Sample Complexity of Offline Distributionally Robust Linear Markov Decision Processes

He Wang
hew2@andrew.cmu.edu
Department of Electrical and Computer
Engineering, Carnegie Mellon University

Laixi Shi
laixis@caltech.edu
Department of Computing and Mathematical
Sciences, California Institute of Technology

Yuejie Chi
yuejiec@andrew.cmu.edu
Department of Electrical and Computer
Engineering, Carnegie Mellon University

Abstract

In offline reinforcement learning (RL), the absence of active exploration calls for attention on the model robustness to tackle the sim-to-real gap, where the discrepancy between the simulated and deployed environments can significantly undermine the performance of the learned policy. To endow the learned policy with robustness in a sample-efficient manner in the presence of high-dimensional state-action space, this paper considers the sample complexity of distributionally robust linear Markov decision processes (MDPs) with an uncertainty set characterized by the total variation distance using offline data. We develop a pessimistic model-based algorithm and establish its sample complexity bound under minimal data coverage assumptions, which outperforms prior art by at least $\widetilde{O}(d)$, where d is the feature dimension. We further improve the performance guarantee of the proposed algorithm by incorporating a carefully-designed variance estimator.

1 Introduction

In reinforcement learning (RL), agents aim to learn an optimal policy that maximizes the expected total rewards, by actively interacting with an *unknown* environment. However, online data collection may be prohibitively expensive or potentially risky in many real-world applications, e.g., autonomous driving (Gu et al., 2022), healthcare (Yu et al., 2021), and wireless security (Uprety and Rawat, 2020). This motivates the study of *offline RL*, which leverages existing historical data (aka batch data) collected in the past to improve policy learning, and has attracted growing attention (Levine et al., 2020). Nonetheless, the performance of the learned policy invoking standard offline RL techniques could drop dramatically when the deployed environment shifts from the one experienced by the historical data even slightly, necessitating the development of robust RL algorithms that are resilient against environmental uncertainty.

In response, recent years have witnessed a surge of interests in *distributionally robust offline RL* (Zhou et al., 2021b; Yang et al., 2022; Shi and Chi, 2022; Blanchet et al., 2024). In particular, given only historical data from a nominal environment, *distributionally robust offline RL* aims to learn a policy that optimizes the *worst-case performance* when the environment falls into some prescribed uncertainty set around the nominal one. Such a framework ensures that the performance of the learned policy does not fail drastically, provided that the *distribution shift* between the nominal and deployment environments is not excessively large.

Nevertheless, most existing provable algorithms in distributionally robust offline RL only focus on the tabular setting with finite state and action spaces (Zhou et al., 2021b; Yang et al., 2022; Shi

and Chi, 2022), where the sample complexity scales linearly with the size of the state-action space, which is prohibitive when the problem is high-dimensional. To expand the reach of distributionally robust offline RL, a few latest works (Ma et al., 2022; Blanchet et al., 2024) attempt to develop sample-efficient solutions by leveraging linear function approximation (Bertsekas, 2012), which is widely used in both theoretic (Jin et al., 2020; 2021; Xiong et al., 2023) and practical (Prashanth and Bhatnagar, 2010; Bellemare et al., 2019) developments of standard RL. However, the existing sample complexity is still far from satisfactory and notably larger than the counterpart of standard offline RL (Jin et al., 2021; Xiong et al., 2023) with linear function approximation, especially in terms of the dependency on the dimension of the feature space d. Therefore, it is natural to ask:

> *Can we design a provably sample-efficient algorithm for distributionally robust offline RL with linear representations?*

1.1 Main contribution

To answer this question, we focus on learning a robust variant of linear Markov decision processes in the offline setting. Throughout this paper, we consider a class of finite-horizon distributionally robust linear MDPs (Lin-RMDPs), where the uncertainty set is characterized by the total variation (TV) distance between the feature representations in the latent space, motivated by its practical (Pan et al., 2024) and theoretical appeals (Shi et al., 2023). The highlights of our contributions can be summarized as follows.

- We propose a distributionally robust variant of pessimistic least-squares value iteration, referred to as `DROP`, which incorporates linear representations of the MDP model and devises a data-driven penalty function to account for data scarcity in the offline setting. We also establish the sub-optimality bound for `DROP` under the minimal offline data assumption (cf. Theorem 1).

- We introduce a clipped single-policy concentrability coefficient $C_{\text{rob}}^{\star} \geq 1$ tailored to characterize the partial feature coverage of the offline data in Lin-RMDPs, and demonstrate that `DROP` attains ϵ-accuracy (for learning the robust optimal policy) as soon as the sample complexity is above $\widetilde{O}(C_{\text{rob}}^{\star} d^2 H^4/\epsilon^2)$ (cf. Corollary 1). Compared with the prior art (Blanchet et al., 2024), `DROP` improves the sample complexity by at least $\widetilde{O}(d)$.

- We further develop a variance-weighted variant of `DROP` by integrating a carefully designed variance estimator, dubbed by `DROP-V`. Due to tighter control of the variance, `DROP-V` exhibits an improved sub-optimality gap under the full feature coverage assumption (see Section 4).

See Table 1 for a summary of our results in terms of the sub-optimality gap.

1.2 Related works

In this section, we mainly discuss works that study sample complexity of linear MDPs and robust RL, which are closely related to this paper.

Finite-sample guarantees for linear MDPs. Considering linear function approximation in RL, a significant body of works study linear MDPs with linear transitions and rewards. Focusing on offline settings, Jin et al. (2021) proposed Pessimistic Value Iteration (PEVI) for offline RL with finite-horizon linear MDPs, which incorporated linear function approximation together with the principle of pessimism (Rashidinejad et al., 2021; Shi et al., 2022; Yan et al., 2023; Woo et al., 2024; Li et al., 2024). However, the temporally statistical dependency arising from backward updates leads to an $O(\sqrt{d})$ amplification in the sub-optimality gap with comparison to the lower bound (Zanette et al., 2021). Subsequently, Min et al. (2021) and Yin et al. (2022) attempted to address this gap. The near-optimal sample complexity in Xiong et al. (2023) is achieved by applying variance estimation

Algorithm	Coverage	Sub-optimality gap
Blanchet et al. (2024)	partial	$\sqrt{\dfrac{C_1^\star d^4 H^4}{K}}$
DROP (this work)	partial	$\sqrt{\dfrac{C_{\mathrm{rob}}^\star d^2 H^4}{K}}$
DROP (this work)	arbitrary	$\sqrt{d}\, H \sum_{i=1}^{d} \sum_{h=1}^{H} \max_{P \in \mathcal{P}^\rho(P^0)} \mathbb{E}_{\pi^\star, P} \|\phi_i(s,a)\mathbb{1}_i\|_{(\Lambda_h)^{-1}}$
DROP-V (this work)	full	$\sqrt{d} \sum_{i=1}^{d} \sum_{h=1}^{H} \max_{P \in \mathcal{P}^\rho(P^0)} \mathbb{E}_{\pi^\star, P} \|\phi_i(s,a)\mathbb{1}_i\|_{(\Sigma_h^\star)^{-1}}$

Table 1: Our results and comparisons with prior art (up to log factors) in terms of sub-optimality gap, for learning robust linear MDPs with an uncertainty set measured by the TV distance. Here, d is the feature dimension, H is the horizon length, K is the number of trajectories, $C_1^\star, C_{\mathrm{rob}}^\star \in [1, \infty)$ are some concentrability coefficients (defined in Section 3.3.1) satisfying $C_{\mathrm{rob}}^\star \leq dC_1^\star$. In addition, $\mathcal{P}^\rho(P^0)$ denotes the uncertainty set around the nominal kernel P^0, $\phi_i(s,a)$ is the i-th coordinate of the feature vector given any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, Λ_h and $\Sigma_h^\star$ are some sort of cumulative sample covariance matrix and variance-weighted cumulative sample covariance matrix, satisfying $H\Lambda_h^{-1} \succeq (\Sigma_h^\star)^{-1}$.

techniques, motivating us to explore variance information for robust offline RL. Beyond these works, there are many offline RL works that investigate model-free algorithms or the infinite-horizon setting that deviates from our focus (Zanette et al., 2021; Uehara and Sun, 2021; Xie et al., 2021). In addition to the offline setting, linear MDPs are also widely explored in other settings such as generative model or online RL (Yang and Wang, 2020; Jin et al., 2020; Agarwal et al., 2023; He et al., 2023).

Distributionally robust RL with tabular MDPs. To promote robustness in the face of environmental uncertainty, a line of research known as distributionally robust RL incorporates distributionally robust optimization tools with RL to ensure robustness in a worst-case manner (Iyengar, 2005; Xu and Mannor, 2012; Wolff et al., 2012; Kaufman and Schaefer, 2013; Tamar et al., 2014; Ho et al., 2018; Smirnova et al., 2019; Derman and Mannor, 2020; Ho et al., 2021; Badrinath and Kalathil, 2021; Goyal and Grand-Clement, 2022; Ding et al., 2024). Recently, a surge of works focus on understanding/improving sample complexity and computation complexity of RL algorithms in the tabular setting (Wang and Zou, 2021; Zhou et al., 2021b; Dong et al., 2022; Yang et al., 2022; Panaganti and Kalathil, 2022; Liu et al., 2022; Li et al., 2022; Xu et al., 2023; Wang et al., 2023a;c; Liang et al., 2023; Wang et al., 2023b; Li and Lan, 2023; Yang et al., 2023; Zhang et al., 2023; Kumar et al., 2024; Blanchet et al., 2024; Shi et al., 2024). Among them, Zhou et al. (2021b); Yang et al. (2022); Shi and Chi (2022); Blanchet et al. (2024) consider distributionally robust RL in the offline setting, which are most related to us. Different from the sample complexity achieved in the tabular setting that largely depends on the size of state and action spaces, this work advances beyond the tabular setting and develops a sample complexity guarantee that only depends on the feature dimension based on the linear MDP model assumption.

Distributionally robust RL with linear MDPs. The prior art (Blanchet et al., 2024) studies the same robust linear MDP (Lin-RMDP) problem as this work, while the sample complexity dependency on the feature dimension d is still worse than that of standard linear MDPs in offline setting (Jin et al., 2021; Xiong et al., 2023), highlighting gap for potential improvements. Besides TV distance considered herein, Ma et al. (2022); Blanchet et al. (2024) consider the Kullback-Leibler (KL) divergence for the uncertainty set. Moreover, Liu and Xu (2024a) explores Lin-RMDPs with an additional structure assumption in the online setting, which diverges from our focus on the offline context. We also note that a concurrent work (Liu and Xu, 2024b) studies offline Lin-RMDPs, which aligns with our interest. However, their focus is limited to the well-explored data coverage

assumption. In addition to linear representations for robust MDPs, Badrinath and Kalathil (2021); Ramesh et al. (2024) consider other classes of function for model approximation.

Notation. Throughout this paper, we define $\Delta(\mathcal{S})$ as the probability simplex over a set $\mathcal{S}$, $[H] := \{1, \ldots, H\}$ and $[d] := \{1, \ldots, d\}$ for some positive integer $H, d > 0$. We also denote $\mathbb{1}_i$ as the vector with appropriate dimension, where the i-th coordinate is 1 and others are 0. We use I_d to represent the identity matrix of the size d. For any vector $x \in \mathbb{R}^d$, we use $\|x\|_2$ and $\|x\|_1$ to represent its l_2 and l_1 norm, respectively. In addition, we denote $\sqrt{x^\top A x}$ as $\|x\|_A$ given any vector $x \in \mathbb{R}^d$ and any semi-definite matrix $A \in \mathbb{R}^{d \times d}$. For any set $\mathcal{D}$, we use $|\mathcal{D}|$ to represent the cardinality (i.e., the size) of the set $\mathcal{D}$. Additionally, we use $\min\{a, b\}_+$ to denote the minimum of a and b when $a, b > 0$, and 0 otherwise. We also let $\lambda_{\min}(A)$ to denote the smallest eigenvalue of any matrix A.

2 Problem Setup

In this section, we introduce the formulation of finite-horizon distributionally robust linear MDPs (Lin-RMDPs), together with the batch data assumptions and the learning objective.

Standard linear MDPs. Consider a finite-horizon standard linear MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P = \{P_h\}_{h=1}^H, r = \{r_h\}_{h=1}^H)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state space and action space respectively, and H is the horizon length. At each step $h \in [H]$, we denote $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ as the transition kernel and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ as the deterministic reward function, which satisfy the following assumption used in Yang and Wang (2019); Jin et al. (2020).

