{10}$$

3.3 `CROP-VI`: constrained robust optimistic-pessimistic value iteration

To compute the empirical Bellman equation (10) recursively over each horizon $h \in [H]$, we propose constrained robust optimistic-pessimistic value iteration (`CROP-VI`), summarized in Algorithm 1.

Algorithm 1 begins by constructing a subsampled dataset $\mathcal{D}^0$ from the original batch $\mathcal{D}$ using the `Two-fold-subsampling` (Algorithm 2), following [60] to mitigate statistical dependencies across time steps. Further details on `Two-fold-subsampling` and its guarantees are provided in Section B.2. Given $\mathcal{D}_0$ and the terminal conditions $\widehat{V}_{r,H+1}(\cdot) = \widehat{V}_{g,H+1}(\cdot) = 0$, each iteration of `CROP-VI` at step h proceeds in two phases:

• **Robust Bellman construction.** Conditioned on the fixed $\widehat{V}_{j,h+1}(j\in\{r,g\})$ from the previous iteration (line 4-9 in Algorithm 1), we build the empirical robust Bellman operators via (7)-(10).

• **Q-functions update.** Using the robust Bellman operators, we then estimate the pessimistic reward Q-function and optimistic constraint Q-function as follows: for all $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\bar{Q}_{r,h}(s,a) = (\widehat{\mathbb{B}}_{r,h}^\rho\widehat{V}_{r,h+1})(s,a) - \Gamma_h(s,a), \ \ \bar{Q}_{g,h}(s,a) = (\widehat{\mathbb{B}}_{g,h}^\rho\widehat{V}_{g,h+1})(s,a) + \Gamma_h(s,a)$$

where $\Gamma_h : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is a state-action-dependent penalty/bonus term defined as $\Gamma_h(s,a) := \gamma_0\sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}$ with $\gamma_0 > 0$ controlling the regularization strength.

Design rationale. In offline RL, the principle of pessimism is routinely invoked to compensate for the limited coverage inherent in batch data [33, 61, 62, 40]. Applying pessimism to both the reward and the constraints guarantees safety, but often at the cost of excessive conservatism. To strike a better balance, we adopt an *asymmetric* strategy: we remain pessimistic with respect to the reward function to prevent over-estimation, while we take an optimistic stance toward the constraint value. This optimism encourages exploratory behavior yet still enforces feasibility, as the update ensures $\widehat{V}_g \geq b$. Although $\widehat{V}_g \geq b$ does not imply $V_g^\star \geq b$, our analysis guarantees that $|\widehat{V}_g - V_g^\star| \leq \epsilon$ with high probability, thereby tightly bounding the gap between the estimated and true constraint values.

4 Theoretical Guarantees

In this section, we establish a high-probability sub-optimality bound for the policy generated by `CROP-VI` and the guarantee of sample complexity under *partial* and *full* feature coverage. We further extend the analysis to accommodate model misspecification.

4.1 Main results

Theorem 1. *Fix $\beta = H/\epsilon$ and consider any $\delta \in (0,1)$. Suppose Assumptions 1 and 2 hold. Set $\lambda_0 = 1$, $\gamma_0 = 6\sqrt{d\xi_0}H$ in Algorithm 1, where $\xi_0 = \log(9HK/\delta)$. Then, with probability at least $1-\delta$, the policy $\widehat{\pi}$ generated by Algorithm 1 satisfies*

$$\text{Sub-Opt}(\widehat{\pi};\zeta,\mathcal{P}^\rho) \le \widetilde{O}(\sqrt{d}H)\sum_{h=1}^{H}\sum_{i=1}^{d}\max_{P\in\mathcal{P}^\rho(P^0)}\mathbb{E}_{\pi^\star,P}\Big[\|\phi_i(s_h,a_h)\mathbb{1}_i\|_{\Lambda_h^{-1}}\Big],\ \text{Violation}(\widehat{\pi};\zeta,\mathcal{P}^\rho)\le\epsilon. \tag{11}$$

Theorem 1 demonstrates that our method achieves ϵ-level constraint violation, while providing an instance-dependent sub-optimality bound without requiring any data coverage assumption. Notably, the sub-optimality bound is independent of the dimensionality of the state space, depending instead on the feature space dimension. The bound is governed by the confidence parameter δ and a general complexity term reflecting how effectively the offline dataset covers the feature space. This result highlights that the sub-optimality is fundamentally determined by the quality of offline data, encapsulated by the expected feature exploration. Building on this foundation, we further analyze the sample complexity required to achieve an ϵ-optimality and ϵ-violation optimal policy under varying data coverage conditions. The proof is postponed to Appendix B.3.1.

4.2 The case of partial feature coverage

We first consider the scenario of partial feature coverage (a.k.a. sufficient feature coverage), where the behavior policy is only required to cover all feature dimensions visited by the optimal robust policy. Formally, we consider the following clipped single-policy concentrability condition, which is used in [40] for Lin-RMDPs. This condition quantifies the worst-case discrepancy between the occupancy measure induced by the optimal robust policy $\pi^\star$ under any kernel $P\in\mathcal{P}^\rho(P^0)$ and that of the behavior policy π^{b} under the nominal kernel P^0 for every feature $i\in[d]$.

Assumption 3 (Robust single-policy clipped concentrability)**.** *The dataset $\mathcal{D}$, collected by the behavior policy π^{b}, satisfies*

$$\forall(i,h,P)\in[d]\times[H]\times\mathcal{P}^\rho\left(P^0\right),\ \frac{u^\top\left(\min\{\mathbb{E}_{d_h^{\star,P}}\phi_i^2(s,a),1/d\}\cdot\mathbb{1}_{i,i}\right)u}{u^\top\left(\mathbb{E}_{d_h^{\text{b}}}[\phi(s,a)\phi(s,a)^\top]\right)u}\le\frac{C_{\text{rob}}^\star}{d},$$

for some finite quantity $C_{\text{rob}}^\star\in[1,\infty)$. In addition, we follow the convention $0/0=0$.

The quantity $C_{\text{rob}}^\star$ is a clipped concentrability constant that measures the relative feature coverage of the batch data compared to an ideal dataset generated by $\pi^\star$. A smaller $C_{\text{rob}}^\star$ implies better alignment with $\pi^\star$, (i.e., higher data quality), while a larger $C_{\text{rob}}^\star$ indicates worse data quality. Importantly, `CROP-VI` does not require prior knowledge of $C_{\text{rob}}^\star$ during implementation; it suffices to assume $C_{\text{rob}}^\star<\infty$, ensuring that π^{b} sufficiently explores every feature dimension that $\pi^\star$ visits.

Under Assumption 3, we establish the following sample complexity guarantee for achieving an ϵ-optimal and ϵ-constrainted robust policy.

Corollary 1 (Partial feature coverage)**.** *Suppose Assumptions 1, 2 and 3 hold. Consider any $\delta\in(0,1)$ and the same hyperparameter settings as in Theorem 1. Let $d_{\min}^{\text{b}}=\min_{h,s,a}\{d_h^{\text{b}}(s,a):d_h^{\text{b}}(s,a)>0\}$. Then, with probability exceeding $1-\delta$, the policy $\widehat{\pi}$ returned by Algorithm 1 achieves*

$$\text{Sub-Opt}(\widehat{\pi};\zeta,\mathcal{P}^\rho)\le\mathcal{O}\left(d^{2/3}H^2\sqrt{C_{\text{rob}}^\star\log(12HK/\delta)/K}\right)$$

provided that $K\ge c_0\log(KH/\delta)/d_{\min}^{\text{b}}$ for some sufficiently large universal constant $c_0>0$. In other words, the learned ϵ-constrained policy $\widehat{\pi}$ is ϵ-optimal if the total number of sample trajectories satisfies $K\ge\widetilde{O}\left(C_{\text{rob}}^\star d^2H^4/\epsilon^2\right)$.

Corollary 1 is a direct consequence of Theorem 1 and establishes a sub-optimality bound that scales with the feature dimension d and the horizon length H. This bound is comparable to those achieved in prior work on standard linear MDPs [33] and unconstrained distributionally robust linear MDPs [40]. Moreover, `CROP-VI` guarantees constraint satisfaction. The proof is deferred to Appendix B.4.

4.3 The case of full feature coverage

We next introduce the following full feature coverage assumption, which is commonly employed in the analysis of offline linear MDPs [33, 51, 34, 50], and requires the behavior policy to uniformly cover the feature space.

Assumption 4 (Full data coverage). *We assume* $\kappa = \min_{h\in[H]} \lambda_{\min}(\mathbb{E}_{d_h^{\mathrm{b}}}[\phi(s,a)\phi(s,a)^\top]) > 0$.

Compared to Assumption 3, Assumption 4 necessitates the behavior policy to be more exploratory to reach every feature dimension, which is a stronger assumption requiring full coverage of all feature dimensions. The following corollary provides the sample complexity guarantee of `CROP-VI` under the full feature coverage.

Corollary 2 (Full feature coverage). *Suppose Assumptions 1, 2 and 4 hold. Consider any* $\delta \in (0,1)$ *and the same hyperparameter settings as in Theorem 1. Let* $d_{\min}^{\mathrm{b}} = \min_{h,s,a}\{d_h^{\mathrm{b}}(s,a) : d_h^{\mathrm{b}}(s,a) > 0\}$. *Then, with probability at least* $1-\delta$, *the* ϵ*-constrained policy* $\widehat{\pi}$ *returned by Algorithm 1 achieves*

$$\text{Sub-Opt}(\widehat{\pi};\zeta,\mathcal{P}^\rho) \le \mathcal{O}\left(H^2\sqrt{d\log(15HK/\delta)\kappa^{-1}K^{-1}}\right),$$

provided that $K \ge \max\{c_0\log(2Hd/\delta)/\kappa^2, c_0\log(KH/\delta)/d_{\min}^{\mathrm{b}}\}$ *for some sufficiently large universal constant* c_0. *In other words, the learned* ϵ*-constrained policy* $\widehat{\pi}$ *is* ϵ*-optimal if the total number of sample trajectories satisfies* $K \ge \widetilde{\mathcal{O}}\Big(dH^4\kappa^{-1}\epsilon^{-2}\Big)$.

Corollary 2 is a direct consequence of Theorem 1. The resulting sample complexity matches the known bounds for standard linear MDPs without robustness and constraint [34, 35], as well as for those without constraint [40], while additionally ensuring constraint satisfaction. The proof is postponed to Appendix B.5.

4.4 Extension

We notice that the soft state-aggregation assumption in Assumption 1 may be overly restrictive in practice. To this end, we relax this assumption by allowing the true transition kernel dynamics to deviate slightly from the state-aggregation transition, as follows.

Assumption 5 (Model misspecification in transition model). *For each* $h \in [H]$, *there exists a kernel* $\tilde{P}_h^0 \in \text{Span}\{\phi(s,a)\}$ *and a deviation coefficient* $\psi \ge 0$ *such that, for all* (s,a), *the true transition kernel* $P_h^0(\cdot|s,a)$ *satisfies* $\|P_h^0(\cdot|s,a) - \tilde{P}_h^0(\cdot|s,a)\|_1 \le \psi$. *Here,* $\text{Span}\{\phi(s,a)\}$ *denotes the set of all linear combinations of the feature vector* $\phi(s,a)$. *We continue to assume, without loss of generality, that the reward/utility functions satisfy* $r_h, g_h \in \text{Span}\{\phi(s,a)\}$ *for all* $h \in [H]$.

Assumption 5 is also considered in [58]. Next, we present the following two corollaries that quantify the impact of model misspecification for both partial and full feature dimension.

Corollary 3 (Partial feature coverage). *Suppose Assumptions 1, 2, 3 and 5 hold. Consider any* $\delta \in (0,1/4)$ *and the same hyperparameter settings as in Theorem 1. Let* $d_{\min}^{\mathrm{b}} = \min_{h,s,a}\{d_h^{\mathrm{b}}(s,a) : d_h^{\mathrm{b}}(s,a) > 0\}$. *Then, with probability exceeding* $1-4\delta$, *the* ϵ*-level constrained policy* $\widehat{\pi}$ *returned by Algorithm 1 achieves*

$$\text{Sub-Opt}(\widehat{\pi};\zeta,\mathcal{P}^\rho) \le 16dH^2\sqrt{\frac{C_{\mathrm{rob}}^\star}{K}}\left(6\log\left(\frac{3HK}{\delta}\right)+\psi\right) + \frac{H(H-1)\psi}{2}$$

provided that $K \ge c_0\log(KH/\delta)/d_{\min}^{\mathrm{b}}$ *for some sufficiently large universal constant* $c_0 > 0$.

Corollary 4 (Full feature coverage). *Suppose Assumptions 1, 2, 3 and 5 hold. Consider any* $\delta \in (0,1/5)$ *and the same hyperparameter settings as in Theorem 1. Let* $d_{\min}^{\mathrm{b}} = \min_{h,s,a}\{d_h^{\mathrm{b}}(s,a) : d_h^{\mathrm{b}}(s,a) > 0\}$. *Then, with probability at least* $1-5\delta$, *the* ϵ*-level constrained policy* $\widehat{\pi}$ *returned by Algorithm 1 achieves*

$$\text{Sub-Opt}(\widehat{\pi};\zeta,\mathcal{P}^\rho) \le \frac{16\sqrt{d}H^2}{\sqrt{\kappa K}}\left(6\log\left(\frac{3HK}{\delta}\right)+\psi\right) + \frac{H(H-1)\psi}{2},$$

provided that $K \ge \max\{c_0\log(2Hd/\delta)/\kappa^2, c_0\log(KH/\delta)/d_{\min}^{\mathrm{b}}\}$ *for some sufficiently large universal constant* c_0.

Corollaries 3 and 4 quantify the impact of model misspecification: when the soft-state aggregation model deviates from the true one by at most ψ in total variation, the sub-optimality of the learned policy degrades by an additive factor of $\mathcal{O}(\psi H^2 \cdot (d\sqrt{C^\star_{\mathrm{rob}}/K} + 1))$ under partial feature coverage, and $\mathcal{O}(\psi H^2 \cdot (\sqrt{d\kappa^{-1}K^{-1}} + 1))$ under full feature coverage.

5 Conclusion

In this paper, we investigate the sample complexity for distributional robust constraint offline RL with linear representations, where the uncertainty sets are characterized by TV distance. To this end, we develop a distributionally robust variant of constraint optimistic-pessimistic least-squares value iteration, called `CROP-VI`. We establish the sub-optimality bound and the constraint violation bound for Lin-RCMDPs under various offline data assumptions. We further extend our analysis to accommodate transition model misspecification. In the future, an important direction is to explore alternative uncertainty sets [62] and to establish the lower bound across the full range of the uncertainty levels.

References

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

[2] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, Trevor Darrell, et al. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

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

[4] Tianyu Shi, Dong Chen, Kaian Chen, and Zhaojian Li. Offline reinforcement learning for autonomous driving with safety and exploration enhancement. *arXiv preprint arXiv:2110.07067*, 2021.

[5] Shangding Gu, Guang Chen, Lijun Zhang, Jing Hou, Yingbai Hu, and Alois Knoll. Constrained reinforcement learning for vehicle motion planning with topological reachability analysis. *Robotics*, 11(4):81, 2022.

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

[7] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys*, 55(1):1–36, 2021.

[8] Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pages 2–35. PMLR, 2021.

[9] Aashma Uprety and Danda B Rawat. Reinforcement learning for IoT security: A comprehensive survey. *IEEE Internet of Things Journal*, 8(11):8693–8706, 2020.

[10] Helin Yang, Zehui Xiong, Jun Zhao, Dusit Niyato, Liang Xiao, and Qingqing Wu. Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications. *IEEE Transactions on Wireless Communications*, 20(1):375–388, 2020.

[11] Xiaozhen Lu, Liang Xiao, Guohang Niu, Xiangyang Ji, and Qian Wang. Safe exploration in wireless security: A safe reinforcement learning algorithm with hierarchical structure. *IEEE Transactions on Information Forensics and Security*, 17:732–743, 2022.

[12] Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

[13] Wenxuan Zhou, Sujay Bajracharya, and David Held. Plas: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*, pages 1719–1735. PMLR, 2021.

[14] Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, and Chelsea Finn. Offline reinforcement learning from images with latent space models. In *Learning for dynamics and control*, pages 1154–1168. PMLR, 2021.

[15] Seyedshams Feyzabadi. *Robot Planning with Constrained Markov Decision Processes*. University of California, Merced, 2017.

[16] Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

[17] Cory Jay Girard. *Structural results for constrained Markov decision processes*. Cornell University, 2018.

[18] Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient Q-learning with low switching cost. *arXiv preprint arXiv:1905.12849*, 2019.

[19] Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, Melissa Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84, 2010.

[20] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[21] Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. *Advances in Neural Information Processing Systems*, 35: 13303–13315, 2022.

[22] Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pages 3274–3307. PMLR, 2022.

[23] Yue Wang, Fei Miao, and Shaofeng Zou. Robust constrained reinforcement learning. *arXiv preprint arXiv:2209.06866*, 2022.

[24] Daniel J Mankowitz, Dan A Calian, Rae Jeong, Cosmin Paduraru, Nicolas Heess, Sumanth Dathathri, Martin Riedmiller, and Timothy Mann. Robust constrained reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:2010.10644*, 2020.

[25] Arnob Ghosh. Sample complexity for obtaining sub-optimality and violation bound for distributionally robust constrained MDP. In *First Reinforcement Learning Safety Workshop*, 2024.

[26] Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Constrained reinforcement learning under model mismatch. *arXiv preprint arXiv:2405.01327*, 2024.

[27] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

[28] Aria HasanzadeZonuzy, Archana Bura, Dileep Kalathil, and Srinivas Shakkottai. Learning with safety constraints: Sample complexity of reinforcement learning for constrained mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7667–7674, 2021.

[29] Francesco Emanuele Stradi, Jacopo Germano, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online learning in cmdps: Handling stochastic and adversarial constraints. In *Forty-first International Conference on Machine Learning*, 2024.

[30] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

[31] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc., 2020.

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

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

[34] Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

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

[36] Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free rl. In *International Conference on Artificial Intelligence and Statistics*, pages 1054–1062. PMLR, 2024.

[38] Yuxin Pan, Yize Chen, and Fangzhen Lin. Adjustable robust reinforcement learning for online 3D bin packing. *Advances in Neural Information Processing Systems*, 36, 2024.

[39] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36, 2023.

[40] He Wang, Laixi Shi, and Yuejie Chi. Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*, 2024.

[41] Sharan Vaswani, Lin Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained mdps. *Advances in Neural Information Processing Systems*, 35:3110–3122, 2022.

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

[43] Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

[44] Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *Advances in Neural Information Processing Systems*, 33:16315–16326, 2020.

[45] Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8030–8037, 2021.

[46] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

[47] Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in neural information processing systems*, 35:1047–1059, 2022.

[48] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International conference on artificial intelligence and statistics*, pages 3304–3312. PMLR, 2021.

[49] Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 243–253. PMLR, 2021.

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

[51] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent MDP and Markov game. In *The Eleventh International Conference on Learning Representations*, 2023.

[52] Zhishuai Liu and Pan Xu. Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. *arXiv preprint arXiv:2402.15399*, 2024.

[53] Zhishuai Liu and Pan Xu. Minimax optimal and computationally efficient algorithms for distributionally robust offline reinforcement learning. *arXiv preprint arXiv:2403.09621*, 2024.

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

[55] Shyam Sundhar Ramesh, Pier Giuseppe Sessa, Yifan Hu, Andreas Krause, and Ilija Bogunovic. Distributionally robust model-based reinforcement learning with large state spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2024.

[56] Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust constrained-mdps: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

[57] Toshinori Kitamura, Tadashi Kozuno, Wataru Kumagai, Kenta Hoshino, Yohei Hosoe, Kazumi Kasaura, Masashi Hamaya, Paavo Parmas, and Yutaka Matsuo. Near-optimal policy identification in robust constrained markov decision processes via epigraph form. *arXiv preprint arXiv:2408.16286*, 2024.

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

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

[60] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.

[61] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 19967–20025. PMLR, 2022.

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

[63] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

[64] Yifei Min, Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Variance-aware off-policy evaluation with linear function approximation. *Advances in Neural Information Processing Systems*, 34:7598–7610, 2021.

[65] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

A Technical Lemmas

Lemma 3 (Hoeffding-type inequality for self-normalized process [63]). *Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process and let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration such that η_t is $\mathcal{F}_t$-measurable. Let $\{x_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where x_t is $\mathcal{F}_{t-1}$ measurable and $x_t \le L$. Let $\Lambda_t = \lambda I_d + \sum_{s=1}^{t} x_s x_s^\top$. Assume that conditioned on $\mathcal{F}_{t-1}$, η_t is mean-zero and R-sub-Gaussian. Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t > 0$, we have*

$$\|\sum_{s=1}^{t} x_s \eta_s\|_{\Lambda_t^{-1}} \le R\sqrt{d\log(1+tL/\lambda) + 2\log(1/\delta)}.$$

Lemma 4 (Lemma 5.1 in [33]). *Under the condition that with probability at least $1-\delta$, the penalty function $\Gamma_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ in Algorithm 1 and satisfying*

$$|(\widehat{\mathbb{B}}_{r,h}^{\rho}\widehat{V}_{r,h+1})(s,a) - (\mathbb{B}_{r,h}^{\rho}\widehat{V}_{r,h+1})(s,a)| \le \Gamma_h(s,a), \quad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H], \tag{12}$$

we have

$$0 \le \iota_h(s,a) \le 2\Gamma_h(s,a), \quad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H].$$

Lemma 5 (Lemma 7 in [40]). *For any positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$ and any constant $c \ge 0$, we have*

$$\mathrm{Tr}\left(A(I+cA)^{-1}\right) \le \sum_{i=1}^{d} \frac{\lambda_i}{1+c\lambda_i},$$

where $\{\lambda_i\}_{i=1}^{d}$ are the eigenvalues of A and $\mathrm{Tr}(\cdot)$ denotes the trace of the given matrix.

Lemma 6 (Lemma H.5 in [64]). *Let $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ be a bounded function such that $\|\phi(s,a)\|_2 \le C$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. For any $K > 0$ and $\lambda > 0$, define $\bar{G}_K = \sum_{k=1}^{K} \phi(s_k,a_k)\phi(s_k,a_k)^\top + \lambda I_d$ where (s_k,a_k) are i.i.d. samples from some distribution ν over $\mathcal{S}\times\mathcal{A}$. Let $G = \mathbb{E}_v[\phi(s,a)\phi(s,a)^\top]$. Then for any $\delta \in (0,1)$, if K satisfies that*

$$K \ge \max\{512C^4\|G^{-1}\|^2\log(2d/\delta), 4\lambda\|G^{-1}\|\}.$$

Then with probability at least $1-\delta$, it holds simultaneously for all $u \in \mathbb{R}^d$ that

$$\|u\|_{\bar{G}_k^{-1}} \le \frac{2}{\sqrt{K}}\|u\|_{G^{-1}}.$$

Lemma 7 (Lemma 6 in [40]). *For any function $f_1 : \mathcal{C} \subseteq \mathbb{R} \to \mathbb{R}$ and $f_2 : \mathcal{C} \subseteq \mathbb{R} \to \mathbb{R}$, we have*

$$\max_{\alpha\in\mathcal{C}} f_1(\alpha) - \max_{\alpha\in\mathcal{C}} f_2(\alpha) \le \max_{\alpha\in\mathcal{C}} \left(f_1(\alpha) - f_2(\alpha)\right).$$