Assumption 1 (Linear MDPs). *A finite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r)$ is called a linear MDP 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 an unknown vector $\theta_h \in \mathbb{R}^d$ at each step h such that*

$$r_h(s, a) = \phi(s, a)^\top \theta_h, \quad P_h(s' \mid s, a) = \phi(s, a)^\top \mu_h^P(s'), \qquad \forall (h, s, a, s') \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}. \quad (1)$$

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

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 time step h. Given the policy π and the transition kernel P, the value function $V_h^{\pi, P}$ and the Q-function $Q_h^{\pi, P}$ at the h-th step are defined as: for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$V_h^{\pi, P}(s) = \mathbb{E}_{\pi, P}\left[\sum_{t=h}^H r_t(s_t, a_t) \mid s_h = s \right], \quad Q_h^{\pi, P}(s, a) = r_h(s, a) + \mathbb{E}_{\pi, P}\left[\sum_{t=h+1}^H r_t(s_t, a_t) \mid s_h = s, a_h = a \right],$$

where the expectation is taken over the randomness of the trajectory induced by the policy π and the transition kernel P.

Lin-RMDPs: distributionally robust linear MDPs. In this work, we consider distributionally robust linear MDPs (Lin-RMDPs), where the transition kernel can be an arbitrary one within an *uncertainty set* around the nominal kernel — an ensemble of probability transition kernels, rather than a fixed transition in standard linear MDPs (Jin et al., 2021; Yin et al., 2022; Xiong et al., 2023). Formally, we denote it by $\mathcal{M}_{\text{rob}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}^\rho(P^0), r)$, where P^0 represents a nominal transition kernel and then $\mathcal{P}^\rho(P^0)$ represents the *uncertainty set* (a ball) around the nominal P^0 with some *uncertainty level* $\rho \geq 0$. For notational simplicity, we denote $\mu_h^0 := \mu_h^{P^0}$ and according to (1), we let

$$\forall (h, s, a, s') \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} : \quad P_{h,s,a}^0(s') := P_h^0(s' \mid s, a) = \phi(s, a)^\top \mu_h^0(s').$$

In this work, we consider the total variation (TV) distance (Tsybakov, 2008) as the divergence metric for the uncertainty set $\mathcal{P}^\rho$ and we assume that Lin-RMDPs satisfy the following d-rectangular assumption (Ma et al., 2022; Blanchet et al., 2024).

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

$$\mathcal{P}^\rho(P^0) := \otimes_{[H],\mathcal{S},\mathcal{A}} \mathcal{P}^\rho(P^0_{h,s,a}), \quad with \ \mathcal{P}^\rho(P^0_{h,s,a}) := \left\{ \phi(s,a)^\top \mu_h(\cdot) : \mu_h \in \mathcal{U}^\rho(\mu^0_h) \right\},$$

$$where \quad \mathcal{U}^\rho(\mu^0_h) := \otimes_{[d]} \mathcal{U}^\rho(\mu^0_{h,i}), \quad with \ \mathcal{U}^\rho(\mu^0_{h,i}) := \left\{ \mu_{h,i} : D_{\mathrm{TV}}\left(\mu_{h,i} \parallel \mu^0_{h,i}\right) \le \rho \ and \ \mu_{h,i} \in \Delta(\mathcal{S}) \right\}.$$

Here, $D_{\mathrm{TV}}(\mu_{h,i} \parallel \mu^0_{h,i})$ represents the TV-distance between two measures over the state space, i.e.

$$D_{\mathrm{TV}}(\mu_{h,i}, \mu^0_{h,i}) = \frac{1}{2}\|\mu_{h,i} - \mu^0_{h,i}\|_1,$$

where $\otimes_{[d]}$ (resp. $\otimes_{[H],\mathcal{S},\mathcal{A}}$) denotes Cartesian products over $[d]$ (resp. $[H]$, $\mathcal{S}$, and $\mathcal{A}$).

Assumption 2 indicates that the uncertainty set can be decoupled into each feature dimension $i \in [d]$ independently, so called d-rectangularity. Note that by letting $d = SA$ and $\phi_i(s,a) = \mathbb{1}_i$ for any $(i,s,a) \in [d] \times \mathcal{S} \times \mathcal{A}$, the Lin-RMDP is instantiated to the tabular RMDP and d rectangularity becomes the (s,a)-rectangularity commonly used in prior literatures (Yang et al., 2022; Shi et al., 2023). Moreover, when the uncertainty level $\rho = 0$, Lin-RMDPs reduce to a subclass of standard linear MDPs satisfying Assumption 1.

Robust value function and robust Bellman equations. Considering Lin-RMDPs with any given policy π, we define *robust value function* (resp. *robust Q-function*) to characterize the *worst-case* performance induced by all possible transition kernels over the uncertainty set, denoted as

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad V^{\pi,\rho}_h(s) := \inf_{P \in \mathcal{P}^\rho(P^0)} V^{\pi,P}_h(s), \qquad Q^{\pi,\rho}_h(s,a) = \inf_{P \in \mathcal{P}^\rho(P^0)} Q^{\pi,P}_h(s,a).$$

They satisfy the following *robust Bellman consistency equations*:

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad Q^{\pi,\rho}_h(s,a) = \mathbb{B}^\rho_h V^{\pi,\rho}_{h+1}(s,a), \quad where \ V^{\pi,\rho}_h(s) = \mathbb{E}_{a \sim \pi_h(\cdot|s)}[Q^{\pi,\rho}_h(s,a)], \quad (2)$$

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

$$[\mathbb{B}^\rho_h f](s,a) := r_h(s,a) + [\mathbb{P}^\rho_h f](s,a), \tag{3}$$

$$[\mathbb{P}^\rho_h f](s,a) := \inf_{\mu_h \in \mathcal{U}^\rho(\mu^0_h)} \int_{\mathcal{S}} \phi(s,a)^\top \mu_h(s') f(s') \mathrm{d}s'. \tag{4}$$

Note that under Assumption 2, the robust Bellman operator inherits the linearity of the Bellman operator in standard linear MDPs (Jin et al., 2020, Proposition 2.3), as shown in the following lemma. The proof is postponed to Appendix A.1.

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

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 by

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad d^{\pi,P}_h(s) := d^{\pi,P}_h(s;\zeta) = \mathbb{P}(s_h = s \mid \zeta, \pi, P), \quad d^{\pi,P}_h(s,a) = d^{\pi,P}_h(s)\pi_h(a \mid s). \tag{5}$$

To continue, we denote $\pi^\star = \{\pi^\star_h\}^H_{h=1}$ as a deterministic *optimal robust policy* (Iyengar, 2005). The resulting *optimal robust value function* and *optimal robust Q-function* are defined as:

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

$$Q^{\star,\rho}_h(s,a) := Q^{\pi^\star,\rho}_h(s,a) = \max_\pi Q^{\pi,\rho}_h(s,a), \qquad \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}.$$

Accordingly, we also have the following *robust Bellman optimality equation*:

$$Q_h^{\star,\rho}(s,a) = [\mathbb{B}_h^\rho V_{h+1}^{\star,\rho}](s,a), \quad \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}. \tag{6}$$

Similar to (5), we also denote the occupancy distribution associated with the optimal robust policy $\pi^\star$ and some transition kernel P by

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad d_h^{\star,P}(s) := d_h^{\pi^\star,P}(s;\zeta), \qquad d_h^{\star,P}(s,a) := d_h^{\pi^\star,P}(s)\pi_h^\star(a \mid s). \tag{7}$$

Offline data. Suppose that we can access a batch dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{h \in [H], \tau \in [K]}$, which consists K i.i.d. trajectories that are generated by executing some (mixed) behavior policy $\pi^{\mathrm{b}} = \{\pi_h^{\mathrm{b}}\}_{h=1}^H$ over some nominal linear MDP $\mathcal{M}^0 = (\mathcal{S}, \mathcal{A}, H, P^0, r)$. Note that $\mathcal{D}$ contains KH transition-reward sample tuples in total, where each sample tuple $(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)$ represents that the agent took the action a_h^τ at the state s_h^τ, received the reward $r_h^\tau = r_h(s_h^\tau, a_h^\tau)$, and then observed the next state $s_{h+1}^\tau \sim P_h^0(\cdot \mid s_h = s_h^\tau, a_h = a_h^\tau)$. To proceed, we define

$$\mathcal{D}_h^0 = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau \in [K]},$$

which contains all samples at the h-th step in $\mathcal{D}$. For simplicity, we abuse the notation $\tau \in \mathcal{D}_h^0$ to denote $(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau) \in \mathcal{D}_h^0$. In addition, we define the induced occupancy distribution w.r.t. the behavior policy π^{b} and the nominal transition kernel P^0 at each step h by

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad d_h^{\mathrm{b}}(s) := d_h^{\pi^{\mathrm{b}},P^0}(s;\zeta) \quad \text{and} \quad d_h^{\mathrm{b}}(s,a) := d_h^{\pi^{\mathrm{b}},P^0}(s,a;\zeta), \tag{8}$$

which is conditioned on the initial distribution ζ.

Learning goal. Given the batch dataset $\mathcal{D}$, the goal of solving the Lin-RMDP $\mathcal{M}_{\mathrm{rob}}$ with a given initial state distribution ζ is to learn an ϵ-optimal robust policy $\widehat{\pi}$ such that the sub-optimality gap satisfies

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) := V_1^{\star,\rho}(\zeta) - V_1^{\widehat{\pi},\rho}(\zeta) \le \epsilon, \tag{9}$$

using as few samples as possible, where ϵ is the targeted accuracy level,

$$V_1^{\star,\rho}(\zeta) = \mathbb{E}_{s_1 \sim \zeta}[V_1^{\star,\rho}(s_1)] \quad \text{and} \quad V_1^{\widehat{\pi},\rho}(\zeta) = \mathbb{E}_{s_1 \sim \zeta}[V_1^{\widehat{\pi},\rho}(s_1)].$$

3 Algorithm and Performance Guarantees

In this section, we propose a model-based approach referred to as *Distributionally Robust Pessimistic Least-squares Value Iteration* (`DROP`), by constructing an empirical Bellman operator for Lin-RMDPs in an offline fashion. Then we analyze the sub-optimality bound of the robust policy learned from `DROP` and discuss the sample complexity under different historical data quality scenarios.

3.1 Empirical robust Bellman operator and strong duality

Recalling the robust Bellman operator in (3) gives that for any value function $V : \mathcal{S} \to [0, H]$,

$$(\mathbb{B}_h^\rho V)(s,a) = r_h(s,a) + \inf_{\mu_h \in \mathcal{U}^\rho(\mu_h^0)} \int_{\mathcal{S}} \phi(s,a)^\top \mu_h(s') V(s') \mathrm{d}s',$$

which can be equivalently rewritten as its dual form:

$$(\mathbb{B}_h^\rho V)(s,a) = \phi(s,a)^\top (\theta_h + \nu_h^{\rho,V}), \tag{10}$$

from the views of strong duality (Iyengar, 2005; Shi et al., 2023) (see the detailed proof in Appendix B.1). Here, $\nu_h^{\rho,V} = [\nu_{h,1}^{\rho,V}, \nu_{h,2}^{\rho,V} \ldots, \nu_{h,d}^{\rho,V}]^\top \in \mathbb{R}^d$ and its i-th coordinate is defined by

$$\nu_{h,i}^{\rho,V} := \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ \mathbb{E}_{s \sim \mu_{h,i}^0}[V]_\alpha(s) - \rho(\alpha - \min_{s'}[V]_\alpha(s')) \right\}, \text{ with } [V]_\alpha(s) = \begin{cases} \alpha, & \text{if } V(s) > \alpha, \\ V(s), & \text{otherwise.} \end{cases} \tag{11}$$

However, notice that we cannot directly apply the above robust Bellman operator and perform value iterations since we cannot have access to the ground-truth nominal linear MDP $\mathcal{M}^0$ (i.e., θ_h and μ_h^0). Therefore, for each time step h, we incorporate ridge regression to obtain the estimator $\widehat{\theta}_h \in \mathbb{R}^d$ and $\widehat{\nu}_h^{\rho,V} \in \mathbb{R}^d$, based on the batch dataset $\mathcal{D}_h^0$ that contains all the samples at h-th step collected in $\mathcal{D}^0$. In particular, for any value function $V : \mathcal{S} \to [0, H]$ and any time step $h \in [H]$, the estimator $\widehat{\theta}_h$ and the i-th coordinate of $\widehat{\nu}_h^{\rho,V}$ are defined by

$$\widehat{\theta}_h = \arg\min_{\theta \in \mathbb{R}^d} \sum_{\tau \in \mathcal{D}_h^0} \left(\phi(s_h^\tau, a_h^\tau)^\top \theta - r_h^\tau \right)^2 + \lambda_0 \|\theta\|_2^2 = \Lambda_h^{-1} \left(\sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) r_h^\tau \right), \tag{12}$$

$$\widehat{\nu}_{h,i}^{\rho,V} = \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ \bar{\nu}_{h,i}^V(\alpha) - \rho \left(\alpha - \min_{s'} [V]_\alpha(s') \right) \right\}, \qquad \forall i \in [d], \tag{13}$$

where $\lambda_0 > 0$ is the regularization coefficient, $\bar{\nu}_{h,i}^V(\alpha)$ is the i-th coordinate of $\bar{\nu}_h^V(\alpha)$ defined by

$$\bar{\nu}_h^V(\alpha) = \arg\min_{\nu \in \mathbb{R}^d} \sum_{\tau \in \mathcal{D}_h^0} \left(\phi(s_h^\tau, a_h^\tau)^\top \nu - [V]_\alpha(s_{h+1}^\tau) \right)^2 + \lambda_0 \|\nu\|_2^2 = \Lambda_h^{-1} \left(\sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) [V]_\alpha(s_{h+1}^\tau) \right), \tag{14}$$

and the cumulative sample covariance matrix is denoted as

$$\Lambda_h = \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top + \lambda_0 I_d. \tag{15}$$

Then following 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 $\mathbb{B}_h^\rho$, using the estimators obtained from (12) and (13): for any value function $V : \mathcal{S} \to [0, H]$,

$$(\widehat{\mathbb{B}}_h^\rho V)(s, a) = \phi(s, a)^\top (\widehat{\theta}_h + \widehat{\nu}_h^{\rho,V}), \qquad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{16}$$

3.2 `DROP`: distributionally robust pessimistic least-squares value iteration

To compute (16) for all time steps $h \in [H]$ recursively, we propose a distributionally robust pessimistic least-squares value iteration algorithm, referred to as `DROP` and summarized as Algorithm 1.