B Analysis for CROP-VI: Algorithm 1

B.1 Proof of Lemma 2

Proof. We prove this by contradiction. Assume the claim is false. Let $\pi^\star$ be the optimal policy for the original problem (3). Feasibility of $\pi^\star$ implies $(b - V_{1,g}^{\pi^\star,\rho}(\zeta))_+ = 0$, hence

$$V_{r,1}^{\pi^\star,\rho}(\zeta) - \lambda\left(b - V_{g,1}^{\pi^\star,\rho}(\zeta)\right)_+ \ge 0, \tag{13}$$

as $r_h(s,a) \in [0,1]$. For any policy π—in particular the optimizer $\widehat{\pi}$ of the rectified problem (5)—we have

$$V_{r,1}^{\pi,\rho}(\zeta) - \lambda\left(b - V_{g,1}^{\pi,\rho}(\zeta)\right)_+ \le H - \lambda\left(b - V_{g,1}^{\pi,\rho}(\zeta)\right)_+. \tag{14}$$

Algorithm 2: `Two-fold-subsampling`

Input: Batch dataset $\mathcal{D}$;

Split Data: Split $\mathcal{D}$ into two haves $\mathcal{D}^{\text{main}}$ and $\mathcal{D}^{\text{aux}}$, where $|\mathcal{D}^{\text{main}}| = |\mathcal{D}^{\text{aux}}| = N_h/2$. Denote $N_h^{\text{main}}(s)$ (resp. $N_h^{\text{aux}}(s)$) as the number of sample transitions from state s at step h in $\mathcal{D}^{\text{main}}$ (resp. $\mathcal{D}^{\text{aux}}$);

Construct the high-probability lower bound $N_h^{\text{trim}}(s)$ **by** $\mathcal{D}^{\text{aux}}$**:** For each $s \in \mathcal{S}$ and $1 \le h \le H$, compute

$$N_h^{\text{trim}}(s) = \max\{N_h^{\text{aux}}(s) - 10\sqrt{N_h^{\text{aux}}(s)\log\frac{KH}{\delta}}, 0\}. \tag{15}$$

Construct the almost temporally statistically independent $\mathcal{D}^{\text{trim}}$**:** Let $\mathcal{D}_h^{\text{main}}(s)$ denote the dataset containing all transition-reward sample tuples at the current state s and step h from $\mathcal{D}^{\text{main}}$. For any $(s,h) \in \mathcal{S} \times [H]$, subsample $\min\{N_h^{\text{trim}}(s), N_h^{\text{main}}(s)\}$ transition-reward sample tuples randomly from $\mathcal{D}_h^{\text{main}}(s)$, denoted as $\mathcal{D}^{\text{main,sub}}$;

Output: $\mathcal{D}^0 = \mathcal{D}^{\text{main,sub}}$.

By the contradictory assumption, $(b - V_{g,1}^{\widehat{\pi},\rho}(\zeta)) > \epsilon$. Substituting into (14) with $\lambda = H/\epsilon$ yields

$$V_{r,1}^{\widehat{\pi},\rho}(\zeta) - \lambda\left(b - V_{g,1}^{\widehat{\pi},\rho}(\zeta)\right)_+ < H - (H/\epsilon)\,\epsilon = 0,$$

which contradicts the lower bound (13). Therefore, the optimal policy of (5) violates the constraint by at most ϵ. Moreover, any infeasible policy must satisfy $b - V_{g,1}^{\pi,\rho}(\zeta) < \epsilon$, so only the feasible policies of the original robust CMDP (3) can be the optimal solution of (5). Hence,

$$V_{r,1}^{\widehat{\pi},\rho} - \lambda\left(b - V_{g,1}^{\widehat{\pi},\rho}\right)_+ \le V_{r,1}^{\pi^\star,\rho} - \lambda\left(b - V_{g,1}^{\pi^\star,\rho}\right)_+,$$

where we have used the fact that $(b - V_{g,1}^{\pi^\star,\rho})_+ = 0$, and $(b - V_{g,1}^{\widehat{\pi},\rho})_+ = 0$. Thus, the optimal solution of (3) is also optimal in (5). □

B.2 Two-fold subsampling method

To mitigate the temporal dependency in the batch dataset $\mathcal{D}$, we adopt the two-fold subsampling idea of [60]. The key idea is to utilize half of the data to establish a valid lower bound of the number of samples, which is employed to achieve the statistical independence in the remaining half of the dataset. The detailed implementation of the two-fold subsampling can be summarized in the Algorithm 2. Recall that we assume the sample trajectories in $\mathcal{D}$ are generated independently. Then, the following lemma, adapted from Lemma 9 in [40]—itself a slight modification of Lemma 3 and Lemma 7 in [60]—establishes that (15) is a valid lower bound of $N_h^{\text{main}}(s)$ for any $s \in \mathcal{S}$ and $h \in [H]$.

Lemma 8 (Lemma 9 in [40])**.** *With probability at least* $1 - 2\delta$*, if* $N_h^{\text{trim}}(s)$ *satisfies* (15) *for every* $s \in \mathcal{S}$ *and* $h \in [H]$*, then* $\mathcal{D}^0 := \mathcal{D}^{\text{main,sub}}$ *contains temporally statistically independent samples and the following bound holds simultaneously, i.e.,*

$$N_h^{\text{trim}}(s) \le N_h^{\text{main}}(s), \qquad \forall (s,h) \in \mathcal{S} \times [H].$$

In addition, with probability at least $1 - 3\delta$*, the following lower bound also holds, i.e.,*

$$N_h^{\text{trim}}(s,a) \ge \frac{K d_h^b(s,a)}{8} - 5\sqrt{K d_h^b(s,a)\log(\frac{KH}{\delta})}, \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

B.3 Proof of Theorem 1

We primarily focus on the model misspecification setting. Notably, when the model misspecification parameter $\psi = 0$, we can recover the results in Theorem 1.

Notations. Before presenting the proof of Theorem 1, we introduce several notations for clarity. We define the model evaluation error at the step h of CROP-VI as

$$\iota_{r,h}(s,a) = \mathbb{B}^{\rho}_{r,h}\widehat{V}_{r,h+1}(s,a) - \widehat{Q}_{r,h}(s,a), \quad \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}. \tag{16}$$

In addition, we denote the estimated weight of the transition kernel at the h-th step by

$$\forall (s,h) \in \mathcal{S} \times [H]: \quad \widehat{\mu}_h(s) = \Lambda_h^{-1} \left(\sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \mathbb{1}(s_{h+1}^\tau = s) \right) \in \mathbb{R}^d, \tag{17}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Accordingly, it holds that $\bar{\nu}_h^{\widehat{V}_r}(\alpha) = \int_{\mathcal{S}} \widehat{\mu}_h(s')[\widehat{V}_{h+1,r}(s')]_\alpha \mathrm{d}s' \in \mathbb{R}^d$. We denote the set of all the possible state occupancy distributions associated with the optimal policy $\pi^\star$ and any $P \in \mathcal{P}^\rho(P^0)$ as

$$\mathcal{D}_h^\star = \left\{ \left[d_h^{\star,P}(s) \right]_{s \in \mathcal{S}} : P \in \mathcal{P}^\rho(P^0) \right\} = \left\{ \left[d_h^{\star,P}(s, \pi_h^\star(s)) \right]_{s \in \mathcal{S}} : P \in \mathcal{P}^\rho(P^0) \right\}, \tag{18}$$

for any time step $h \in [H]$.

B.3.1 Proof sketch for Theorem 1

We first claim that Theorem 1 holds as long as the following theorem can be established.

Theorem 2. *Consider $\delta \in (0,1)$. Suppose that the dataset $\mathcal{D}_0$ in Algorithm 1 contains $N_h < K$ transition-reward sample tuples at every $h \in [H]$. Assume that conditional on $\{N_h\}_{h\in[H]}$, all the sample tuples in $\mathcal{D}_h^0$ are statistically independent. Suppose that Assumption 1, 2, and 5 hold. In CROP-VI, we set*

$$\lambda_0 = 1, \quad \gamma_0 = 6\sqrt{d\xi_0}H + \sqrt{d}\psi H, \quad \textit{where } \xi_0 = \log(3HK/\delta). \tag{19}$$

Here, $\delta \in (0,1)$ is the confidence parameter and K is the upper bound of N_h for any $h \in [H]$. Then, $\{\widehat{\pi}_h\}_{h=1}^H$ generated by Algorithm 1, with the probability at least $1-\delta$, satisfies

$$\text{Sub-Opt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \le \tilde{\mathcal{O}}(\sqrt{d}H \cdot \max\{1,\psi\}) \sum_{h=1}^H \sum_{i=1}^d \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{d_h^\star} \left[\|\phi_i(s_h,a_h)\mathbb{1}_i\|_{\Lambda_h^{-1}} \right] + \frac{H(H-1)\psi}{2}.$$

As the construction in Algorithm 2, $\{N_h^{\mathrm{trim}}(s)\}_{s\in\mathcal{S},h\in[H]}$ is computed using $\mathcal{D}^{\mathrm{aux}}$ that is independent of $\mathcal{D}^0$. From Lemma 8, $N_h^{\mathrm{trim}}(s)$ is a valid sampling number for any $s \in \mathcal{S}$ and $h \in [H]$ such that $|\mathcal{D}_h^0| = \sum_{s\in\mathcal{S}} N_h^{\mathrm{trim}}(s) \le K$, and $\mathcal{D}_h^0$ can be treated as containing temporally statistically independent samples with probability exceeding $1-2\delta$. Therefore, by invoking Theorem 2 with $N_h := |\mathcal{D}_h^0|$, we have

$$\text{Sub-Opt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \le \tilde{\mathcal{O}}(\sqrt{d}H \cdot \max\{1,\psi\}) \sum_{h=1}^H \sum_{i=1}^d \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{d_h^\star} \left[\|\phi_i(s_h,a_h)\mathbb{1}_i\|_{\Lambda_h^{-1}} \right] + \frac{H(H-1)\psi}{2},$$

with probability exceeding $1-3\delta$.

B.3.2 Proof of Theorem 2

The sub-optimality bound. The argument unfolds in three steps.

Step 1: establishing the pessimistic property of the reward function and the optimistic property of the utility function. We first substantiate the pessimism of the reward function, which heavily depends on the following lemmas.

Lemma 9 (Adapted from Lemma 10 in [40])**.** *Suppose all the assumptions in Theorem 2 hold and follow all the parameters setting in* (19)*. Then for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, with probability at least $1-\delta$, the value/utility function $\{\widehat{V}_{j,h}\}_{h=1}^H$ for $j = r, g$ generated by CROP-VI satisfies*

$$|(\widehat{\mathbb{B}}^\rho_{j,h}\widehat{V}_{j,h+1})(s,a) - (\mathbb{B}^\rho_{j,h}\widehat{V}_{j,h+1})(s,a)| \le \Gamma_h(s,a) := \gamma_0 \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} + (H-h)\psi. \tag{20}$$

In the following, we will show that the following relations hold:

$$Q^{\star,\rho}_{r,h}(s,a) \geq Q^{\widehat{\pi},\rho}_{r,h}(s,a) \geq \widehat{Q}_{r,h}(s,a) \quad \text{and} \quad V^{\star,\rho}_{r,h}(s) \geq V^{\widehat{\pi},\rho}_{r,h}(s) \geq \widehat{V}_{r,h}(s), \tag{21}$$

for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, if the condition (20) holds. It implies that $\widehat{Q}_{r,h}(s,a)$ and $\widehat{V}_{r,h}(s)$ is the pessimistic estimates of $Q^{\widehat{\pi},\rho}_{r,h}(s,a)$ and $V^{\widehat{\pi},\rho}_{r,h}(s)$ for any $s \in \mathcal{S}$, respectively. Notice that if $Q^{\widehat{\pi},\rho}_{r,h}(s,a) \geq \widehat{Q}_{r,h}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, one simultaneously has the following relation:

$$V^{\widehat{\pi},\rho}_{r,h}(s) = Q^{\widehat{\pi},\rho}_{r,h}(s,\widehat{\pi}_h(s)) \geq \widehat{Q}_{r,h}(s,\widehat{\pi}_h(s)) = \widehat{V}_{r,h}(s), \qquad \forall (s,h) \in \mathcal{S} \times [H].$$

Therefore, we shall verify that

$$Q^{\widehat{\pi},\rho}_{r,h}(s,a) \geq \widehat{Q}_{r,h}(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{22}$$

by induction, and $V^{\widehat{\pi},\rho}_{r,h}(s) \geq \widehat{V}_{r,h}(s)$ will spontaneously hold for $s \in \mathcal{S}$.