In Algorithm 1, we first construct a dataset $\mathcal{D}_0$ by subsampling from $\mathcal{D}$ by `Two-fold-subsampling` (cf. Algorithm 2), inspired by Li et al. (2024) to tackle the statistical dependency between different time steps $h \in [H]$ in the original batch dataset $\mathcal{D}$. As the space is limited, we defer the details of `Two-fold-subsampling` and the corresponding statistical independence property to Section B.2. With $\mathcal{D}_0$ in hand and initializations $\widehat{Q}_{H+1}(\cdot, \cdot) = 0$ and $\widehat{V}_{H+1}(\cdot) = 0$, the updates at time step h in `DROP` can be boiled down to the following two steps. The first one is to construct the empirical robust Bellman operator via (12)-(16) conditioned on a fixed $\widehat{V}_{h+1}$ from the previous iteration (see the line 3-8 in Algorithm 1). Then we can estimate the robust Q-function from the pessimistic value iteration as below:

$$\bar{Q}_h(s, a) = \widehat{\mathbb{B}}_h^\rho (\widehat{V}_{h+1})(s, a) - \underbrace{\gamma_0 \sum_{i=1}^d \|\phi_i(s, a) \mathbb{1}_i\|_{\Lambda_h^{-1}}}_{\text{penalty function } \Gamma_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}}, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A},$$

where $\gamma_0 > 0$ is the coefficient of the penalty term.

The above pessimistic principle is widely advocated in both standard and robust offline RL (Jin et al., 2021; Xiong et al., 2023; Shi and Chi, 2022). When dealing with the uncertainty set characterized by TV distance, previous penalty designs tailored for standard linear MDPs (Jin et al., 2021; Xiong et al., 2023) and robust linear MDPs specifically addressing KL divergence (Ma et al., 2022), are

---

Algorithm 1 Distributionally Robust Pessimistic Least-Squares Value Iteration (`DROP`)

Input: Dataset $\mathcal{D}$; feature map $\phi(s, a)$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$; parameters $\lambda_0, \gamma_0 > 0$.

1: Construct a temporally independent dataset $\mathcal{D}_0 = \text{Two-fold-subsampling}(\mathcal{D})$ (Algorithm 2).

Initialization: Set $\widehat{Q}_{H+1}(\cdot, \cdot) = 0$ and $\widehat{V}_{H+1}(\cdot) = 0$.

2: **for** step $h = H, H - 1, \cdots, 1$ **do**

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

4: $\quad$ $\widehat{\theta}_h = \Lambda_h^{-1} \left(\sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) r_h^\tau \right)$.

5: $\quad$ **for** feature $i = 1, \cdots, d$ **do**

6: $\quad\quad$ Update $\widehat{\nu}_{h,i}^{\rho, \widehat{V}}$ via (13).

7: $\quad$ **end for**

8: $\quad$ $\widehat{w}_h^{\rho, \widehat{V}} = \widehat{\theta}_h + \widehat{\nu}_h^{\rho, \widehat{V}}$.

9: $\quad$ $\bar{Q}_h(\cdot, \cdot) = \phi(\cdot, \cdot)^\top \widehat{w}_h^{\rho, \widehat{V}} - \gamma_0 \sum_{i=1}^d \left\| \phi_i(\cdot, \cdot) \mathbb{1}_i \right\|_{\Lambda_h^{-1}}$.

10: $\quad$ $\widehat{Q}_h(\cdot, \cdot) = \min \left\{ \bar{Q}_h(\cdot, \cdot), H - h + 1 \right\}_+$.

11: $\quad$ $\widehat{\pi}_h(\cdot) = \arg\max_{a \in \mathcal{A}} \widehat{Q}_h(\cdot, a)$.

12: $\quad$ $\widehat{V}_h(\cdot) = \widehat{Q}_h(\cdot, \widehat{\pi}_h(\cdot))$.

13: **end for**

Ouput: $\widehat{V} := \{\widehat{V}\}_{h=1}^{H+1}$, $\widehat{\pi} := \{\widehat{\pi}_h\}_{h=1}^H$.

---

no longer applicable. To this end, we carefully devise the penalty function, denoted as Γ_h, for Lin-RMDPs with TV distance. Compared to the prior art (Blanchet et al., 2024) which promotes pessimism by solving an inner constrained optimization problem, our proposed penalty function efficiently addresses the uncertainty in every feature dimension $i \in [d]$, which plays a crucial role in improving the sub-optimality gap.

3.3 Performance guarantees of `DROP`

Next, we provide the theoretical guarantees for `DROP`, under different assumptions on the batch data quality. We start without any coverage assumption on the batch dataset $\mathcal{D}$, where the proof is postponed to Appendix B.3.

Theorem 1. *Consider any $\delta \in (0, 1)$. Suppose that Assumption 1 and 2 hold. By setting*

$$\lambda_0 = 1, \quad \gamma_0 = 6\sqrt{d\xi_0} H, \quad \text{where } \xi_0 = \log(3HK/\delta),$$

one has with the probability at least $1 - 3\delta$, the policy $\widehat{\pi}$ generated by Algorithm 1 satisfies

$$\text{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \leq \widetilde{O}(\sqrt{d}H) \sum_{h=1}^H \sum_{i=1}^d \max_{P \in \mathcal{P}^\rho(P^0)} \mathbb{E}_{\pi^\star, P} \left[\left\| \phi_i(s_h, a_h) \mathbb{1}_i \right\|_{\Lambda_h^{-1}} \right]. \tag{17}$$

Since we do not impose any coverage assumption on the batch data, Theorem 1 demonstrates an "instance-dependent" sub-optimality gap, which can be controlled by some general term (the right hand side of (17)) and the confidence level δ. The sub-optimality gap largely depends on the quality of the batch data — specifically, how well it explores the feature space within $\mathbb{R}^d$. Building upon the above foundational result, we proceed to examine the sample complexity required to achieve an ϵ-optimal policy, considering different data qualities in the subsequent discussion.

3.3.1 The case of partial feature coverage

We first consider the partial feature coverage, which only compares the behavior policy with the optimal policy, in terms of the ability to explore each feature dimension. To accommodate with our Lin-RMDPs, we propose a tailored partial data coverage assumption, which depicts the worst-case

dissimilarity between the optimal robust policy $\pi^\star$ in any transition kernel $P \in \mathcal{P}^\rho(P^0)$ and the behavior policy π^b in the nominal kernel P^0 over every feature space dimension $i \in [d]$, as follows.

Assumption 3 (Robust single-policy clipped concentrability). *The behavior policy of the batch dataset $\mathcal{D}$ satisfies*

$$\max_{(u,h,i,P)\in\mathbb{R}^d\times[H]\times[d]\times\mathcal{P}^\rho(P^0)} \frac{u^\top \left(\min\{\mathbb{E}_{d_h^{\star,P}}\phi_i^2(s,a), 1/d\} \cdot \mathbb{1}_{i,i}\right) u}{u^\top \left(\mathbb{E}_{d_h^b}[\phi(s,a)\phi(s,a)^\top]\right) u} \le \frac{C_{\mathrm{rob}}^\star}{d}, \tag{18}$$

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

The quantity $C_{\mathrm{rob}}^\star$ measures the expected quality of batch data, by comparing to the desired dataset associated with the optimal robust policy. Intuitively, $C_{\mathrm{rob}}^\star$ decreases as the batch dataset contains more expert data, and increases when the quality of the dataset is poorer — generated from some policy far from the optimal policy. Note that prior knowledge of $C_{\mathrm{rob}}^\star$ is not required when implementing DROP in practice. Here, we assume $C_{\mathrm{rob}}^\star < \infty$, which requires that the behavior policy is able to explore the same feature dimensions as the optimal robust policy. Under Assumption 3, the following corollary provides the provable sample complexity of DROP to achieve an ϵ-optimal robust policy. The proof is postponed to Appendix B.4.

Corollary 1 (Partial feature coverage). *With Theorem 1 and Assumption 3 hold and consider any $\delta \in (0, 1)$. Let $d_{\min}^b = \min_{h,s,a}\{d_h^b(s,a) : d_h^b(s,a) > 0\}$. With probability exceeding $1 - 4\delta$, the policy $\widehat{\pi}$ returned by Algorithm 1 achieves*

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \le 96\sqrt{\frac{d^2 H^4 C_{\mathrm{rob}}^\star \log(3HK/\delta)}{K}}$$

as long as $K \ge c_0 \log(KH/\delta)/d_{\min}^b$ for some sufficiently large universal constant $c_0 > 0$. In other words, the learned policy $\widehat{\pi}$ is ϵ-optimal if the total number of sample trajectories satisfies

$$K \ge \widetilde{O}\Big(\frac{C_{\mathrm{rob}}^\star d^2 H^4}{\epsilon^2}\Big). \tag{19}$$

Notice that Corollary 1 implies the sub-optimality bound of DROP is comparable to that of the prior art in standard linear MDPs (Jin et al., 2021, Corollary 4.5), in terms of the feature dimension d and the horizon length H.

Comparisons to prior art for Lin-RMDPs. With high probability, the existing Assumption 6.2 proposed in Blanchet et al. (2024) can be transferred into the following condition (see (66)-(68) in Appendix B.4):

$$\max_{(u,h,i,P)\in\mathbb{R}^d\times[H]\times[d]\times\mathcal{P}^\rho(P^0)} \frac{u^\top \left(\mathbb{E}_{d_h^{\star,P}}\phi_i^2(s,a) \cdot \mathbb{1}_{i,i}\right) u}{u^\top \left(\mathbb{E}_{d_h^b}[\phi(s,a)\phi(s,a)^\top]\right) u} \le C_1^\star \in [1, \infty). \tag{20}$$

Thanks to the clipping operation in (18), $C_{\mathrm{rob}}^\star \le d \cdot C_1^\star$ for any given batch dataset $\mathcal{D}$. Notice that both $C_{\mathrm{rob}}^\star$ and $C_1^\star$ are lower bounded by 1. The proposed algorithm in Blanchet et al. (2024) can achieve ϵ-accuracy provided that the total number of sample trajectories obeys

$$K \ge \widetilde{O}\Big(\frac{C_1^\star d^4 H^4}{\epsilon^2}\Big).$$

It shows that the sample complexity (19) of DROP improves the one in Blanchet et al. (2024) by at least $\widetilde{O}(d)$.

3.3.2 The case of full feature coverage

Then, we introduce the following full feature coverage assumption that is also widely used in standard offline linear MDPs (Jin et al., 2021; Xiong et al., 2023; Yin et al., 2022; Ma et al., 2022), which requires the behavior policy exploring the feature space uniformly well.

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

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 DROP under the full feature coverage, where the proof is postponed to Appendix B.5.

Corollary 2 (Full feature coverage). *With Theorem 1 and Assumption 4 hold and consider any $\delta \in (0, 1)$. Let $d_{\min}^b = \min_{h,s,a}\{d_h^b(s, a) : d_h^b(s, a) > 0\}$. With probability at least $1 - 5\delta$, the policy $\widehat{\pi}$ returned by Algorithm 1 achieves*

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \leq 96\sqrt{\frac{dH^4 \log(3HK/\delta)}{\kappa K}},$$

as long as $K \geq \max\{c_0 \log(2Hd/\delta)/\kappa^2, c_0 \log(KH/\delta)/d_{\min}^b\}$ for some sufficiently large universal constant c_0. In other words, the learned policy $\widehat{\pi}$ is ϵ-optimal if the total number of sample trajectories satisfies

$$K \geq \widetilde{O}\Big(\frac{dH^4}{\kappa\epsilon^2}\Big). \tag{21}$$

Notice that the sample complexity in (21) matches the prior arts in standard linear MDPs (Yin et al., 2022; Xiong et al., 2023) when robustness is not considered. A careful reader may observe that (21) has better dependency on d compared to (19). While noting that the upper bound of κ is $1/d$ (Wang et al., 2021), the sample complexity of DROP (cf. (21)) is at least $\widetilde{O}(d^2H^4/\epsilon^2)$.

4 Tightening the Sample Complexity by Leveraging Variance Estimation

To tighten our results, we further explore the variance information to reweight the ridge regression in DROP and develop its variance-aware variant called DROP-V. The key idea is to design a tighter penalty term with the estimated variance, which is widely used in standard linear MDPs (Zhou et al., 2021a; Min et al., 2021; Yin et al., 2022; Xiong et al., 2023) to achieve near-optimal results.

4.1 DROP-V: a variance-aware variant of DROP

We first highlight the design features of DROP-V that are different from DROP, which can boiled down to the following two steps.

Constructing a variance estimator. First, we run the Algorithm 1 on a sub-dataset $\tilde{\mathcal{D}}^0 \in \mathcal{D}$ to obtain the estimated value function $\{\tilde{V}_h\}_{h=1}^{H+1}$. Then with $\{\tilde{V}_h\}_{h=1}^{H+1}$ at our hands, we design the variance estimator $\widehat{\sigma}_h^2 : \mathcal{S} \times \mathcal{A} \to [1, H^2]$ by

$$\widehat{\sigma}_h^2(s, a) = \max\{[\phi(s, a)^\top \nu_{h,1}]_{[0,H^2]} - ([\phi(s, a)^\top \nu_{h,2}]_{[0,H]})^2, 1\}, \quad \forall(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{22}$$

where $\nu_{h,1}$ and $\nu_{h,2}$ are obtained from ridge regression as follows:

$$\nu_{h,1} = \arg\min_{\nu \in \mathbb{R}^d} \sum_{\tau \in \tilde{\mathcal{D}}_h^0} \big(\phi(s_h^\tau, a_h^\tau)^\top \nu - (\tilde{V}_{h+1}(s_{h+1}^\tau))^2\big)^2 + \|\nu\|_2^2, \tag{23}$$

$$\nu_{h,2} = \arg\min_{\nu \in \mathbb{R}^d} \sum_{\tau \in \tilde{\mathcal{D}}_h^0} \big(\phi(s_h^\tau, a_h^\tau)^\top \nu - \tilde{V}_{h+1}(s_{h+1}^\tau)\big)^2 + \|\nu\|_2^2. \tag{24}$$

Notice that the technique of variance estimation is also used in Xiong et al. (2023); Yin et al. (2022) for standard linear MDPs, while we address the temporal dependency via a carefully designed three-fold subsampling approach detailed in Appendix C.2.

Variance-weighted ridge regression. Incorporating with the variance estimator $\widehat{\sigma}_h^2$ computed on $\tilde{\mathcal{D}}^0$, we replace the ridge regression updates (i.e. (12) and (14)) by their variance-weighted counterparts as follows:

$$\widehat{\theta}_h^\sigma = \arg\min_{\theta\in\mathbb{R}^d} \sum_{\tau\in\mathcal{D}_h^0} \frac{\left(\phi(s_h^\tau, a_h^\tau)^\top \theta - r_h^\tau\right)^2}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} + \lambda_1\|\theta\|_2^2 = \Sigma_h^{-1}\Big(\sum_{\tau\in\mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau) r_h^\tau}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)}\Big), \tag{25}$$

$$\bar{\nu}_h^{\rho,\sigma,V}(\alpha) = \arg\min_{\nu\in\mathbb{R}^d} \sum_{\tau\in\mathcal{D}_h^0} \frac{\left(\phi(s_h^\tau, a_h^\tau)^\top \nu - [V]_\alpha(s_{h+1}^\tau)\right)^2}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} + \lambda_1\|\nu\|_2^2 = \Sigma_h^{-1}\Big(\sum_{\tau\in\mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)[V]_\alpha(s_{h+1}^\tau)}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)}\Big), \tag{26}$$

for any value function $V : \mathcal{S} \to [0, H]$ and $h \in [H]$, where

$$\Sigma_h := \sum_{\tau\in\mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} + \lambda_1 I_d$$

with the regularization coefficient λ_1 and $\mathcal{D}^0$ is another sub-dataset of $\mathcal{D}$ that is independent of $\tilde{\mathcal{D}}^0$. Accordingly, DROP-V constructs an empirical variance-aware robust Bellman operator as

$$\big(\widehat{\mathbb{B}}_h^{\rho,\sigma} V\big)(s,a) = \phi(s,a)^\top \big(\widehat{\theta}_h^\sigma + \widehat{\nu}_h^{\rho,\sigma,V}\big), \tag{27}$$

where the i-th coordinate of $\widehat{\nu}_h^{\rho,\sigma,V}$ is defined as

$$\widehat{\nu}_{h,i}^{\rho,\sigma,V} = \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \big\{\bar{\nu}_{h,i}^{\rho,\sigma,V}(\alpha) - \rho(\alpha - \min_{s'} V(s'))\big\}. \tag{28}$$

Similar to DROP, we also perform the pessimistic value iterations, where the estimated Q function is updated by

$$\bar{Q}_h(s,a) = \widehat{\mathbb{B}}_h^{\rho,\sigma}(\widehat{V}_{h+1})(s,a) - \underbrace{\gamma_1 \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Sigma_h^{-1}}}_{\text{penalty function } \Gamma_h^\sigma : \mathcal{S}\times\mathcal{A}\to\mathbb{R}}, \quad \forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H],$$

with a fixed, estimated $\widehat{V}_{h+1}$ obtained from previous iteration and some coefficient $\gamma_1 > 0$.

The rest of DROP-V follows the procedure described in Algorithm 1. To avoid redundancy, the detailed implementation of DROP-V is provided in Appendix C.1.

4.2 Performance guarantees of DROP-V

Then, we are ready to present our improved results, where the proof is postponed to Appendix C.3.

Theorem 2. *Suppose that Assumption 1, 2, and 4 hold and consider any $\delta \in (0,1)$. In DROP-V, we set*

$$\lambda_1 = 1/H^2, \gamma_1 = \xi_1\sqrt{d}, \quad \text{where } \xi_1 = 66\log(3HK/\delta).$$

Provided that $\sqrt{d} \geq H$ and $K \geq \widetilde{O}\big(\frac{H^4 + H^6 d\kappa}{\kappa^2} + \frac{1}{d_{\min}^b}\big)$, then with probability exceeding $1 - \delta$, the robust policy $\widehat{\pi}$ learned by DROP-V satisfies

$$\text{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \leq \widetilde{O}(\sqrt{d}) \sum_{h=1}^H \sum_{i=1}^d \max_{P\in\mathcal{P}^\rho(P^0)} \mathbb{E}_{\pi^\star, P}\left[\|\phi_i(s_h, a_h)\mathbb{1}_i\|_{(\Sigma_h^\star)^{-1}}\right], \tag{29}$$

where the variance-weighted cumulative sample covariance matrix $\Sigma_h^\star$ is defined by

$$\Sigma_h^\star = \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top}{\max\{1, \mathrm{Var}_{P_h^0}[V_{h+1}^{\star,\rho}](s_h^\tau, a_h^\tau)\}} + \frac{1}{H^2}I_d, \tag{30}$$

and $\mathrm{Var}_{P_h^0}[V](s,a) = \int_\mathcal{S} P_{h,s,a}^0(s')V^2(s')\mathrm{d}s' - (\int_\mathcal{S} P_{h,s,a}^0(s')V(s')\mathrm{d}s')^2$ *represents the conditional variance for any value function* $V : \mathcal{S} \to [0, H]$ *and any* $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

Compared to the instance-dependent sub-optimality bound of `DROP` (cf. Theorem 1), the above guarantee for `DROP-V` is tighter since $H^2\Lambda_h^{-1} \succeq (\Sigma_h^\star)^{-1}$. The underlying reason for the improvement is the use of the variance estimator to control the conditional variance and the tighter penalty function designed via the Bernstein-type inequality, while that of `DROP` depends on the Hoeffding-type counterpart.

5 Conclusion

In this paper, we investigate the sample complexity for distributionally robust offline RL when the model has linear representations and the uncertainty set can be characterized by TV distance. We develop a robust variant of pessimistic value iteration with linear function approximation, called `DROP`. We establish the sub-optimality guarantees for `DROP` under various offline data assumptions. Compared to the prior art, `DROP` notably improves the sample complexity by at least $\widetilde{O}(d)$, under the partial feature coverage assumption. We further incorporate `DROP` with variance estimation to develop an enhanced `DROP`, referred to as `DROP-V`, which improves the sub-optimality bound of `DROP`. In the future, it will be of interest to consider different choices of the uncertainty sets and establish the lower bound for the entire range of the uncertainty level.

Acknowledgement

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24.

Agarwal, A., Jin, Y., and Zhang, T. (2023). VOQL: Towards optimal regret in model-free RL with nonlinear function approximation. *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063.

Badrinath, K. P. and Kalathil, D. (2021). Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR.

Bellemare, M. G., Le Roux, N., Castro, P. S., and Moitra, S. (2019). Distributional reinforcement learning with linear function approximation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2203–2211. PMLR.

Bertsekas, D. (2012). *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific.

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. (2024). Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36.

Derman, E. and Mannor, S. (2020). Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*.

Ding, W., Shi, L., Chi, Y., and Zhao, D. (2024). Seeing is not believing: Robust reinforcement learning against spurious correlation. *Advances in Neural Information Processing Systems*, 36.

Dong, J., Li, J., Wang, B., and Zhang, J. (2022). Online policy optimization for robust MDP. *arXiv preprint arXiv:2209.13841*.

Goyal, V. and Grand-Clement, J. (2022). Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*.

Gu, S., Chen, G., Zhang, L., Hou, J., Hu, Y., and Knoll, A. (2022). Constrained reinforcement learning for vehicle motion planning with topological reachability analysis. *Robotics*, 11(4):81.

He, J., Zhao, H., Zhou, D., and Gu, Q. (2023). Nearly minimax optimal reinforcement learning for linear Markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR.

Ho, C. P., Petrik, M., and Wiesemann, W. (2018). Fast Bellman updates for robust MDPs. In *International Conference on Machine Learning*, pages 1979–1988. PMLR.

Ho, C. P., Petrik, M., and Wiesemann, W. (2021). Partial policy iteration for l_1-robust Markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46.

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

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096.

Kaufman, D. L. and Schaefer, A. J. (2013). Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410.

Kumar, N., Derman, E., Geist, M., Levy, K. Y., and Mannor, S. (2024). Policy gradient for rectangular robust Markov decision processes. *Advances in Neural Information Processing Systems*, 36.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2024). Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260.

Li, Y. and Lan, G. (2023). First-order policy optimization for robust policy evaluation. *arXiv preprint arXiv:2307.15890*.

Li, Y., Zhao, T., and Lan, G. (2022). First-order policy optimization for robust Markov decision process. *arXiv preprint arXiv:2209.10579*.

Liang, Z., Ma, X., Blanchet, J., Zhang, J., and Zhou, Z. (2023). Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*.

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. (2022). Distributionally robust Q-learning. In *International Conference on Machine Learning*, pages 13623–13643. PMLR.

Liu, Z. and Xu, P. (2024a). Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. *arXiv preprint arXiv:2402.15399*.

Liu, Z. and Xu, P. (2024b). Minimax optimal and computationally efficient algorithms for distributionally robust offline reinforcement learning. *arXiv preprint arXiv:2403.09621*.

Ma, X., Liang, Z., Blanchet, J., Liu, M., Xia, L., Zhang, J., Zhao, Q., and Zhou, Z. (2022). Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*.

Min, Y., Wang, T., Zhou, D., and Gu, Q. (2021). Variance-aware off-policy evaluation with linear function approximation. *Advances in Neural Information Processing Systems*, 34:7598–7610.

Pan, Y., Chen, Y., and Lin, F. (2024). Adjustable robust reinforcement learning for online 3D bin packing. *Advances in Neural Information Processing Systems*, 36.

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR.

Prashanth, L. and Bhatnagar, S. (2010). Reinforcement learning with function approximation for traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):412–421.

Ramesh, S. S., Sessa, P. G., Hu, Y., Krause, A., and Bogunovic, I. (2024). Distributionally robust model-based reinforcement learning with large state spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 100–108. PMLR.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716.

Shi, L. and Chi, Y. (2022). Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). 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.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2023). The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36.

Shi, L., Mazumdar, E., Chi, Y., and Wierman, A. (2024). Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. *arXiv preprint arXiv:2404.18909*.

Smirnova, E., Dohmatob, E., and Mary, J. (2019). Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*.

Tamar, A., Mannor, S., and Xu, H. (2014). Scaling up robust MDPs using function approximation. In *International conference on machine learning*, pages 181–189. PMLR.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation.* Springer Publishing Company, Incorporated, 1st edition.

Uehara, M. and Sun, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*.