- *At step* $h = H+1$: From the initialization step in Algorithm 1, we have $Q^{\widehat{\pi},\rho}_{r,H+1}(s,a) = \widehat{Q}_{r,H+1}(s,a) = 0$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, and (22) holds.
- *For any step* $h \leq H$: Suppose $Q^{\widehat{\pi},\rho}_{r,h+1}(s,a) \geq \widehat{Q}_{r,h+1}(s,a)$. From (21), we have $V^{\widehat{\pi},\rho}_{r,h+1}(s) \geq \widehat{V}_{r,h+1}(s)$. Therefore, if $\widehat{Q}_{r,h}(s,a) = 0$, $Q^{\widehat{\pi},\rho}_{r,h}(s,a) \geq 0 = \widehat{Q}_{r,h}(s,a)$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Otherwise,

$$\begin{aligned}
\widehat{Q}_{r,h}(s,a) &\leq (\widehat{\mathbb{B}}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a) - \Gamma_h(s,a) \\
&= (\mathbb{B}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a) + (\widehat{\mathbb{B}}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a) - (\mathbb{B}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a) - \Gamma_h(s,a) \\
&\leq (\mathbb{B}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a) + |(\widehat{\mathbb{B}}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a) - (\mathbb{B}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a)| - \Gamma_h(s,a) \\
&\leq (\mathbb{B}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a) + \Gamma_h(s,a) - \Gamma_h(s,a) \\
&\leq (\mathbb{B}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a) \\
&\leq (\mathbb{B}^{\rho}_{r,h}V^{\widehat{\pi},\rho}_{r,h+1})(s,a) = Q^{\widehat{\pi},\rho}_{r,h}(s,a),
\end{aligned}$$

  where the first inequality is from the definition of $\widehat{Q}_{r,h}(s,a)$ (cf. Line 14 in Algorithm 1), and third inequality is based on the condition (20).

Combining these two cases, for any $h \in [H+1]$, we could verify the pessimistic property, i.e. the equation (21).

Next, we establish the optimism of the utility function, leveraging Lemma 9. Building on this result, we will demonstrate that the following relations hold:

$$Q^{\widehat{\pi},\rho}_{g,h}(s,a) \leq \widehat{Q}_{g,h}(s,a) \quad \text{and} \quad V^{\widehat{\pi},\rho}_{g,h}(s) \leq \widehat{V}_{g,h}(s), \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{23}$$

if the condition (20) holds. It implies that $\widehat{Q}_{g,h}(s,a)$ and $\widehat{V}_{g,h}(s)$ is the optimistic estimates of $Q^{\widehat{\pi},\rho}_{g,h}(s,a)$ and $V^{\widehat{\pi},\rho}_{g,h}(s)$ for any $s \in \mathcal{S}$, respectively. Notice that if $Q^{\widehat{\pi},\rho}_{g,h}(s,a) \leq \widehat{Q}_{g,h}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, one simultaneously has the following relation:

$$V^{\widehat{\pi},\rho}_{g,h}(s) = Q^{\widehat{\pi},\rho}_{g,h}(s,\widehat{\pi}_h(s)) \leq \widehat{Q}_{g,h}(s,\widehat{\pi}_h(s)) = \widehat{V}_{g,h}(s), \qquad \forall (s,h) \in \mathcal{S} \times [H].$$

Therefore, we shall verify that

$$Q^{\widehat{\pi},\rho}_{g,h}(s,a) \leq \widehat{Q}_{g,h}(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{24}$$

by induction, and $V^{\widehat{\pi},\rho}_{g,h}(s) \leq \widehat{V}_{g,h}(s)$ will spontaneously hold for $s \in \mathcal{S}$.

- *At step* $h = H+1$: From the initialization step in Algorithm 1, we have $Q^{\widehat{\pi},\rho}_{g,H+1}(s,a) = \widehat{Q}_{g,H+1}(s,a) = 0$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, and (24) holds.

- *For any step* $h \le H$: Suppose $Q^{\widehat{\pi},\rho}_{g,h+1}(s,a) \le \widehat{Q}_{g,h+1}(s,a)$. From (23), we have $V^{\widehat{\pi},\rho}_{g,h+1}(s) \le \widehat{V}_{g,h+1}(s)$. Therefore, if $\widehat{Q}_{g,h}(s,a) = H-h+1$, $Q^{\widehat{\pi},\rho}_{g,h}(s,a) \le H-h+1 = \widehat{Q}_{g,h}(s,a)$, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$. Otherwise,

$$\begin{aligned}
Q^{\widehat{\pi},\rho}_{g,h}(s,a) &= (\mathbb{B}^{\rho}_{g,h} V^{\widehat{\pi},\rho}_{g,h+1})(s,a) \\
&\le (\mathbb{B}^{\rho}_{g,h} \widehat{V}_{g,h+1})(s,a) \\
&\le (\widehat{\mathbb{B}}^{\rho}_{g,h} \widehat{V}_{g,h+1})(s,a) + (\mathbb{B}^{\rho}_{g,h} \widehat{V}_{g,h+1})(s,a) - (\widehat{\mathbb{B}}^{\rho}_{g,h} \widehat{V}_{g,h+1})(s,a) \\
&\le (\widehat{\mathbb{B}}^{\rho}_{g,h} \widehat{V}_{g,h+1})(s,a) + |(\widehat{\mathbb{B}}^{\rho}_{g,h} \widehat{V}_{g,h+1})(s,a) - (\mathbb{B}^{\rho}_{g,h} \widehat{V}_{g,h+1})(s,a)| \\
&\le (\mathbb{B}^{\rho}_{g,h} \widehat{V}_{g,h+1})(s,a) + \Gamma_h(s,a) = \widehat{Q}_{g,h}(s,a),
\end{aligned}$$

  where the first inequality is from the definition of $\widehat{Q}_{g,h}(s,a)$ (cf. Line 15 in Algorithm 1), and third inequality is based on the condition (20).

Combining these two cases, for any $h \in [H+1]$, we could verify the optimistic property, i.e. the equation (23).

Step 2: bounding the suboptimality gap. Notice that for any $h \in [H]$ and any $s \in \mathcal{S}$,

$$V^{\star,\rho}_{r,h}(s) - V^{\widehat{\pi},\rho}_{r,h}(s) = \underbrace{V^{\star,\rho}_{r,h}(s) - \widehat{V}^{\pi^\star,\rho}_{r,h}(s)}_{T_1} + \underbrace{\widehat{V}^{\pi^\star,\rho}_{r,h}(s) - \widehat{V}^{\widehat{\pi},\rho}_{r,h}(s)}_{T_2} + \underbrace{\widehat{V}^{\widehat{\pi},\rho}_{r,h}(s) - V^{\widehat{\pi},\rho}_{r,h}(s)}_{T_3}.$$

We first control T_1. For any $s \in \mathcal{S}$,

$$V^{\star,\rho}_{r,h}(s) - \widehat{V}^{\pi^\star,\rho}_{r,h}(s) = Q^{\star}_{r,h}(s,\pi^\star(s)) - \widehat{Q}^{\pi^\star}_{r,h}(s,\pi^\star(s)). \tag{25}$$

From the definition of the model evaluation error (i.e., equation (16)) and the robust Bellman optimality equation, we have

$$\begin{aligned}
&\widehat{Q}_{r,h}(s,a) = (\mathbb{B}^{\rho}_{r,h}\widehat{V}_{r,h+1})(s,a) - \iota_{r,h}(s,a), && \forall (s,a)\in\mathcal{S}\times\mathcal{A}, \\
&Q^{\star,\rho}_{r,h}(s,a) = (\mathbb{B}^{\rho}_{r,h} V^{\star,\rho}_{r,h+1})(s,a) && \forall (s,a)\in\mathcal{S}\times\mathcal{A},
\end{aligned}$$

which leads to

$$\begin{aligned}
&Q^{\star,\rho}_{r,h}(s,\pi^\star_h(s)) - \widehat{Q}_{r,h}(s,\pi^\star_h(s)) \\
=&(\mathbb{P}^{\rho}_h V^{\star,\rho}_{r,h+1})(s,\pi^\star_h(s)) - (\mathbb{P}^{\rho}_h \widehat{V}_{r,h+1})(s,\pi^\star_h(s)) + \iota_{r,h}(s,\pi^\star_h(s)), \quad \forall s\in\mathcal{S}.
\end{aligned} \tag{26}$$

Denote

$$P^{\inf,\widehat{V}_r}_{h,s,\pi^\star_h(s)}(\cdot) := \operatorname*{arg\,min}_{P(\cdot)\in\mathcal{P}^{\rho}(P^0_{h,s,\pi^\star_h(s)})} \int_S P(s')\widehat{V}_{r,h+1}(s')\mathrm{d}s'.$$

Therefore, (26) becomes

$$\begin{aligned}
&Q^{\star,\rho}_{r,h}(s,\pi^\star_h(s)) - \widehat{Q}_{r,h}(s,\pi^\star_h(s)) \\
\le &\int_{\mathcal{S}} P^{\inf,\widehat{V}_r}_{h,s,\pi^\star_h(s)}(s') \left(V^{\star,\rho}_{r,h+1}(s') - \widehat{V}_{r,h+1}(s')\right)\mathrm{d}s' + \iota_{r,h}(s,\pi^\star_h(s)), \ \ \forall (s,a)\in\mathcal{S}\times\mathcal{A}.
\end{aligned} \tag{27}$$

Substituting (27) into (28), one has

$$V^{\pi^\star,\rho}_{r,h}(s) - \widehat{V}^{\widehat{\pi},\rho}_{r,h}(s) \le \int_{\mathcal{S}} P^{\inf,\widehat{V}_r}_{h,s,\pi^\star_h(s)}(s') \left(V^{\star,\rho}_{r,h+1}(s') - \widehat{V}_{r,h+1}(s')\right)\mathrm{d}s' + \iota_{r,h}(s,\pi^\star_h(s)).$$

For any $h\in[H]$, define $\widehat{P}^{\inf}_{h,s} : \mathcal{S}\to\mathcal{S}$ and $\iota^\star_{r,h}\in\mathcal{S}\to\mathbb{R}$ by

$$\widehat{P}^{\inf}_h(s) = P^{\inf,\widehat{V}_r}_{h,s,\pi^\star_h(s)}(\cdot) \quad \text{and} \quad \text{and} \quad \iota^\star_{r,h}(s) := \iota_{r,h}(s,\pi^\star(s)), \quad \forall s\in\mathcal{S}.$$

By telescoping sum, we finally obtain that for any $s \in \mathcal{S}$,

$$\begin{aligned} V_{r,h}^{\star,\rho}(s) - \widehat{V}_{r,h}(s) =& \langle \mathbb{1}_s, V_{r,h}^{\star,\rho} - \widehat{V}_{r,h} \rangle \\ &\leq \left(\prod_{j=h}^{H} \widehat{P}_j^{\inf} \right) \left(V_{r,H+1}^{\star,\rho} - \widehat{V}_{r,H+1} \right)(s) + \sum_{t=h}^{H} \left(\prod_{j=h}^{t-1} \widehat{P}_j^{\inf} \right) \iota_{r,t}^{\star}(s) \\ &= \sum_{t=h}^{H} \left(\prod_{j=h}^{t-1} \widehat{P}_j^{\inf} \right) \iota_{r,t}^{\star}(s), \end{aligned}$$

where the equality is from $V_{r,H+1}^{\star,\rho}(s) = \widehat{V}_{r,H+1}(s) = 0$ and $\left(\prod_{j=t}^{t-1} \widehat{P}_j^{\inf} \right)(s) = \mathbb{1}_s$.