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

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wang, R., Foster, D., and Kakade, S. M. (2021). What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023a). A finite sample complexity bound for distributionally robust Q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3370–3398. PMLR.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023b). On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023c). Sample complexity of variance-reduced distributionally robust Q-learning. *arXiv preprint arXiv:2305.18420*.

Wang, Y. and Zou, S. (2021). Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34.

Wolff, E. M., Topcu, U., and Murray, R. M. (2012). Robust control of uncertain Markov decision processes with temporal logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3372–3379. IEEE.

Woo, J., Shi, L., Joshi, G., and Chi, Y. (2024). Federated offline reinforcement learning: Collaborative single-policy coverage suffices. *arXiv preprint arXiv:2402.05876*.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694.

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. (2023). Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent MDP and Markov game. In *The Eleventh International Conference on Learning Representations*.

Xu, H. and Mannor, S. (2012). Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300.

Xu, Z., Panaganti, K., and Kalathil, D. (2023). Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9728–9754. PMLR.

Yan, Y., Li, G., Chen, Y., and Fan, J. (2023). The efficacy of pessimism in asynchronous Q-learning. *IEEE Transactions on Information Theory*.

Yang, L. and Wang, M. (2019). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004.

Yang, L. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR.

Yang, W., Wang, H., Kozuno, T., Jordan, S. M., and Zhang, Z. (2023). Avoiding model estimation in robust Markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*.

Yang, W., Zhang, L., and Zhang, Z. (2022). Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. (2022). Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*.

Yu, C., Liu, J., Nemati, S., and Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys*, 55(1):1–36.

Zanette, A., Wainwright, M. J., and Brunskill, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640.

Zhang, R., Hu, Y., and Li, N. (2023). Regularized robust MDPs and risk-sensitive MDPs: Equivalence, policy gradient, and sample complexity. *arXiv preprint arXiv:2306.11626*.

Zhou, D., Gu, Q., and Szepesvari, C. (2021a). Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4532–4576. PMLR.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. (2021b). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR.

A Technical Lemmas

A.1 Proof of Lemma 1

For a Lin-RMDP satisfying Assumption 2, the robust linear transition operator defined in (4) obeys: for any time step $h \in [H]$,

$$
\begin{aligned}
(\mathbb{P}_h^\rho f)(s,a) &= \inf_{\mu_h \in \mathcal{U}^\rho(\mu_h^0)} \int_{\mathcal{S}} \phi(s,a)^\top \mu_h(s') f(s') \mathrm{d}s' \\
&= \inf_{\mu_h \in \mathcal{U}^\rho(\mu_h^0)} \sum_{i=1}^d \phi_i(s,a) \int_{\mathcal{S}} \mu_{h,i}(s') f(s') \mathrm{d}s' \\
&= \sum_{i=1}^d \phi_i(s,a) \inf_{\mu_{h,i} \in \mathcal{U}^\rho(\mu_{h,i}^0)} \int_{\mathcal{S}} \mu_{h,i}(s') f(s') \mathrm{d}s' \\
&= \phi(s,a)^\top \left(\inf_{\mu_h \in \mathcal{U}^\rho(\mu_h^0)} \int_{\mathcal{S}} \mu_h(s') f(s') \mathrm{d}s' \right),
\end{aligned}
$$

where the penultimate equality is due to $\phi_i(s,a) \geq 0, \forall (i,s,a) \in [d] \times \mathcal{S} \times \mathcal{A}$ and $\mathcal{U}^\rho(\mu_h^0) := \otimes_{[d]} \mathcal{U}^\rho(\mu_{h,i}^0)$. Therefore, the robust linear Bellman operator defined in (3) is linear in the feature map ϕ: for $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$
\begin{aligned}
(\mathbb{B}_h^\rho f)(s,a) &= r_h(s,a) + [\mathbb{P}_h^\rho f](s,a) \\
&= \phi(s,a)^\top \theta_h + \phi(s,a)^\top \left(\inf_{\mu_h \in \mathcal{U}^\rho(\mu_h^0)} \int_{\mathcal{S}} \mu_h(s') f(s') \mathrm{d}s' \right) \\
&= \phi(s,a)^\top \underbrace{\left(\theta_h + \inf_{\mu_h \in \mathcal{U}^\rho(\mu_h^0)} \int_{\mathcal{S}} \mu_h(s') f(s') \mathrm{d}s' \right)}_{:= w_h^\rho},
\end{aligned}
$$

where the second equality is due to Assumption 1.

A.2 Preliminary facts

A.2.1 Useful concenrtation inequalities

Lemma 2 (Hoeffding-type inequality for self-normalized process (Abbasi-Yadkori et al., 2011))**.** *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 $x_t \leq L$. 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*

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

Lemma 3 (Bernstein-type inequality for self-normalized process (Zhou et al., 2021a)). *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 $x_t \leq L$. Let $\Lambda_t = \lambda I_d + \sum_{s=1}^{t} x_s x_s^{\top}$. Assume that*

$$|\eta_t| \leq R, \quad \mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0, \quad \mathbb{E}[\eta_t^2 | \mathcal{F}_{t-1}] \leq \sigma^2.$$

Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $t > 0$, we have

$$\|\sum_{s=1}^{t} x_s \eta_s\|_{\Lambda_t^{-1}} \leq 8\sigma \sqrt{d \log\left(1 + \frac{tL^2}{\lambda d}\right) \log(\frac{4t^2}{\delta})} + 4R \log(\frac{4t^2}{\delta}).$$

Lemma 4 (Lemma H.5, Min et al. (2021)). *Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a bounded function such that $\|\phi(s,a)\|_2 \leq C$ 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 $G = \mathbb{E}_{\nu}[\phi(s,a)\phi(s,a)^{\top}]$. Then for any $\delta \in (0,1)$, if K satisfies that*

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

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

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

Lemma 5 (Lemma 5.1, Jin et al. (2021)). *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*

$$|(\widehat{\mathbb{B}}_h^{\rho}\widehat{V}_{h+1})(s,a) - (\mathbb{B}_h^{\rho}\widehat{V}_{h+1})(s,a)| \leq \Gamma_h(s,a), \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{31}$$

we have

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

A.2.2 Useful facts

Lemma 6. *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*

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

Proof. Let $\alpha_1^{\star} = \arg\max_{\alpha \in \mathcal{C}} f_1(\alpha)$. Then,

$$\begin{aligned}
\max_{\alpha \in \mathcal{C}} f_1(\alpha) - \max_{\alpha \in \mathcal{C}} f_2(\alpha) &\leq f_1(\alpha_1^{\star}) - \max_{\alpha \in \mathcal{C}} f_2(\alpha) \\
&\leq f_1(\alpha_1^{\star}) - f_2(\alpha_1^{\star}) \\
&\leq \max_{\alpha \in \mathcal{C}}\left(f_1(\alpha) - f_2(\alpha)\right).
\end{aligned}$$

$\square$

Lemma 7. *For any positive semi-definite matrix $A \in \mathbb{R}^{d \times d}$ and any constant $c \geq 0$, we have*

$$\mathrm{Tr}\left(A(I + cA)^{-1}\right) \leq \sum_{i=1}^{d} \frac{\lambda_i}{1 + c\lambda_i}, \tag{32}$$

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

Proof. Note that

$$
\begin{aligned}
A(I + cA)^{-1} &= A(I + cA)^{-1} + \frac{1}{c}(I + cA)^{-1} - \frac{1}{c}(I + cA)^{-1} \\
&= \frac{1}{c}I - \frac{1}{c}(I + cA)^{-1}.
\end{aligned}
$$

In addition, the eigenvalues of $(I + cA)^{-1}$ are $\{\frac{1}{1+c\lambda_i}\}_{i=1}^d$. Therefore,

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

$\square$

Lemma 8 (modified Lemma 4, Shi et al. (2023)). *Consider any probability vector $\mu^0 \in \Delta(\mathcal{S})$, any fixed uncertainty level ρ and the uncertainty set $\mathcal{U}^\rho(\mu^0)$ satisfying Assumption 2. For any value function $V : \mathcal{S} \to [0, H]$, recalling the definition of $[V]_\alpha$ in (11), one has*

$$
\inf_{\mu \in \mathcal{U}^\rho(\mu^0)} \int_{\mathcal{S}} \mu(s')V(s')\mathrm{d}s' = \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \{\mathbb{E}_{s' \sim \mu^0}[V]_\alpha(s') - \rho(\alpha - \min_{s'}[V]_\alpha(s'))\}. \tag{33}
$$

B Analysis for `DROP`: Algorithm 1

B.1 Proof of equation (10)

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

$$
\begin{aligned}
(\mathbb{B}_h^\rho V)(s, a) &= r_h(s, a) + \inf_{\mu_h \in \mathcal{U}^\rho(\mu_h^0)} \int_{\mathcal{S}} \phi(s, a)^\top \mu_h(s')V(s')\mathrm{d}s' \\
&= \phi(s, a)^\top \theta_h + \sum_{i=1}^d \phi_i(s, a) \inf_{\mu_{h,i} \in \mathcal{U}^\rho(\mu_{h,i}^0)} \int_{\mathcal{S}} \mu_{h,i}(s')V(s')\mathrm{d}s'.
\end{aligned}
$$

According to Lemma 8, for any value function $V : \mathcal{S} \to [0, H]$ and any uncertainty set $\mathcal{U}^\rho(\mu_{h,i}^0), \forall (h, i) \in [H] \times [d]$ that satisfies Assumption 2, one has

$$
\inf_{\mu_{h,i} \in \mathcal{U}^\rho(\mu_{h,i}^0)} \int_S \mu_{h,i}(s')V(s')\mathrm{d}s' = \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ \int_{\mathcal{S}} \mu_{h,i}(s')[V]_\alpha(s')\mathrm{d}s' - \rho(\alpha - \min_{s'}[V]_\alpha(s')) \right\}.
$$

Denote $\nu_h^{\rho,V} = [\nu_{h,1}^{\rho,V}, \nu_{h,2}^{\rho,V} \ldots, \nu_{h,d}^{\rho,V}]^\top \in \mathbb{R}^d$, where $\nu_{h,i}^{\rho,V} = \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ \int_{\mathcal{S}} \mu_{h,i}(s')[V]_\alpha(s')\mathrm{d}s' - \rho(\alpha - \min_{s'}[V]_\alpha(s')) \right\}$ for any $i \in [d]$. Therefore,

$$
(\mathbb{B}_h^\rho V)(s, a) = \phi(s, a)^\top \left(\theta_h + \nu_h^{\rho,V}\right).
$$

B.2 Two-fold subsampling method

To tackle the temporal dependency in batch dataset $\mathcal{D}$, we apply the insights from the subsampling approach inspired by Li et al. (2024). 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 shows that (34) is a valid lower bound of $N_h^{\mathrm{main}}(s)$ for any $s \in \mathcal{S}$ and $h \in [H]$, which can be obtained by a slight modification on Lemma 3 and Lemma 7 in Li et al. (2024)

Algorithm 2 `Two-fold-subsampling`

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

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

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

$$N_h^{\mathrm{trim}}(s) = \max\{N_h^{\mathrm{aux}}(s) - 10\sqrt{N_h^{\mathrm{aux}}(s)\log\frac{KH}{\delta}}, 0\}. \tag{34}$$

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

Ouput: $\mathcal{D}^0 = \mathcal{D}^{\mathrm{main,sub}}$.

Lemma 9. *With probability at least $1 - 2\delta$, if $N_h^{\mathrm{trim}}(s)$ satisfies (34) for every $s \in \mathcal{S}$ and $h \in [H]$, then $\mathcal{D}^0 := \mathcal{D}^{\mathrm{main,sub}}$ contains temporally statistically independent samples and the following bound holds simultaneously, i.e.,*

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

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

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

Proof. Let $\mathcal{S}_D$ be the collection of all the states appearing in any batch dataset $\mathcal{D}$, where $|\mathcal{S}_D| \le K$. By changing the union bound over $\mathcal{S}$ to $\mathcal{S}_D$ in the proof of Li et al. (2024, Lemma 3), the remaining proof still holds, since $N_h^{\mathrm{trim}}(s) = N_h^{\mathrm{main}}(s) = 0, \forall s \notin \mathcal{S}_D$. Together with Li et al. (2024, Lemma 7), $\mathcal{D}^0$ contains temporally statistically independent samples if $N_h^{\mathrm{trim}}(s) \le N_h^{\mathrm{main}}(s), \forall(s,h) \in \mathcal{S} \times [H]$. $\square$

B.3 Proof of Theorem 1

Notations. Before starting the proof of Theorem 1, we introduce several notations for the convenience of the following analysis. First, we use

$$\iota_h(s,a) = \mathbb{B}_h^\rho \widehat{V}_{h+1}(s,a) - \widehat{Q}_h(s,a), \quad \forall(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}, \tag{35}$$

to represent the model evaluation error at the h-th step of `DROP`. In addition, we denote the estimated weight of the transition kernel at the h-th step by

$$\forall(s,h) \in \mathcal{S} \times [H]: \quad \widehat{\mu}_h(s) = \Lambda_h^{-1}\left(\sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)\mathbb{1}(s_{h+1}^\tau = s)\right) \in \mathbb{R}^d, \tag{36}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Accordingly, it holds that $\bar{\nu}_h^{\widehat{V}}(\alpha) = \int_{\mathcal{S}} \widehat{\mu}_h(s')[\widehat{V}_{h+1}(s')]_\alpha \mathrm{d}s' \in \mathbb{R}^d$. 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

$$\mathcal{D}_h^\star = \left\{\left[d_h^{\star,P}(s)\right]_{s \in \mathcal{S}} : P \in \mathcal{P}^\rho(P^0)\right\} = \left\{\left[d_h^{\star,P}(s, \pi_h^\star(s))\right]_{s \in \mathcal{S}} : P \in \mathcal{P}^\rho(P^0)\right\}, \tag{37}$$

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 3. *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 and 2 hold. In* DROP, *we set*

$$\lambda_0 = 1, \quad \gamma_0 = 6\sqrt{d\xi_0}H, \quad \text{where } \xi_0 = \log(3HK/\delta). \tag{38}$$

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

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \leq \tilde{O}(\sqrt{d}H) \sum_{h=1}^H \sum_{i=1}^d \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{d_h^\star}\left[\left\|\phi_i(s_h, a_h)\mathbb{1}_i\right\|_{\Lambda_h^{-1}}\right].$$

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 14, $N_h^{\mathrm{trim}}(s)$ is a valid sampling number for any $s \in \mathcal{S}$ and $h \in [H]$ such that $|\mathcal{D}_h^0| = \sum_{s\in\mathcal{S}} N_h^{\mathrm{trim}}(s) \leq K$, and $\mathcal{D}_h^0$ can be treated as containing temporally statistically independent samples with probability exceeding $1 - 2\delta$. Therefore, by invoking Theorem 3 with $N_h := |\mathcal{D}_h^0|$, we have

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \leq \tilde{O}(\sqrt{d}H) \sum_{h=1}^H \sum_{i=1}^d \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{d_h^\star}\left[\left\|\phi_i(s_h, a_h)\mathbb{1}_i\right\|_{\Lambda_h^{-1}}\right],$$

with probability exceeding $1 - 3\delta$.

B.3.2 Proof of Theorem 3

The proof of Theorem 3 can be summarized as following key steps.

Step 1: establishing the pessimistic property. To substantiate the pessimism, we heavily depend on the following lemma, where the proof is postponed to Appendix B.3.3.

Lemma 10. *Suppose all the assumptions in Theorem 3 hold and follow all the parameters setting in (38). Then for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, with probability at least $1 - \delta$, the value function $\{\widehat{V}\}_{h=1}^H$ generated by* DROP *satisfies*

$$|(\widehat{\mathbb{B}}_h^\rho \widehat{V}_{h+1})(s, a) - (\mathbb{B}_h^\rho \widehat{V}_{h+1})(s, a)| \leq \Gamma_h(s, a) := \gamma_0 \sum_{i=1}^d \left\|\phi_i(s, a)\mathbb{1}_i\right\|_{\Lambda_h^{-1}}. \tag{39}$$

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

$$Q_h^{\star,\rho}(s, a) \geq Q_h^{\widehat{\pi},\rho}(s, a) \geq \widehat{Q}_h(s, a) \quad \text{and} \quad V_h^{\star,\rho}(s) \geq V_h^{\widehat{\pi},\rho}(s) \geq \widehat{V}_h(s), \quad \forall(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{40}$$

if the condition (39) holds. It implies that $\widehat{Q}_h(s, a)$ and $\widehat{V}_h(s)$ is the pessimistic estimates of $Q_h^{\widehat{\pi},\rho}(s, a)$ and $V_h^{\widehat{\pi},\rho}(s)$ for any $s \in \mathcal{S}$, respectively. Notice that if $Q_h^{\widehat{\pi},\rho}(s, a) \geq \widehat{Q}_h(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, one simultaneously has the following relation:

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

Therefore, we shall verify that

$$Q_h^{\widehat{\pi},\rho}(s, a) \geq \widehat{Q}_h(s, a), \quad \forall(s, a) \in \mathcal{S} \times \mathcal{A}, \tag{41}$$

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

- *At step $h = H + 1$:* From the initialization step in Algorithm 1, we have $Q_{H+1}^{\widehat{\pi},\rho}(s,a) = \widehat{Q}_{H+1}(s,a) = 0$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, and (41) holds.

- *For any step $h \leq H$:* Suppose $Q_{h+1}^{\widehat{\pi},\rho}(s,a) \geq \widehat{Q}_{h+1}(s,a)$. From (40), we have $V_{h+1}^{\widehat{\pi},\rho}(s) \geq \widehat{V}_{h+1}(s)$. Therefore, if $\widehat{Q}_h(s,a) = 0$, $Q_h^{\widehat{\pi},\rho}(s,a) \geq 0 = \widehat{Q}_h(s,a)$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Otherwise,

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

where the first inequality is from the definition of $\widehat{Q}_h(s,a)$ (cf. Line 10 in Algorithm 1), and third inequality is based on the condition (39).

Combining these two cases, for any $h \in [H+1]$, we could verify the pessimistic property, i.e. the equation (40).

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

$$
\begin{aligned}
V_h^{\star,\rho}(s) - V_h^{\widehat{\pi},\rho}(s) &= V_h^{\star,\rho}(s) - \widehat{V}_h(s) + \widehat{V}_h(s) - V_h^{\widehat{\pi},\rho}(s) \\
&\leq V_h^{\star,\rho}(s) - \widehat{V}_h(s),
\end{aligned} \tag{42}
$$

where the inequality is due to the pessimistic property in (40).

In the following, we will control the value difference, i.e., $V_h^{\star,\rho}(s) - \widehat{V}_h(s)$, for any $(s,h) \in \mathcal{S} \times [H]$. First, recall the definition of $\widehat{V}_h$ (cf. Line 12 in Algorithm 1) and the robust Bellman consistency equation (2). For all $s \in \mathcal{S}$,

$$
\begin{aligned}
V_h^{\star,\rho}(s) - \widehat{V}_h(s) &= Q_h^{\star,\rho}(s, \pi_h^\star(s)) - \widehat{Q}_h(s, \widehat{\pi}_h(s)) \\
&\leq Q_h^{\star,\rho}(s, \pi_h^\star(s)) - \widehat{Q}_h(s, \pi_h^\star(s)),
\end{aligned} \tag{43}
$$

where the last inequality is from $\widehat{\pi}_h$ is the greedy policy with respect to $\widehat{Q}_h$ (cf. Line 11 in Algorithm 1.) From the definition of the model evaluation error (i.e., equation (35)) and the robust Bellman optimality equation (6), we have

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

which leads to

$$
Q_h^{\star,\rho}(s, \pi_h^\star(s)) - \widehat{Q}_h(s, \pi_h^\star(s)) = (\mathbb{P}_h^\rho V_{h+1}^{\star,\rho})(s, \pi_h^\star(s)) - (\mathbb{P}_h^\rho \widehat{V}_{h+1})(s, \pi_h^\star(s)) + \iota_h(s, \pi_h^\star(s)), \quad \forall s \in \mathcal{S}. \tag{44}
$$

Denote

$$
P_{h,s,\pi_h^\star(s)}^{\inf, \widehat{V}}(\cdot) := \arg\min_{P(\cdot) \in \mathcal{P}^\rho(P_{h,s,\pi_h^\star(s)}^0)} \int_S P(s') \widehat{V}_{h+1}(s') \mathrm{d}s'. \tag{45}
$$

Therefore, (44) becomes

$$Q_h^{\star,\rho}(s, \pi_h^\star(s)) - \widehat{Q}_h(s, \pi_h^\star(s)) \le \int_{\mathcal{S}} P_{h,s,\pi_h^\star(s)}^{\inf,\widehat{V}}(s') \left(V_{h+1}^{\star,\rho}(s') - \widehat{V}_{h+1}(s') \right) \mathrm{d}s' + \iota_h(s, \pi_h^\star(s)), \ \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{46}$$

Substituting (46) into (43), one has

$$V_h^{\star,\rho}(s) - \widehat{V}_h(s) \le \int_{\mathcal{S}} P_{h,s,\pi_h^\star(s)}^{\inf,\widehat{V}}(s') \left(V_{h+1}^{\star,\rho}(s') - \widehat{V}_{h+1}(s') \right) \mathrm{d}s' + \iota_h(s, \pi_h^\star(s)). \tag{47}$$

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

$$\widehat{P}_h^{\inf}(s) = P_{h,s,\pi_h^\star(s)}^{\inf,\widehat{V}}(\cdot) \quad \text{and} \quad \text{and} \quad \iota_h^\star(s) := \iota_h(s, \pi^\star(s)), \quad \forall s \in \mathcal{S}. \tag{48}$$

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

$$V_h^{\star,\rho}(s) - \widehat{V}_h(s) = \langle \mathbb{1}_s, V_h^{\star,\rho} - \widehat{V}_h \rangle \le \left(\prod_{t=h}^{H} \widehat{P}_j^{\inf} \right) \left(V_{H+1}^{\star,\rho} - \widehat{V}_{H+1} \right)(s) + \sum_{t=h}^{H} \left(\prod_{j=h}^{t-1} \widehat{P}_j^{\inf} \right) \iota_t^\star(s)$$

$$= \sum_{t=h}^{H} \left(\prod_{j=h}^{t-1} \widehat{P}_j^{\inf} \right) \iota_t^\star(s),$$

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

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

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

Together with (42), the sub-optimality gap defined in (9) satisfies

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \le \mathbb{E}_{s_1 \sim \zeta} V_1^{\star,\rho}(s_1) - \mathbb{E}_{s_1 \sim \zeta} \widehat{V}_1(s_1) \le \sum_{t=1}^{H} \mathbb{E}_{s_t \sim d_{1:t}^\star} \iota_t^\star(s_t). \tag{49}$$

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

$$\Gamma_h^\star(s) = \Gamma_h(s, \pi_h^\star(s)), \quad \forall s \in \mathcal{S}. \tag{50}$$

Combining Lemma 5 together with Lemma 10 will lead to

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \le 2 \sum_{h=1}^{H} \mathbb{E}_{s_h \sim d_{1:h}^\star} \Gamma_h^\star(s_h).$$

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 (37), we have $d_{1:h}^\star \in \mathcal{D}_h^\star$ and correspondingly,

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \le 2 \sum_{h=1}^{H} \mathbb{E}_{s_h \sim d_{1:h}^\star} \Gamma_h^\star(s_h)$$

$$\le 2\gamma_0 \sum_{h=1}^{H} \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{(s_h, a_h) \sim d_h^\star} \left[\sum_{i=1}^{d} \|\phi_i(s_h, a_h) \mathbb{1}_i\|_{\Lambda_h^{-1}} \right],$$

with probability exceeding $1 - \delta$.

B.3.3 Proof of Lemma 10

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

Lemma 11. *Suppose Assumption 1 and 2 hold. Then, for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the estimated value function $\widehat{V}_{h+1}$ generated by* DROP *satisfies*

$$|(\widehat{\mathbb{B}}_h^\rho \widehat{V}_{h+1})(s,a) - (\mathbb{B}_h^\rho \widehat{V}_{h+1})(s,a)|$$

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

(51)

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

We observe that the second term (i.e., $T_{1,h}$) in (51) will become dominating, as long as λ_0 is sufficiently small. In the following analysis, we will control $T_{1,h}$ via uniform concentration and the concentration of self-normalized process.

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

$$|\alpha - \alpha^\dagger| \leq \epsilon_0. \tag{52}$$

Then we can rewrite $T_{1,h}^2$ as

$$T_{1,h}^2 = \max_{\alpha \in [\min_s \widehat{V}_{h+1}(s), \max_s \widehat{V}_{h+1}(s)]} \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \left(\epsilon_h^\tau(\alpha, \widehat{V}_{h+1}) - \epsilon_h^\tau(\alpha^\dagger, \widehat{V}_{h+1}) + \epsilon_h^\tau(\alpha^\dagger, \widehat{V}_{h+1}) \right) \|_{\Lambda_h^{-1}}^2$$

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

$$\leq 8\epsilon_0^2 K^2 / \lambda_0 + 2\underbrace{\| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha^\dagger, \widehat{V}_{h+1}) \|_{\Lambda_h^{-1}}^2}_{T_{2,h}}, \tag{53}$$

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,

$$T_{2,h} \leq \sup_{\alpha \in \mathcal{N}(\epsilon_0, H)} \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, \widehat{V}_{h+1}) \|_{\Lambda_h^{-1}}^2. \tag{54}$$

Noted that the samples in $\mathcal{D}^0$ are temporally statistically independent, i.e., $\widehat{V}_{h+1}$ is independent of $\mathcal{D}_h^0$, or to say, $\widehat{\mu}_h$. Therefore, we can directly control $T_{2,h}$ via the following lemma.