We now bound the term T_2 which we describe next.

$$\begin{aligned} T_2 &= \widehat{Q}_{r,h}(s, \pi_h^{\star}(s)) - \widehat{Q}_{r,h}(s, \widehat{\pi}_h(s)) \\ &\leq \widehat{Q}_{r,h}(s, \pi_h^{\star}(s)) - \lambda \left(b - \widehat{Q}_{g,h}(s, \pi^{\star}(s)) \right)_+ - \left[\widehat{Q}_{r,h}(s, \widehat{\pi}_h(s)) - \lambda \left(b - \widehat{Q}_{g,h}(s, \pi^{\star}(s)) \right)_+ \right] \\ &\leq \widehat{Q}_{r,h}(s, \widehat{\pi}_h(s)) - \lambda \left(b - \widehat{Q}_{g,h}(s, \widehat{\pi}(s)) \right)_+ - \left[\widehat{Q}_{r,h}(s, \widehat{\pi}_h(s)) - \lambda \left(b - \widehat{Q}_{g,h}(s, \pi^{\star}(s)) \right)_+ \right] \\ &\leq \lambda \left(b - \widehat{Q}_{g,h}(s, \pi^{\star}(s)) \right)_+ - \lambda \left(b - \widehat{Q}_{g,h}(s, \widehat{\pi}(s)) \right)_+ . \end{aligned}$$

Thus, if the optimism holds constraints, i.e., $\widehat{Q}_{g,h}(s, \pi^{\star}(s)) \geq Q_{g,h}(s, \pi^{\star}(s))$, then

$$T_2 \leq \lambda \left(b - Q_{g,h}(s, \pi^{\star}(s)) \right)_+ - \lambda \left(b - \widehat{Q}_{g,h}(s, \widehat{\pi}(s)) \right)_+ \leq -\lambda \left(b - \widehat{Q}_{g,h}(s, \widehat{\pi}(s)) \right)_+ \leq 0.$$

Then, we bound the term T_3. We have

$$\widehat{V}_{r,h}^{\widehat{\pi},\rho}(s) - V_{r,h}^{\widehat{\pi},\rho}(s) = \widehat{Q}_{r,h}(s, \widehat{\pi}(s)) - Q_{r,h}(s, \widehat{\pi}(s)) \leq 0, \ \ \forall s \in \mathcal{S}, \tag{28}$$

where the inequality follows from the pessimistic property of $Q_{r,h}^{\widehat{\pi},\rho}(s,a)$.

Step 3: finishing up. For any $d_h^{\star} \in \mathcal{D}_h^{\star}$, denote

$$d_{h:t}^{\star} = d_h^{\star} \left(\prod_{j=h}^{t-1} \widehat{P}_j^{\inf} \right) \in \mathcal{D}_t^{\star}.$$

The sub-optimality gap defined in (4) satisfies

$$\text{Sub-Opt}(\widehat{\pi}; \zeta, \mathcal{P}^{\rho}) \leq \mathbb{E}_{s_1 \sim \zeta} V_{r,1}^{\star,\rho}(s_1) - \mathbb{E}_{s_1 \sim \zeta} \widehat{V}_{r,1}(s_1) \leq \sum_{t=1}^{H} \mathbb{E}_{s_t \sim d_{1:t}^{\star}} \iota_t^{\star}(s_t) + \frac{(H-1)H\psi}{2}.$$

For any $h \in [H]$, we let $\Gamma_h^{\star} : \mathcal{S} \to \mathbb{R}$ satisfy

$$\Gamma_h^{\star}(s) = \Gamma_h(s, \pi_h^{\star}(s)), \quad \forall s \in \mathcal{S}.$$

Combining Lemma 4 together with Lemma 9 will lead to

$$\text{Sub-Opt}(\widehat{\pi}; \zeta, \mathcal{P}^{\rho}) \leq 2 \sum_{h=1}^{H} \mathbb{E}_{s_h \sim d_{1:h}^{\star}} \Gamma_h^{\star}(s_h) + \frac{(H-1)H\psi}{2}.$$

Note that $\Gamma_h^\star(s) = \gamma_0 \sum_{i=1}^d \|\phi_i(s, \pi_h^\star(s))\mathbb{1}_i\|_{\Lambda_h^{-1}}$ for any $(s,h) \in \mathcal{S} \times [H]$. Following the definition (18), we have $d_{1:h}^\star \in \mathcal{D}_h^\star$ and correspondingly,

$$\begin{aligned}
\text{Sub-Opt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) &\leq 2\sum_{h=1}^H \mathbb{E}_{s_h \sim d_{1:h}^\star} \Gamma_h^\star(s_h) + \frac{(H-1)H\psi}{2} \\
&\leq 2\gamma_0 \sum_{h=1}^H \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{(s_h,a_h)\sim d_h^\star} \left[\sum_{i=1}^d \|\phi_i(s_h,a_h)\mathbb{1}_i\|_{\Lambda_h^{-1}}\right] + \frac{(H-1)H\psi}{2} \\
&\leq 2\gamma_0 \sum_{h=1}^H \sum_{i=1}^d \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{d_h^\star} \left[\|\phi_i(s_h,a_h)\mathbb{1}_i\|_{\Lambda_h^{-1}}\right] + \frac{(H-1)H\psi}{2}.
\end{aligned}$$

with probability exceeding $1-\delta$.

Violation bound. Now, we show the violation bound. We prove this by contradiction. Suppose that the statement is not true. Consider π^* the optimal solution for the original problem (3). Since this is feasible, then $\left(b - V_{1,g}^{\pi^\star,\rho}(\zeta)\right)_+ = 0$, then

$$V_{r,1}^{\pi^\star,\rho}(\zeta) - \lambda\left(b - V_{g,1}^{\pi^\star,\rho}(\zeta)\right)_+ \geq 0, \tag{29}$$

as $r_h(s,a) \in [0,1]$. For any π (including the optimal policy $\widehat{\pi}$) of (5), we have

$$V_{r,1}^{\pi,\rho}(\zeta) - \lambda\left(b - V_{g,1}^{\pi,\rho}(\zeta)\right)_+ \leq H - \lambda\left(b - V_{g,1}^{\pi,\rho}(\zeta)\right)_+. \tag{30}$$

For the optimal policy $\widehat{\pi}$, we have $(b - V_{g,1}^{\widehat{\pi},\rho}(\zeta)) > \epsilon$ (by contradiction). Hence, by (30), we have

$$V_{r,1}^{\widehat{\pi},\rho}(\zeta) - \lambda\left(b - V_{g,1}^{\widehat{\pi},\rho}(\zeta)\right)_+ < H - (H/\epsilon)\epsilon = 0.$$

However, it contradicts (29) as $\pi^\star$ can achieve a better value for the objective in (5).

B.3.3 Proof of Lemma 9

To control $|(\widehat{\mathbb{B}}_{j,h}^\rho \widehat{V}_{j,h+1})(s,a) - (\mathbb{B}_{j,h}^\rho \widehat{V}_{j,h+1})(s,a)|$ for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and $j \in \{r,g\}$, we first show the following lemma, where the proof can be found in Appendix B.3.4.

Lemma 10 (Adapted from Lemma 11 in [40])**.** *Suppose Assumption 1, 2 and 5 hold. Then, for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, the estimated value/utility function $\widehat{V}_{j,h+1}(j=r,g)$ generated by* `CROP-VI` *satisfies*

$$\begin{aligned}
&|(\widehat{\mathbb{B}}_{j,h}^\rho \widehat{V}_{j,h+1})(s,a) - (\mathbb{B}_{j,h}^\rho \widehat{V}_{j,h+1})(s,a)| \\
&\leq \left(\sqrt{\lambda_0 d}H + \sqrt{d}\xi H\right)\sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} + (H-h)\xi \\
&\quad + \underbrace{\max_{\alpha\in[\min_s \widehat{V}_{j,h+1}(s), \max_s \widehat{V}_{j,h+1}(s)]} \|\sum_{\tau\in\mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)\epsilon_h^\tau(\alpha, \widehat{V}_{j,h+1})\|_{\Lambda_h^{-1}}}_{T_{4,h}} \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}},
\end{aligned} \tag{31}$$

where $\epsilon_h^\tau(\alpha, V_j) = \int_{\mathcal{S}} P_h^0(s'|s_h^\tau, a_h^\tau)[V_j]_\alpha(s')\mathrm{d}s' - [V_j]_\alpha(s_{h+1}^\tau)$ for any value/utility function $V_j : \mathcal{S} \to [0,H]$, $\alpha \in [\min_s V_j(s), \max_s V_j(s)]$ and $\tau \in \mathcal{D}_h^0$.

We observe that the second term (i.e., $T_{4,h}$) in (31) will become dominating, as long as λ_0 and ψ are sufficiently small. In the following analysis, we will control $T_{4,h}$ via uniform concentration and the concentration of the self-normalized process.

Notice that α and $\widehat{V}_{j,h+1}$ are coupled with each other, which makes controlling $T_{4,h}$ intractable. To this end, we propose the minimal ϵ_0-covering set for α. Since $\widehat{V}_{j,h+1}(s) \in [0,H]$ for any $s \in \mathcal{S}$, we

construct $\mathcal{N}(\epsilon_0, H)$ as the minimal ϵ_0-cover of the $[0, H]$ whose size satisfies $|\mathcal{N}(\epsilon_0, H)| \leq \frac{3H}{\epsilon_0}$. In other words, for any $\alpha \in [0, H]$, there exists $\alpha^\dagger \in \mathcal{N}(\epsilon_0, H)$, we have

$$|\alpha - \alpha^\dagger| \leq \epsilon_0.$$

Then we can rewrite $T_{4,h}^2$ as

$$\begin{aligned}
T_{4,h}^2 &= \max_{\alpha \in [\min_s \widehat{V}_{j,h+1}(s), \max_s \widehat{V}_{j,h+1}(s)]} \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \left(\epsilon_h^\tau(\alpha, \widehat{V}_{j,h+1}) - \epsilon_h^\tau(\alpha^\dagger, \widehat{V}_{j,h+1}) + \epsilon_h^\tau(\alpha^\dagger, \widehat{V}_{j,h+1}) \right) \|_{\Lambda_h^{-1}}^2 \\
&\leq \max_{\alpha \in [0,H]} 2 \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \left(\epsilon_h^\tau(\alpha, \widehat{V}_{j,h+1}) - \epsilon_h^\tau(\alpha^\dagger, \widehat{V}_{j,h+1}) \right) \|_{\Lambda_h^{-1}}^2 + 2 \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha^\dagger, \widehat{V}_{j,h+1}) \|_{\Lambda_h^{-1}}^2 \\
&\leq 8\epsilon_0^2 K^2/\lambda_0 + 2 \underbrace{\| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha^\dagger, \widehat{V}_{j,h+1}) \|_{\Lambda_h^{-1}}^2}_{T_{5,h}}, \qquad (32)
\end{aligned}$$

for some $\alpha^\dagger \in \mathcal{N}(\epsilon_0, H)$, where the proof of the last inequality is postponed to Appendix B.3.5. Alternatively,

$$T_{5,h} \leq \sup_{\alpha \in \mathcal{N}(\epsilon_0, H)} \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, \widehat{V}_{j,h+1}) \|_{\Lambda_h^{-1}}^2. \qquad (33)$$

Noted that the samples in $\mathcal{D}^0$ are temporally statistically independent, i.e., $\widehat{V}_{j,h+1}$ is independent of $\mathcal{D}_h^0$, or to say, $\widehat{\mu}_h$. Therefore, we can directly control $T_{5,h}$ via the following lemma.