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

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

The proof of Lemma 12 is postponed to Appendix B.3.6. Then applying Lemma 12 and the union bound over $\mathcal{N}(\epsilon_0, H)$, we have

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

for any fixed $h \in [H]$. According to Vershynin (2018), one has $|\mathcal{N}(\epsilon_0, H)| \leq \frac{3H}{\epsilon_0}$. Taking the union bound for any $h \in [H]$, we arrive at

$$\sup_{\alpha \in \mathcal{N}(\epsilon_0, H)} \| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, \widehat{V}_{h+1})\|_{\Lambda_h^{-1}}^2 \leq 2H^2 \log(\frac{3H^2}{\epsilon_0 \delta}) + H^2 d\log(1 + \frac{K}{\lambda_0}), \tag{55}$$

with probability exceeding $1 - \delta$, where we utilize $N_h \leq K$ for every $h \in [H]$ on the right-hand side. Combining (53), (54) and (55), we have

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

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

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

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

$$T_{1,h}^2 \leq 8H^2 + 4H^2 \xi_0 + 2dH^2 \xi_0 \leq 16dH^2 \xi_0.$$

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

$$|(\widehat{\mathbb{B}}_h^\rho \widehat{V}_{h+1})(s,a) - (\mathbb{B}_h^\rho \widehat{V}_{h+1})(s,a)| \leq \left(\sqrt{d}H + 4\sqrt{d\xi_0}H\right) \sum_{i=1}^d \|\phi_i(s,a)\|_{\Lambda_h^{-1}}$$

$$\leq \gamma_0 \sum_{i=1}^d \|\phi_i(s_h, a_h)\|_{\Lambda_h^{-1}} = \Gamma_h(s,a),$$

where $\gamma_0 = 6\sqrt{d\xi_0}H$ and the above inequality satisfies (31).

B.3.4 Proof of Lemma 11

It follows Lemma 1 and (16) that

$$|(\widehat{\mathbb{B}}_h^\rho \widehat{V}_{h+1})(s,a) - (\mathbb{B}_h^\rho \widehat{V}_{h+1})(s,a)| = |\phi(s,a)^\top (\widehat{w}_h^{\rho, \widehat{V}} - w_h^{\rho, \widehat{V}})|$$
$$= \underbrace{|\phi(s,a)^\top \left(\widehat{\theta}_h - \theta_h\right)|}_{\text{(i)}} + \underbrace{|\phi(s,a)^\top \left(\widehat{\nu}_h^{\rho, \widehat{V}} - \nu_h^{\rho, \widehat{V}}\right)|}_{\text{(ii)}}, \qquad \forall (s,a,h) \times \mathcal{S} \times \mathcal{A} \times [H]. \tag{56}$$

We first bound the term (i), for any $h \in [H]$. By the update (12), we have

$$\text{(i)} = |\phi(s,a)^\top \Lambda_h^{-1} \left(\sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) r_h^\tau\right) - \phi(s,a)^\top \theta_h|$$
$$= |\phi(s,a)^\top \Lambda_h^{-1} (\Lambda_h - \lambda_0 I) \theta_h - \phi(s,a)^\top \theta_h|$$
$$= |\lambda_0 \phi(s,a)^\top \Lambda_h^{-1} \theta_h|,$$

where the second equality is from Assumption 1 and (15). Applying the Cauchy-Schwarz inequality leads to

$$\texttt{(i)} \leq \lambda_0 \left\| \phi(s,a) \right\|_{\Lambda_h^{-1}} \left\| \theta_h \right\|_{\Lambda_h^{-1}} \leq \sqrt{d\lambda_0} \sum_{i=1}^{d} \left\| \phi_i(s,a) \mathbb{1}_i \right\|_{\Lambda_h^{-1}}, \tag{57}$$

where the last inequality is $\|\theta_h\| \leq \sqrt{d}$ and $\|\Lambda_h^{-1}\| \leq 1/\lambda_0$ such that

$$\|\theta_h\|_{\Lambda_h^{-1}} \leq \|\Lambda_h^{-1}\|^{1/2}\|\theta_h\| \leq \sqrt{d/\lambda_0}, \qquad \forall h \in [H].$$

Next, to bound the term (ii), we define the following notations for simplicity. Let $\epsilon_h^\tau(\alpha, V) = \int_{\mathcal{S}} P_h^0(s'|s_h^\tau, a_h^\tau)[V]_\alpha(s')\mathrm{d}s' - [V]_\alpha(s_{h+1}^\tau)$ for any $V : \mathcal{S} \to [0, H]$ and $\alpha \in [\min_s V(s), \max_s V(s)]$. Also, we define two auxiliary functions:

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

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

With the new notations in hand, we can bound (ii) by

$$\texttt{(ii)} = \left| \sum_{i=1}^{d} \phi_i(s,a) \left(\widehat{\nu}_{h,i}^{\rho,\widehat{V}} - \nu_{h,i}^{\rho,\widehat{V}} \right) \right|$$

$$\leq \sum_{i=1}^{d} \phi_i(s,a) \max_{\alpha \in [\min_s \widehat{V}_{h+1}(s), \max_s \widehat{V}_{h+1}(s)]} \left| \widehat{g}_{h,i}(\alpha) - g_{h,i}^0(\alpha) \right|$$

$$\leq \sum_{i=1}^{d} \phi_i(s,a) \max_{\alpha \in [\min_s \widehat{V}_{h+1}(s), \max_s \widehat{V}_{h+1}(s)]} \left| \int_{\mathcal{S}} \left(\widehat{\mu}_{h,i}(s') - \mu_{h,i}^0(s') \right) [\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' \right|$$

$$\leq \sum_{i=1}^{d} \max_{\alpha \in [\min_s \widehat{V}_{h+1}(s), \max_s \widehat{V}_{h+1}(s)]} \left| \phi_i(s,a) \int_{\mathcal{S}} \left(\widehat{\mu}_{h,i}(s') - \mu_{h,i}^0(s') \right) [\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' \right|, \tag{58}$$

where the first inequality is due to (11), (13) as well as Lemma 6, 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,

$$\left| \int_{\mathcal{S}} \mu_{h,i}^0(s')[\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' - \int_{\mathcal{S}} \widehat{\mu}_{h,i}(s')[\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' \right|$$

$$= \left| \int_{\mathcal{S}} \mu_{h,i}^0(s')[\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' - \mathbb{1}_i^\top \Lambda_h^{-1} \left(\sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)[\widehat{V}_{h+1}]_\alpha(s_{h+1}^\tau) \right) \right|$$

$$= \left| \mathbb{1}_i^\top \Lambda_h^{-1} \left(\Lambda_h \int_{\mathcal{S}} \mu_h^0(s')[\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' - \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)[\widehat{V}_{h+1}]_\alpha(s_{h+1}^\tau) \right) \right|$$

$$= \left| \mathbb{1}_i^\top \Lambda_h^{-1} \left[\lambda_0 \int_{\mathcal{S}} \mu_h^0(s')[\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' + \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \left(\int_{\mathcal{S}} P_h^0(s'|s_h^\tau, a_h^\tau)[\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' - [\widehat{V}_{h+1}]_\alpha(s_{h+1}^\tau) \right) \right] \right|$$

$$= \left| \mathbb{1}_i^\top \Lambda_h^{-1} \left[\lambda_0 \int_{\mathcal{S}} \mu_h^0(s')[\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' + \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, \widehat{V}_{h+1}) \right] \right| \tag{59}$$

where the first equality is from (36), the third one is due to (15) and we let

$$\epsilon_h^\tau(\alpha, V) = \int_{\mathcal{S}} P_h^0(s'|s_h^\tau, a_h^\tau)[V]_\alpha(s')\mathrm{d}s' - [V]_\alpha(s_{h+1}^\tau),$$

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

$$
\left| \phi_i(s, a) \cdot (\widehat{\mu}_{h,i} - \mu_{h,i}^0)[\widehat{V}_{h+1}]_\alpha \right|
$$

$$
\leq \left| \phi_i(s, a) \mathbb{1}_i^\top \Lambda_h^{-1} \left(\lambda_0 \int_\mathcal{S} \mu_h^0(s')[\widehat{V}_{h+1}]_\alpha(s') \mathrm{d}s' + \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, \widehat{V}_{h+1}) \right) \right|
$$

$$
\leq \|\phi_i(s, a) \mathbb{1}_i\|_{\Lambda_h^{-1}} \left(\lambda_0 \underbrace{\left\| \int_\mathcal{S} \mu_h^0(s')[\widehat{V}_{h+1}]_\alpha(s') \mathrm{d}s' \right\|_{\Lambda_h^{-1}}}_{\texttt{(iii)}} + \left\| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, \widehat{V}_{h+1}) \right\|_{\Lambda_h^{-1}} \right), \quad (60)
$$

where the last inequality holds due to the Cauchy-Schwarz inequality.

Moreover, the term (iii) in (60) can be further simplified to

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

since $|\widehat{V}_{h+1}(s)| \leq H$ for any $s \in \mathcal{S}$ and $\|\Lambda_h^{-1}\| \leq 1/\lambda_0$. Then we have

$$
\texttt{(ii)} \leq \left(\sqrt{\lambda_0} H + \max_{\alpha \in [\min_s \widehat{V}_{h+1}(s), \max_s \widehat{V}_{h+1}(s)]} \left\| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\alpha, \widehat{V}_{h+1}) \right\|_{\Lambda_h^{-1}} \right) \sum_{i=1}^d \|\phi_i(s, a) \mathbb{1}_i\|_{\Lambda_h^{-1}},
$$

$$
(61)
$$

for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Combining (61) with (57) finally leads to (51), which completes our proof.

B.3.5 Proof of (53)

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

$$
|\epsilon_h^\tau(\alpha, V) - \epsilon_h^\tau(\alpha^\dagger, V)| \leq 2|\alpha - \alpha^\dagger| \leq 2\epsilon_0.
$$

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

$$
\left\| \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau) \left(\epsilon_h^\tau(\alpha, V) - \epsilon_h^\tau(\alpha^\dagger, V) \right) \right\|_{\Lambda_h^{-1}}^2
$$

$$
= \sum_{\tau, \tau' \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^{\tau'}, a_h^{\tau'}) \left[\left(\epsilon_h^\tau(\alpha, V) - \epsilon_h^\tau(\alpha^\dagger, V) \right) \left(\epsilon_h^{\tau'}(\alpha, V) - \epsilon_h^{\tau'}(\alpha^\dagger, V) \right) \right]
$$

$$
\leq \sum_{\tau, \tau' \in \mathcal{D}_h^0} \phi(s_h, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^{\tau'}, a_h^{\tau'}) \cdot 4\epsilon_0^2
$$

$$
\leq 4\epsilon_0^2 N_h^2 / \lambda_0,
$$

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

$$
\sum_{\tau, \tau' \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^{\tau'}, a_h^{\tau'}) = \sum_{\tau, \tau' \in \mathcal{D}_h^0} \|\phi(s_h^\tau, a_h^\tau)\|_2 \cdot \|\phi(s_h^{\tau'}, a_h^{\tau'})\|_2 \cdot \|\Lambda_h^{-1}\| \leq N_h^2 / \lambda_0. \quad (62)
$$

Thus,

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

due to the fact $N_h \leq K$ for any $h \in [H]$, which completes the proof of (53).

B.3.6 Proof of Lemma 12

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

$$\mathcal{F}_{h,\tau} = \sigma(\{(s_h^j, a_h^j)\}_{j=1}^{(\tau+1)\wedge|N_h|}, \{r_h^j, s_{h+1}^j\}_{j=1}^\tau).$$

As shown in Jin et al. (2021, 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 $\epsilon_h^\tau(\alpha, V)$ is $\mathcal{F}_{h,\tau-1}$-measurable. Hence $\{\epsilon_h^\tau(\alpha, V)\}_{\tau\in\mathcal{D}_h^0}$ is stochastic process adapted to the filtration $\{\mathcal{F}_{h,\tau}\}_{\tau\in\mathcal{D}_h^0}$. Then, we have

$$\mathbb{E}_{\mathcal{D}_h^0}[\epsilon_h^\tau(\alpha, V)|\mathcal{F}] = \int_{\mathcal{S}} P_h^0(s'|s_h^\tau, a_h^\tau)[V]_\alpha - \mathbb{E}_{\mathcal{D}_h^0}\left[[V(s_{h+1}^\tau)]_\alpha|\{(s_h^j, a_h^j)\}_{j=1}^{(\tau)\wedge N_h}, \{r_h^j, s_{h+1}^j\}_{j=1}^{\tau-1}\right]$$

$$= \int_{\mathcal{S}} P_h^0(s'|s_h^\tau, a_h^\tau)[V]_\alpha - \mathbb{E}_{\mathcal{D}_h^0}\left[[V(s_{h+1}^\tau)]_\alpha\right] = 0.$$

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

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

Hence, for the fixed $h \in [H]$ and all $\tau \in [H]$, the random variable $\epsilon_h^\tau(\alpha, V)$ is mean-zero and H-sub-Gaussian conditioning on $\mathcal{F}_{h,\tau-1}$. Then, we invoke the Lemma 2 with $\eta_\tau = \epsilon_h^\tau(\alpha, V)$ and $x_\tau = \phi(s_h^\tau, a_h^\tau)$. For any $\delta > 0$, we have

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

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 12.