Lemma 11 (Concentration of self-normalized process, adapted from Lemma 12 in [40])**.** *Let $V_j : \mathcal{S} \to [0, H]$ be any fixed vector that is independent of $\widehat{\mu}_h$ and $\alpha \in [0, H]$ be a fixed constant. For any fixed $h \in [H]$ and any $\delta \in (0, 1)$, we have*

$$P_{\mathcal{D}} \left(\| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, V_j) \|_{\Lambda_h^{-1}}^2 > H^2 \left(2 \cdot \log(1/\delta) + d \cdot \log(1 + N_h/\lambda_0) \right) \right) \leq \delta.$$

The proof of Lemma 11 is postponed to Appendix B.3.6. Then applying Lemma 11 and the union bound over $\mathcal{N}(\epsilon_0, H)$, we have

$$\begin{aligned}
&P_{\mathcal{D}} \left(\sup_{\alpha \in \mathcal{N}(\epsilon_0, H)} \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, \widehat{V}_{j,h+1}) \|_{\Lambda_h^{-1}}^2 \geq H^2 (2 \log(H |\mathcal{N}(\epsilon_0, H)|/\delta) + d \log(1 + N_h/\lambda_0)) \right) \\
&\leq \delta/H,
\end{aligned}$$

for any fixed $h \in [H]$. According to [65], one has $|\mathcal{N}(\epsilon_0, H)| \leq \frac{3H}{\epsilon_0}$. Taking the union bound for any $h \in [H]$, we arrive at

$$\sup_{\alpha \in \mathcal{N}(\epsilon_0, H)} \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, \widehat{V}_{j,h+1}) \|_{\Lambda_h^{-1}}^2 \leq 2H^2 \log(\frac{3H^2}{\epsilon_0 \delta}) + H^2 d \log(1 + \frac{K}{\lambda_0}), \qquad (34)$$

with probability exceeding $1 - \delta$, where we utilize $N_h \leq K$ for every $h \in [H]$ on the right-hand side.

Combining (32), (33) and (34), we have

$$T_{4,h}^2 \leq 8\epsilon_0^2 K^2/\lambda_0 + 4H^2 \log(\frac{3H^2}{\epsilon_0 \delta}) + 2H^2 d \log(1 + \frac{K}{\lambda_0}),$$

with probability at least $1 - \delta$. Let $\epsilon_0 = H/K$ and $\lambda_0 = 1$. Then,

$$T_{4,h}^2 \leq 8H^2 + 4H^2 \log(\frac{3HK}{\delta}) + 2H^2 d \log(1 + K) \leq 8H^2 + 4H^2 \log(3HK/\delta) + 2dH^2 \log(2K).$$

Let $\xi_0 = \log(3HK/\delta) \geq 1$. Note that $\log(2K) \leq \log(3HK/\delta) = \xi_0$. Then, we have

$$T_{4,h}^2 \leq 8H^2 + 4H^2 \xi_0 + 2dH^2 \xi_0 \leq 16 dH^2 \xi_0.$$

Therefore, with probability exceeding $1-\delta$, one has

$$\begin{aligned}
&|(\widehat{\mathbb{B}}^{\rho}_{j,h}\widehat{V}_{j,h+1})(s,a)-(\mathbb{B}^{\rho}_{j,h}\widehat{V}_{j,h+1})(s,a)| \\
\leq&\left(\sqrt{d}H+4\sqrt{d\xi_0}H+\sqrt{d}\psi H\right)\sum_{i=1}^{d}\|\phi_i(s,a)\|_{\Lambda_h^{-1}}+(H-h)\psi \\
\leq&\gamma_0\sum_{i=1}^{d}\|\phi_i(s_h,a_h)\|_{\Lambda_h^{-1}}+(H-h)\psi=\Gamma_h(s,a),
\end{aligned}$$

where $\gamma_0=6\sqrt{d}H(\xi_0+\psi)$ and the above inequality satisfies (12).

B.3.4 Proof of Lemma 10

Proof. From the DRO Bellman optimality equation, we denote

$$\begin{aligned}
&(\mathbb{B}_{j,h}\widehat{V}_{j,h+1})(s,a) \\
=&j_h(s,a)+\inf_{P_{h+1}\in\mathcal{P}^{\rho}(P^0_{h+1})}\mathbb{E}_{P_{h+1}(\cdot|s,a)}[\widehat{V}_{j,h+1}(s')] \\
=&j_h(s,a)+\inf_{\widetilde{P}_{h+1}\in\mathcal{P}^{\rho}(\widetilde{P}^0_{h+1})}\mathbb{E}_{\widetilde{P}_{h+1}(\cdot|s,a)}[\widehat{V}_{j,h+1}(s')] \\
&\quad+(\inf_{P_{h+1}\in\mathcal{P}^{\rho}(P^0_{h+1})}\mathbb{E}_{P_{h+1}(\cdot|s,a)}[\widehat{V}_{j,h+1}(s')]-\inf_{\widetilde{P}_{h+1}\in\mathcal{P}^{\rho}(\widetilde{P}^0_{h+1})}\mathbb{E}_{\widetilde{P}_{h+1}(\cdot|s,a)}[\widehat{V}_{j,h+1}(s')]) \\
=&\phi(s,a)^{\top}\theta_{j,h} \\
&+\sum_{i=1}^{d}\phi_i(s,a)\max_{\alpha\in[\min_s\widehat{V}_{j,h+1}(s),\max_s\widehat{V}_{j,h+1}(s)]}\int_{\mathcal{S}}\mu^0_{h,i}(s')[\widehat{V}_{j,h+1}]_{\alpha}(s')\mathrm{d}s'-\rho(\alpha-\min_{s'}[\widehat{V}_{j,h+1}]_{\alpha}(s')) \\
&+(H-h)\xi.
\end{aligned}$$

Combined with the empirical Bellman operator in our algorithm,

$$(\mathbb{B}^{\rho}_{j,h}\widehat{V}_{j,h+1})(s,a)-(\widehat{\mathbb{B}}^{\rho}_{j,h}\widehat{V}_{j,h+1})(s,a)=\phi(s,a)^{\top}(\theta_{j,h}-\widehat{\theta}_{j,h})+\phi(s,a)^{\top}(\widehat{\nu}_h^{\rho,\widehat{V}_j}-\nu_h^{\rho,\widehat{V}_j})+(H-h)\psi.$$

Step 1: We analyze the error in the reward estimation, i.e., $\phi(s,a)^{\top}(\theta_{j,h}-\widehat{\theta}_{j,h})$.

$$\begin{aligned}
\phi(s,a)^{\top}(\theta_{j,h}-\widehat{\theta}_{j,h})&=\phi(s,a)^{\top}\Lambda_h^{-1}\Lambda_h\theta_{j,h}-\phi(s,a)^{\top}\Lambda_h^{-1}[\sum_{\tau=1}^{N}\phi(s_h^{\tau},a_h^{\tau})j_h(s_h^{\tau},a_h^{\tau})] \\
&=\phi(s,a)^{\top}\Lambda_h^{-1}\Lambda_h\theta_{j,h}-\phi(s,a)^{\top}\Lambda_h^{-1}[\sum_{\tau=1}^{N}\phi(s_h^{\tau},a_h^{\tau})\phi(s_h^{\tau},a_h^{\tau})^{\top}\theta_{j,h}] \\
&=\phi(s,a)^{\top}\Lambda_h^{-1}\Lambda_h\theta_{j,h}-\phi(s,a)^{\top}\Lambda_h^{-1}(\Lambda_h-\lambda_0 I)\theta_{j,h} \\
&=\lambda_0\phi(s,a)^{\top}\Lambda_h^{-1}\theta_{j,h} \\
&\leq\lambda_0\|\theta_{j,h}\|_{\Lambda_h^{-1}}\|\phi(s,a)\|_{\Lambda_h^{-1}} \\
&\leq\sqrt{d\lambda_0}\|\phi(s,a)\|_{\Lambda_h^{-1}} \\
&\leq\sqrt{d\lambda_0}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}},
\end{aligned}$$

where the last inequality is from

$$\|\theta_{j,h}\|_{\Lambda_h^{-1}}=\sqrt{\theta_{j,h}^{\top}\Lambda_h^{-1}\theta_{j,h}}\leq\|\Lambda_h^{-1}\|^{1/2}\|\theta_{j,h}\|\leq\sqrt{d/\lambda_0},$$

by using the fact that $\|\Lambda_h^{-1}\|\leq\lambda_0^{-1}$.

Step 2: We turn to the estimation error from the transition model, i.e., $\widehat{\nu}_h^{\rho,\widehat{V}_j} - \nu_h^{\rho,\widehat{V}_j}$. We define two auxiliary functions:

$$\widehat{g}_{h,i}(\alpha; j) = \int_{\mathcal{S}} \widehat{\mu}_{h,i}(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' - \rho(\alpha - \min_{s'}[\widehat{V}_{j,h+1}]_\alpha(s')),$$

and

$$g^0_{h,i}(\alpha; j) = \int_{\mathcal{S}} \mu^0_{h,i}(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' - \rho(\alpha - \min_{s'}[\widehat{V}_{j,h+1}]_\alpha(s')).$$

Then

$$\begin{aligned}
&\left|\sum_{i\in[d]} \phi_i(s,a)(\widehat{\nu}^{\rho,\widehat{V}_j}_{h,i} - \nu^{\rho,\widehat{V}_j}_{h,i})\right| \\
&\leq \sum_{i\in[d]} \left|\phi_i(s,a)(\widehat{\nu}^{\rho,\widehat{V}_j}_{h,i} - \nu^{\rho,\widehat{V}_j}_{h,i})\right| \\
&= \sum_{i\in[d]} \phi_i(s,a) \max_{\alpha\in[\min_s \widehat{V}_{j,h+1}(s), \max_s \widehat{V}_{j,h+1}(s)]} \left|\widehat{g}_{h,i}(\alpha; j) - g^0_{h,i}(\alpha; j)\right| \\
&\leq \sum_{i\in[d]} \phi_i(s,a) \max_{\alpha\in[\min_s \widehat{V}_{j,h+1}(s), \max_s \widehat{V}_{j,h+1}(s)]} \left|\int_{\mathcal{S}} \left(\widehat{\mu}_{h,i}(s') - \mu^0_{h,i}(s')\right)[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s'\right| \\
&= \sum_{i\in[d]} \max_{\alpha\in[\min_s \widehat{V}_{j,h+1}(s), \max_s \widehat{V}_{j,h+1}(s)]} \left|\phi_i(s,a)\int_{\mathcal{S}} \left(\widehat{\mu}_{h,i}(s') - \mu^0_{h,i}(s')\right)[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s'\right|,
\end{aligned}$$

where the first inequality is due to (6), (8) as well as Lemma 7, and the last inequality is based on $\phi_i(s,a) \geq 0$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ from Assumption 1. Moreover,

$$\begin{aligned}
&\left|\int_{\mathcal{S}} \mu^0_{h,i}(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' - \int_{\mathcal{S}} \widehat{\mu}_{h,i}(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s'\right| \\
&= \left|\int_{\mathcal{S}} \mu^0_{h,i}(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' - \mathbb{1}_i^\top \Lambda_h^{-1}\left(\sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)[\widehat{V}_{j,h+1}]_\alpha(s^\tau_{h+1})\right)\right| \\
&= \left|\mathbb{1}_i^\top \Lambda_h^{-1}\left(\Lambda_h \int_{\mathcal{S}} \mu^0_h(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' - \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)[\widehat{V}_{j,h+1}]_\alpha(s^\tau_{h+1})\right)\right| \\
&= \left|\mathbb{1}_i^\top \Lambda_h^{-1}\left[\lambda_0 \int_{\mathcal{S}} \mu^0_h(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' + \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\left(\int_{\mathcal{S}} \widetilde{P}^0_h(s'|s^\tau_h, a^\tau_h)[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' - [\widehat{V}_{j,h+1}]_\alpha(s^\tau_{h+1})\right)\right]\right| \\
&= \left|\mathbb{1}_i^\top \Lambda_h^{-1}\left[\lambda_0 \int_{\mathcal{S}} \mu^0_h(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' + \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\left(\int_{\mathcal{S}} P^0_h(s'|s^\tau_h, a^\tau_h)[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' - [\widehat{V}_{j,h+1}]_\alpha(s^\tau_{h+1})\right)\right]\right| \\
&\quad + \left|\mathbb{1}_i^\top \Lambda_h^{-1}\left[\sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\left(\int_{\mathcal{S}} \left(\widetilde{P}^0_h(s'|s^\tau_h, a^\tau_h) - P^0_h(s'|s^\tau_h, a^\tau_h)\right)[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s'\right)\right]\right| \\
&= \left|\mathbb{1}_i^\top \Lambda_h^{-1}\left[\lambda_0 \int_{\mathcal{S}} \mu^0_h(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' + \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\epsilon^\tau_h(\alpha, \widehat{V}_{j,h+1})\right]\right| + \left|\psi H \mathbb{1}_i^\top \Lambda_h^{-1} \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\right|
\end{aligned}$$

where the first equality is from (17), the third one is due to (9) and we let

$$\epsilon^\tau_h(\alpha, V_j) = \int_{\mathcal{S}} P^0_h(s'|s^\tau_h, a^\tau_h)[V_j]_\alpha(s')\mathrm{d}s' - [V_j]_\alpha(s^\tau_{h+1}),$$

for any $V_j : \mathcal{S} \to [0, H]$ and $\alpha \in [\min_s V_j(s), \max_s V_j(s)]$. Then, we have

$$
\begin{aligned}
&\left|\phi_i(s,a) \cdot (\widehat{\mu}_{h,i} - \mu^0_{h,i})[\widehat{V}_{j,h+1}]_\alpha\right| \\
\leq&\left|\phi_i(s,a)\mathbb{1}_i^\top \Lambda_h^{-1}\Big(\lambda_0 \int_{\mathcal{S}} \mu^0_h(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s' + \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\epsilon^\tau_h(\alpha, \widehat{V}_{j,h+1}) + \psi H \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\Big)\right| \\
\leq&\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \underbrace{\Big(\lambda_0 \Big\| \int_{\mathcal{S}} \mu^0_h(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s'\Big\|_{\Lambda_h^{-1}}}_{\texttt{(i)}} + \Big\| \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\epsilon^\tau_h(\alpha, \widehat{V}_{j,h+1})\Big\|_{\Lambda_h^{-1}}\Big) \\
&+ \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \underbrace{\Big(\psi H \| \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\|_{\Lambda_h^{-1}}\Big)}_{\texttt{(ii)}}, \qquad (35)
\end{aligned}
$$

where the last inequality holds due to the Cauchy-Schwarz inequality. Moreover, the term `(i)` in (35) can be further simplified to

$$
\texttt{(i)} \leq \lambda_0 \|\Lambda_h^{-1}\|^{\frac{1}{2}} \| \int_{\mathcal{S}} \mu^0_h(s')[\widehat{V}_{j,h+1}]_\alpha(s')\mathrm{d}s'\| \leq \sqrt{\lambda_0} H,
$$

since $|\widehat{V}_{h+1}(s)| \leq H$ for any $s \in \mathcal{S}$ and $\|\Lambda_h^{-1}\| \leq 1/\lambda_0$. And the term `(ii)` in (35) can be further simplified to

$$
\begin{aligned}
\texttt{(ii)} &= \psi H \sqrt{\sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)^\top \Lambda_h^{-1} \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)}, \\
&= \psi H \sqrt{\mathrm{Tr}(\Lambda_h^{-1} \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h) \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)^\top)} \\
&= \psi H \sqrt{\mathrm{Tr}(\Lambda_h^{-1}(\Lambda_h - \lambda_0 I_d))} \\
&\leq \psi H \sqrt{\mathrm{Tr}(\Lambda_h^{-1}\Lambda_h)} \\
&= \sqrt{d}\psi H.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\left|\sum_{i\in[d]} \phi_i(s,a)(\widehat{\nu}^{\rho,\widehat{V}_j}_{h,i} - \nu^{\rho,\widehat{V}_j}_{h,i})\right| \\
\leq&\left(\sqrt{\lambda_0}H + \max_{\alpha\in[\min_s \widehat{V}_{j,h+1}(s), \max_s \widehat{V}_{j,h+1}(s)]} \| \sum_{\tau\in\mathcal{D}^0_h} \phi(s^\tau_h, a^\tau_h)\epsilon^\tau_h(\alpha, \widehat{V}_{j,h+1})\|_{\Lambda_h^{-1}} + \sqrt{d}\psi H\right) \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}},
\end{aligned}
$$

for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

B.3.5 Proof of (32)

Since $\epsilon^\tau_h(\alpha, V_j)$ is 2-Lipschitz with respect to α for any $V_j : \mathcal{S} \to [0, H]$, i.e.

$$
|\epsilon^\tau_h(\alpha, V_j) - \epsilon^\tau_h(\alpha^\dagger, V_j)| \leq 2|\alpha - \alpha^\dagger| \leq 2\epsilon_0.
$$

Therefore, for any $\alpha \in [0, H]$, one has

$$\begin{aligned}
&\left\| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \left(\epsilon_h^\tau(\alpha, V_j) - \epsilon_h^\tau(\alpha^\dagger, V_j) \right) \right\|_{\Lambda_h^{-1}}^2 \\
=& \sum_{\tau, \tau' \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^{\tau'}, a_h^{\tau'}) \left[\left(\epsilon_h^\tau(\alpha, V_j) - \epsilon_h^\tau(\alpha^\dagger, V_j) \right) \left(\epsilon_h^{\tau'}(\alpha, V_j) - \epsilon_h^{\tau'}(\alpha^\dagger, V_j) \right) \right] \\
\leq& \sum_{\tau, \tau' \in \mathcal{D}_h^0} \phi(s_h, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^{\tau'}, a_h^{\tau'}) \cdot 4\epsilon_0^2 \\
\leq& 4\epsilon_0^2 N_h^2 / \lambda_0,
\end{aligned}$$

where the last inequality is based on $\|\phi(s, a)\| \leq 1$ and $\lambda_{\min}(\Lambda_h) \geq \lambda_0$ for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ such that

$$\sum_{\tau, \tau' \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^{\tau'}, a_h^{\tau'}) = \sum_{\tau, \tau' \in \mathcal{D}_h^0} \|\phi(s_h^\tau, a_h^\tau)\|_2 \cdot \|\phi(s_h^{\tau'}, a_h^{\tau'})\|_2 \cdot \|\Lambda_h^{-1}\| \leq N_h^2 / \lambda_0.$$

Thus,

$$\max_{\alpha \in [0, H]} \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \left(\epsilon_h^\tau(\alpha, \widehat{V}_{j,h+1}) - \epsilon_h^\tau(\alpha^\dagger, \widehat{V}_{j,h+1}) \right) \|_{\Lambda_h^{-1}}^2 \leq 4\epsilon_0^2 N_h^2 / \lambda_0 \leq 4\epsilon_0^2 K^2 / \lambda_0,$$

due to the fact $N_h \leq K$ for any $h \in [H]$, which completes the proof of (32).

B.3.6 Proof of Lemma 11

For any fixed $h \in [H]$ and $\tau \in \mathcal{D}_h^0$, we define the σ-algebra

$$\mathcal{F}_{h,\tau} = \sigma(\{(s_h^k, a_h^k)\}_{k=1}^{(\tau+1) \wedge |N_h|}, \{j_h^k, s_{h+1}^k\}_{k=1}^{\tau}), \quad j \in \{r, g\}.$$

As shown in Jin et al. [33, Lemma B.2], for any $\tau \in \mathcal{D}_h^0$, we have $\phi(s_h^\tau, a_h^\tau)$ is $\mathcal{F}_{h,\tau-1}$-measurable and $\epsilon_h^\tau(\alpha, V_j)$ is $\mathcal{F}_{h,\tau-1}$-measurable. Hence $\{\epsilon_h^\tau(\alpha, V_j)\}_{\tau \in \mathcal{D}_h^0}$ is stochastic process adapted to the filtration $\{\mathcal{F}_{h,\tau}\}_{\tau \in \mathcal{D}_h^0}$. Then, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_h^0}[\epsilon_h^\tau(\alpha, V_j) | \mathcal{F}] &= \int_{\mathcal{S}} P_h^0(s' | s_h^\tau, a_h^\tau)[V_j]_\alpha - \mathbb{E}_{\mathcal{D}_h^0} \left[[V_j(s_{h+1}^\tau)]_\alpha | \{(s_h^k, a_h^k)\}_{k=1}^{(\tau) \wedge N_h}, \{j_h^k, s_{h+1}^k\}_{k=1}^{\tau-1} \right] \\
&= \int_{\mathcal{S}} P_h^0(s' | s_h^\tau, a_h^\tau)[V_j]_\alpha - \mathbb{E}_{\mathcal{D}_h^0} \left[[V_j(s_{h+1}^\tau)]_\alpha \right] = 0.
\end{aligned}$$

Note that $\epsilon_h^\tau(\alpha, V_j) = \int_{\mathcal{S}} P_h^0(s' | s_h^\tau, a_h^\tau)[V_j]_\alpha - [V_j(s_{h+1}^\tau)]_\alpha$ for any $V_j \in [0, H]^{\mathcal{S}}$ and $\alpha \in [0, H]$. Then, we have

$$|\epsilon_h^\tau(\alpha, V_j)| \leq H.$$

Hence, for the fixed $h \in [H]$ and all $\tau \in [H]$, the random variable $\epsilon_h^\tau(\alpha, V_j)$ is mean-zero and H-sub-Gaussian conditioning on $\mathcal{F}_{h,\tau-1}$. Then, we invoke the Lemma 3 with $\eta_\tau = \epsilon_h^\tau(\alpha, V_j)$ and $x_\tau = \phi(s_h^\tau, a_h^\tau)$. For any $\delta > 0$, we have

$$P_{\mathcal{D}} \left(\| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, V_j) \|_{\Lambda_h^{-1}}^2 > 2H^2 \log(\frac{\det(\Lambda_h^{1/2})}{\delta \det(\lambda_0 I_d)^{1/2}}) \right) \leq \delta.$$

Together with the facts that $\det(\Lambda_h^{1/2}) = (\lambda_0 + N_h)^{d/2}$ and $\det(\lambda_0 I_d)^{1/2} = \lambda_0^{d/2}$, we can conclude the proof of Lemma 11.

B.4 Proof of Corollary 1

Before continuing, we introduce some additional notations that will be used in the following analysis. For any $(h, i) \in [H] \times [d]$, define $\Phi_{h,i}^\star : \mathcal{S} \to \mathbb{R}^{d \times d}$ and $b_{h,i}^\star : \mathcal{S} \to \mathbb{R}$ by

$$\begin{aligned}
\Phi_{h,i}^\star(s) &= (\phi_i(s, \pi_h^\star(s)) \mathbb{1}_i)(\phi_i(s, \pi_h^\star(s)) \mathbb{1}_i)^\top \in \mathbb{R}^{d \times d}, \\
b_{h,i}^\star(s) &= (\phi_i(s, \pi_h^\star(s)) \mathbb{1}_i)^\top \Lambda_h^{-1} (\phi_i(s, \pi_h^\star(s)) \mathbb{1}_i).
\end{aligned}$$

With these notations in hand and recalling (11) in Theorem 1, one has

$$\begin{aligned} V_{r,1}^{\star,\rho}(\zeta) - V_{r,1}^{\widehat{\pi},\rho}(\zeta) &\leq 2\gamma_0 \sum_{h=1}^{H}\sum_{i=1}^{d} \sup_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{s\sim d_h^\star}\sqrt{b_{h,i}^\star(s)} \\ &\leq 2\gamma_0 \sum_{h=1}^{H}\sum_{i=1}^{d} \sup_{d_h^\star \in \mathcal{D}_h^\star} \sqrt{\mathbb{E}_{s\sim d_h^\star} b_{h,i}^\star(s)} \\ &= 2\gamma_0 \sum_{h=1}^{H} \sup_{d_h^\star \in \mathcal{D}_h^\star} \sum_{i=1}^{d} \sqrt{\mathbb{E}_{s\sim d_h^\star} b_{h,i}^\star(s)}, \end{aligned} \tag{36}$$

where the second inequality is due to the Jensen's inequality and concavity.

In the following, we will control the key term $\sum_{i=1}^{d}\sqrt{\mathbb{E}_{s\sim d_h^\star} b_{h,i}^\star(s)}$ for any $d_h^\star \in \mathcal{D}_h^\star$. Before continuing, we first denote

$$\mathcal{C}_h^{\mathrm{b}} = \{(s,a) : d_h^{\mathrm{b}}(s,a) > 0\}.$$

Considering any (s,a) s.t. $d_h^{\mathrm{b}}(s,a) > 0$ and from Lemma 8, the following lower bound holds with probability at least $1-3\delta$, i.e.,

$$N_h(s,a) \geq \frac{K d_h^{\mathrm{b}}(s,a)}{8} - 5\sqrt{K d_h^{\mathrm{b}}(s,a)\log(\frac{KH}{\delta})} \geq \frac{K d_h^{\mathrm{b}}(s,a)}{16}, \tag{37}$$

as long as

$$K \geq c_0 \frac{\log(KH/\delta)}{d_{\min}^{\mathrm{b}}} \geq c_0 \frac{\log(KH/\delta)}{d_h^{\mathrm{b}}(s,a)} \tag{38}$$

for some sufficiently large c_0 and $d_{\min}^{\mathrm{b}} = \min_{h,s,a}\{d_h^{\mathrm{b}}(s,a) : d_h^{\mathrm{b}}(s,a) > 0\}$. Therefore,

$$\begin{aligned} \Lambda_h &= \sum_{(s,a)\in\mathcal{C}_h^{\mathrm{b}}} N_h(s,a)\phi(s,a)\phi(s,a)^\top + I_d \\ &\succeq \sum_{(s,a)\in\mathcal{C}_h^{\mathrm{b}}} \frac{K d_h^{\mathrm{b}}(s,a)}{16}\phi(s,a)\phi(s,a)^\top + I_d \\ &\succeq \frac{K}{16}\mathbb{E}_{d_h^{\mathrm{b}}}[\phi(s,a)\phi(s,a)^\top] + I_d. \end{aligned}$$

From Assumption 3,

$$\mathbb{E}_{d_h^{\mathrm{b}}}[\phi(s,a)\phi(s,a)^\top] \succeq \max_{P\in\mathcal{P}^\rho(P^0)} \frac{d\cdot\min\{\mathbb{E}_{d_h^{\star,P}}\phi_i^2(s,a), 1/d\}}{C_{\mathrm{rob}}^\star}\mathbb{1}_{i,i}, \quad \forall i\in[d]$$

Thus, for any $i\in[d]$,

$$\Lambda_h \succeq I_d + \frac{Kd\cdot\min\{\mathbb{E}_{d_h^\star}\phi_i^2(s,\pi_h^\star(s)), 1/d\}}{16C_{\mathrm{rob}}^\star}\cdot\mathbb{1}_{i,i}.$$

Here, $\mathbb{1}_{i,j}$ represents a matrix with the (i,j)-th coordinate as 1 and all other elements as 0. Consequently,

$$\begin{aligned} \mathbb{E}_{s\sim d_h^\star} b_{h,i}^\star(s) = \mathbb{E}_{s\sim d_h^\star}\mathrm{Tr}(\Phi_{h,i}^\star(s)\Lambda_h^{-1}) &= \mathrm{Tr}(\mathbb{E}_{s\sim d_h^\star}\Phi_{h,i}^\star(s)\Lambda_h^{-1}) \\ &\leq \frac{\mathbb{E}_{d_h^\star}\phi_i^2(s,\pi_h^\star(s))}{1+Kd\cdot\min\{\mathbb{E}_{d_h^\star}\phi_i^2(s,\pi_h^\star(s)), 1/d\}/16C_{\mathrm{rob}}^\star}, \end{aligned} \tag{39}$$

where the second equality is because the trace is a linear mapping and the last inequality holds by Lemma 5. We further define $\mathcal{E}_{h,\mathsf{larger}} = \{i : \mathbb{E}_{(s,a)\sim d_h^\star}\phi_i^2(s,a) \geq \frac{1}{d}\}$. Due to Assumption 1, we first claim that

$$|\mathcal{E}_{h,\mathsf{larger}}| \leq \sqrt{d}, \tag{40}$$

where the proof can be found at the end of this subsection.

By utilizing Assumption 3, we discuss the following three cases.

- If $\mathbb{E}_{(s,a)\sim d_h^\star}\phi_i^2(s,a)=0$ ($i\notin\mathcal{E}_{h,\mathsf{larger}}$), it is easily observed that (39) can be controlled by $\langle d_h^\star, b_{h,i}^\star\rangle\le 0$.

- If $0<\mathbb{E}_{(s,a)\sim d_h^\star}\phi_i^2(s,a)\le\frac{1}{d}$ ($i\notin\mathcal{E}_{h,\mathsf{larger}}$),we have

$$(39)\le\frac{16C_{\mathrm{rob}}^\star\cdot\mathbb{E}_{d_h^\star}\phi_i^2(s,\pi_h^\star(s))}{Kd\cdot\mathbb{E}_{d_h^\star}\phi_i^2(s,\pi_h^\star(s))}=\frac{16C_{\mathrm{rob}}^\star}{Kd}.$$

- If $i\in\mathcal{E}_{h,\mathsf{larger}}$, i.e., $\frac{1}{d}\le\mathbb{E}_{(s,a)\sim d_h^\star}\phi_i^2(s,a)\le 1$, we have

$$(39)\le\frac{16C_{\mathrm{rob}}^\star\cdot\mathbb{E}_{d_h^\star}\phi_i^2(s,\pi_h^\star(s))}{K}\le\frac{16C_{\mathrm{rob}}^\star}{K},$$

  where the last inequality holds due to $\phi_i^2(s,\pi_h^\star(s))\le 1$.

Summing up the above three cases and (40), we have

$$\begin{aligned}\sum_{i=1}^d\sqrt{\mathbb{E}_{s\sim d_h^\star}b_{h,i}^\star(s)}&\le\sum_{i\in\mathcal{E}_{h,\mathsf{larger}}}\sqrt{\mathbb{E}_{s\sim d_h^\star}b_{h,i}^\star(s)}+\sum_{i\notin\mathcal{E}_{h,\mathsf{larger}}}\sqrt{\mathbb{E}_{s\sim d_h^\star}b_{h,i}^\star(s)}\\&\le|\mathcal{E}_{h,\mathsf{larger}}|\sqrt{\frac{16C_{\mathrm{rob}}^\star}{K}}+|d-\mathcal{E}_{h,\mathsf{larger}}|\sqrt{\frac{16C_{\mathrm{rob}}^\star}{Kd}}\\&\le 8\sqrt{C_{\mathrm{rob}}^\star}\sqrt{\frac{d}{K}}.\end{aligned}$$

Together with (36) and setting $\gamma_0=6\sqrt{d}H\sqrt{\log(3HK/\delta)}$, one obtains

$$\begin{aligned}V_{r,1}^{\star,\rho}(\zeta)-V_{r,1}^{\widehat{\pi},\rho}(\zeta)&\le 2\gamma_0\sum_{h=1}^H\sup_{d_h^\star\in\mathcal{D}_h^\star}\sum_{i=1}^d\sqrt{\mathbb{E}_{s\sim d_h^\star}b_{h,i}^\star(s)}\\&\le 96dH^2\sqrt{C_{\mathrm{rob}}^\star/K}\sqrt{\log(3HK/\delta)},\end{aligned}$$

with probability at least $1-4\delta$, as long as $K\ge c_0\frac{\log(KH/\delta)}{d_{\min}^{\mathrm{b}}}$ for some universal constant c_0.

Proof of (40)**.** Let $\tilde{\mathcal{E}}_{h,\mathsf{larger}}=\{i:\mathbb{E}_{(s,a)\sim d_h^\star}\phi_i(s,a)\ge\frac{1}{\sqrt{d}}\}$.

- We first show that $|\tilde{\mathcal{E}}_{h,\mathsf{larger}}|$ should be no larger than $\sqrt{d}$ by contradiction. Suppose $|\tilde{\mathcal{E}}_{h,\mathsf{larger}}|>\sqrt{d}$. Then, there are more than $\sqrt{d}$ coordinates of $\mathbb{E}_{(s,a)\sim d_h^\star}\phi(s,a)\in\mathbb{R}^d$ that is larger than $1/\sqrt{d}$. In other words,

$$\sum_{i\in\tilde{\mathcal{E}}_{h,\mathsf{larger}}}\mathbb{E}_{(s,a)\sim d_h^\star}\phi_i(s,a)>1,$$

  which is equivalent to

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\|_1\ge\mathbb{E}_{(s,a)\sim d_h^\star}\|\phi(s,a)\|_1\ge\mathbb{E}_{(s,a)\sim d_h^\star}\sum_{i\in\tilde{\mathcal{E}}_{h,\mathsf{larger}}}\phi_i(s,a)>1,$$

  where the last inequality is from the linearity of the expectation mapping. It contradicts to our Assumption 2, which implies $\|\phi(s,a)\|_1=1$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}\times[H]$.

- Then, we show that $\tilde{\mathcal{E}}_{h,\mathsf{larger}}\subseteq\mathcal{E}_{h,\mathsf{larger}}$: For every element $i\in\tilde{\mathcal{E}}_{h,\mathsf{larger}}$, we have

$$\frac{1}{d}\le(\mathbb{E}_{(s,a)\sim d_h^\star}\phi_i(s,a))^2\le\mathbb{E}_{(s,a)\sim d_h^\star}\phi_i^2(s,a),$$

  where the second inequality is due to the Jensen's inequality. Thus, $\tilde{\mathcal{E}}_{h,\mathsf{larger}}\subseteq\mathcal{E}_{h,\mathsf{larger}}$.

Combining these two arguments, we show that $|\mathcal{E}_{h,\mathsf{larger}}|\ge\sqrt{d}$.

B.5 Proof of Corollary 2

We first establish the following lemma to control the sub-optimality, under the full feature coverage.

Lemma 12. *Consider $\delta \in (0,1)$. Suppose Assumption 2, Assumption 4 and all conditions in Lemma 6 hold. For any $h \in [H]$, if $N_h \geq \max\{512\log(2Hd/\delta)/\kappa^2, 4/\kappa\}$, we have*

$$\sum_{i=1}^{d} \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \leq \frac{2}{\sqrt{N_h\kappa}}, \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A},$$

with probability exceeding $1-\delta$.

Proof. From Lemma 6 and Assumption 4, one has

$$\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \leq \frac{2\phi_i(s,a)}{\sqrt{N_h\kappa}}, \quad \forall (i,s,a) \in [d]\times\mathcal{S}\times\mathcal{A},$$

as long as $N_h \geq \max\{512\log(2Hd/\delta)/\kappa^2, 4/\kappa\}$. In addition,

$$1 = \int_{\mathcal{S}} P_h^0(s'|s,a)\mathrm{d}s' = \int_{\mathcal{S}} \phi(s,a)^\top \mu_h^0(s')\mathrm{d}s' = \sum_{i=1}^{d}\phi_i(s,a)\int_{\mathcal{S}}\mu_{h,i}^0(s')\mathrm{d}s' = \sum_{i=1}^{d}\phi_i(s,a),$$

where the last equality is implied by Assumption 2. Therefore,

$$\sum_{i=1}^{d} \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \leq \sum_{i=1}^{d}\frac{2\phi_i(s,a)}{\sqrt{N_h\kappa}} \leq \frac{2}{\sqrt{N_h\kappa}}.$$

□

From (37), we have $N_h \geq \frac{K}{16}$ with probability exceeding $1-3\delta$, as long as K obeys (38). Together will Lemma 12, with probability exceeding $1-4\delta$, one has

$$\sum_{i=1}^{d} \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \leq \frac{8}{\sqrt{K\kappa}}, \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H],$$

as long as $K \geq \max\{c_0\log(2Hd/\delta)/\kappa^2, c_0\log(KH/\delta)/d_{\min}^{\mathrm{b}}\}$ for some sufficiently large universal constant c_0. It follows Theorem 1 that

$$\text{Sub-Opt}(\widehat{\pi};\zeta,\mathcal{P}^\rho) \leq 96\sqrt{d}H^2\sqrt{\frac{\log(3HK/\delta)}{K\kappa}},$$

which completes the proof.

□