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

$$\Phi_{h,i}^\star(s) = (\phi_i(s, \pi_h^\star(s))\mathbb{1}_i)(\phi_i(s, \pi_h^\star(s))\mathbb{1}_i)^\top \in \mathbb{R}^{d\times d}, \tag{63}$$

$$b_{h,i}^\star(s) = (\phi_i(s, \pi_h^\star(s))\mathbb{1}_i)^\top \Lambda_h^{-1}(\phi_i(s, \pi_h^\star(s))\mathbb{1}_i). \tag{64}$$

With these notations in hand and recalling (17) in Theorem 1, one has

$$V_1^{\star,\rho}(\zeta) - V_1^{\widehat{\pi},\rho}(\zeta) \leq 2\gamma_0 \sum_{h=1}^H \sum_{i=1}^d \sup_{d_h^\star\in\mathcal{D}_h^\star} \mathbb{E}_{s\sim d_h^\star} \sqrt{b_{h,i}^\star(s)}$$

$$\leq 2\gamma_0 \sum_{h=1}^H \sum_{i=1}^d \sup_{d_h^\star\in\mathcal{D}_h^\star} \sqrt{\mathbb{E}_{s\sim d_h^\star} b_{h,i}^\star(s)}$$

$$= 2\gamma_0 \sum_{h=1}^H \sup_{d_h^\star\in\mathcal{D}_h^\star} \sum_{i=1}^d \sqrt{\mathbb{E}_{s\sim d_h^\star} b_{h,i}^\star(s)}, \tag{65}$$

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. $d_h^b(s, a) > 0$ and from Lemma 9, the following lower bound holds with probability at least $1 - 3\delta$, i.e.,

$$N_h(s, a) \geq \frac{K d_h^b(s, a)}{8} - 5\sqrt{K d_h^b(s, a) \log(\frac{KH}{\delta})} \geq \frac{K d_h^b(s, a)}{16}, \tag{66}$$

as long as

$$K \geq c_0 \frac{\log(KH/\delta)}{d_{\min}^b} \geq c_0 \frac{\log(KH/\delta)}{d_h^b(s, a)} \tag{67}$$

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

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

From Assumption 3,

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

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

$$\Lambda_h \succeq I_d + \frac{K d \cdot \min\{\mathbb{E}_{d_h^\star} \phi_i^2(s, \pi_h^\star(s)), 1/d\}}{16 C_{\mathrm{rob}}^\star} \cdot \mathbb{1}_{i,i}. \tag{68}$$

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

$$\begin{aligned}
\mathbb{E}_{s \sim d_h^\star} b_{h,i}^\star(s) = \mathbb{E}_{s \sim d_h^\star} \mathrm{Tr}(\Phi_{h,i}^\star(s) \Lambda_h^{-1}) &= \mathrm{Tr}(\mathbb{E}_{s \sim d_h^\star} \Phi_{h,i}^\star(s) \Lambda_h^{-1}) \\
&\leq \frac{\mathbb{E}_{d_h^\star} \phi_i^2(s, \pi_h^\star(s))}{1 + K d \cdot \min\{\mathbb{E}_{d_h^\star} \phi_i^2(s, \pi_h^\star(s)), 1/d\}/16 C_{\mathrm{rob}}^\star},
\end{aligned} \tag{69}$$

where the second equality is because the trace is a linear mapping and the last inequality holds by Lemma 7. 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

$$|\mathcal{E}_{h,\mathsf{larger}}| \leq \sqrt{d}, \tag{70}$$

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 (69) can be controlled by $\langle d_h^\star, b_{h,i}^\star \rangle \leq 0$.

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

$$(69) \leq \frac{16 C_{\mathrm{rob}}^\star \cdot \mathbb{E}_{d_h^\star} \phi_i^2(s, \pi_h^\star(s))}{K d \cdot \mathbb{E}_{d_h^\star} \phi_i^2(s, \pi_h^\star(s))} = \frac{16 C_{\mathrm{rob}}^\star}{K d}. \tag{71}$$

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

$$(69) \leq \frac{16 C_{\mathrm{rob}}^\star \cdot \mathbb{E}_{d_h^\star} \phi_i^2(s, \pi_h^\star(s))}{K} \leq \frac{16 C_{\mathrm{rob}}^\star}{K}, \tag{72}$$

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

Summing up the above three cases and (70), we have

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

Together with (65) and setting $\gamma_0 = 6\sqrt{d}H\sqrt{\log(3HK/\delta)}$, one obtains

$$
\begin{aligned}
V_1^{\star,\rho}(\zeta) - V_1^{\widehat{\pi},\rho}(\zeta) &\leq 2\gamma_0 \sum_{h=1}^{H} \sup_{d_h^\star\in\mathcal{D}_h^\star} \sum_{i=1}^{d} \sqrt{\mathbb{E}_{s\sim d_h^\star} b_{h,i}^\star(s)} \\
&\leq 96dH^2 \sqrt{C_{\text{rob}}^\star/K} \sqrt{\log(3HK/\delta)},
\end{aligned}
$$

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

Proof of (70). Let $\tilde{\mathcal{E}}_{h,\text{larger}} = \{i : \mathbb{E}_{(s,a)\sim d_h^\star} \phi_i(s,a) \geq \frac{1}{\sqrt{d}}\}$.

- We first show that $|\tilde{\mathcal{E}}_{h,\text{larger}}|$ should be no larger than $\sqrt{d}$ by contradiction. Suppose $|\tilde{\mathcal{E}}_{h,\text{larger}}| > \sqrt{d}$. 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,

$$
\sum_{i\in\tilde{\mathcal{E}}_{h,\text{larger}}} \mathbb{E}_{(s,a)\sim d_h^\star} \phi_i(s,a) > 1, \tag{73}
$$

  which is equivalent to

$$
\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|_1 \geq \mathbb{E}_{(s,a)\sim d_h^\star} \|\phi(s,a)\|_1 \geq \mathbb{E}_{(s,a)\sim d_h^\star} \sum_{i\in\tilde{\mathcal{E}}_{h,\text{larger}}} \phi_i(s,a) > 1, \tag{74}
$$

  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 $\tilde{\mathcal{E}}_{h,\text{larger}} \subseteq \mathcal{E}_{h,\text{larger}}$: For every element $i \in \tilde{\mathcal{E}}_{h,\text{larger}}$, we have

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

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

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

B.5 Proof of Corollary 2

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

Lemma 13. *Consider $\delta \in (0,1)$. Suppose Assumption 2, Assumption 4 and all conditions in Lemma 4 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 4 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,

$$1 = \int_{\mathcal{S}} P_h^0(s'|s,a)\mathrm{d}s' = \int_{\mathcal{S}} \phi(s,a)^\top \mu_h^0(s')\mathrm{d}s' = \sum_{i=1}^d \phi_i(s,a) \int_{\mathcal{S}} \mu_{h,i}^0(s')\mathrm{d}s' = \sum_{i=1}^d \phi_i(s,a), \qquad (75)$$

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

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

$\square$

From (66), we have $N_h \geq \frac{K}{16}$ with probability exceeding $1 - 3\delta$, as long as K obeys (67). Together will Lemma 13, 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

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

which completes the proof.

C Analysis for `DROP-V`: Algorithm 3

C.1 The implementation of `DROP-V`

The implementation of `DROP-V` is detailed in Algorithm 3, which can be divided into three steps. First, we carefully design `Three-fold-subsampling` (cf. Algorithm 4), to generate two almost temporally statistically independent datasets, $\tilde{\mathcal{D}}^0, \mathcal{D}^0$, which are also independent from each other. The theoretical analysis of `Three-fold-subsampling` is postponed to Appendix C.2. The second step is to construct a variance estimator $\widehat{\sigma}_h^2$ for any $h \in [H]$ via $\tilde{\mathcal{D}}^0$, which is independent of $\mathcal{D}_h^0$. The key idea is to utilize the intermediate results $\{\tilde{V}_h\}_{h=1}^{H+1}$ of running `DROP` on $\tilde{\mathcal{D}}^0$ to approximate the variance as (22). With the variance estimator at our hands, the last step is to apply the weighted ridge regression to construct the empirical variance-aware robust Bellman operator via (25)-(28), which is slightly different from `DROP`.

C.2 Theoretical guarantee for `Three-fold-subsampling`

As the three-fold subsampling method presented in Appendix C.1, it is slightly different from the two-fold variant. Thus, we establish the following lemma to show that (76) is a valid high-probability lower bound of $N^{\mathrm{main}}(s)$ for any $s \in \mathcal{S}$ and $h \in [H]$, which follows the proof of Lemma 3 in Li et al. (2024).

Lemma 14. *Consider $\delta \in (0,1)$. With probability at least $1 - 3\delta$, if $N_h^{\mathrm{trim}}(s)$ satisfies (76) for every $s \in \mathcal{S}$ and $h \in [H]$, then the following bounds hold, i.e.,*

$$N_h^{\mathrm{trim}}(s) \leq N_h^{\mathrm{main}}(s), \qquad N_h^{\mathrm{trim}}(s) \leq N_h^{\mathrm{var}}(s), \qquad \forall(s,h) \in \mathcal{S} \times [H]. \qquad (77)$$

Algorithm 3 Distributionally Robust Pessimistic Least-Squares Value Iteration with Variance Estimation (`DROP-V`)

Input: Datasets $\tilde{\mathcal{D}}^0, \mathcal{D}^0 \leftarrow$ `Three-fold-subsampling`$(\mathcal{D})$; feature map $\phi(s,a)$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$; $\gamma_1, \lambda_1 > 0$.

Construct the variance estimator: Obtain $(\tilde{V}, \tilde{\pi}) \leftarrow$ `DROP`$(\tilde{\mathcal{D}}^0, \phi)$

1: For every $h \in [H]$, compute $\tilde{\Lambda}_h = \sum_{\tau \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + I_d$ and

$$\nu_{h,1} = (\tilde{\Lambda}_h)^{-1}\left(\sum_{\tau \in \tilde{\mathcal{D}}_h^0} \phi(s_h^\tau, a_h^\tau)\tilde{V}_{h+1}^2(s_{h+1}^\tau)\right), \quad \nu_{h,2} = (\tilde{\Lambda}_h)^{-1}\left(\sum_{\tau \in \tilde{\mathcal{D}}_h^0} \phi(s_h^\tau, a_h^\tau)\tilde{V}_{h+1}(s_{h+1}^\tau)\right).$$

2: Update $\widehat{\sigma}_h^2(s,a)$ via (22), for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Initialization: Set $\widehat{Q}_{H+1}(\cdot,\cdot) = 0$ and $\widehat{V}_{H+1}(\cdot) = 0$.

3: **for** step $h = H, H-1, \cdots, 1$ **do**

4: $\quad \Sigma_h = \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} + \lambda_1 I_d.$

5: $\quad \widehat{\theta}_h^\sigma = \Sigma_h^{-1}\left(\sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)r_h^\tau}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)}\right).$

6: $\quad$ **for** feature $i = 1, \cdots, d$ **do**

7: $\quad\quad$ Update $\widehat{\nu}_{h,i}^{\rho,\sigma,\widehat{V}}$ via (28).

8: $\quad$ **end for**

9: $\quad \widehat{w}_h^{\rho,\sigma,\widehat{V}} = \widehat{\theta}_h + \widehat{\nu}_h^{\rho,\sigma,\widehat{V}}.$

10: $\quad \bar{Q}_h(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{w}_h^{\rho,\sigma,\widehat{V}} - \gamma_1 \sum_{i=1}^d \|\phi_i(\cdot,\cdot)\mathbb{1}_i\|_{\Sigma_h^{-1}}.$

11: $\quad \widehat{Q}_h(\cdot,\cdot) = \min\left\{\bar{Q}_h, H-h+1\right\}_+.$

12: $\quad \widehat{\pi}_h(\cdot) = \arg\max_{a \in \mathcal{A}} \widehat{Q}_h(\cdot,a).$

13: $\quad \widehat{V}_h(\cdot) = \widehat{Q}_h(\cdot, \widehat{\pi}_h(\cdot)).$

14: **end for**

Ouput: $\widehat{\pi} := \{\widehat{\pi}_h\}_{h=1}^H$

Algorithm 4 `Three-fold-subsampling`

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

1: **Split Data:** Split $\mathcal{D}$ into three haves $\mathcal{D}^{\text{aux}}$, $\mathcal{D}^{\text{main}}$ and $\mathcal{D}^{\text{var}}$, where $|\mathcal{D}^{\text{aux}}| = |\mathcal{D}^{\text{main}}| = |\mathcal{D}^{\text{var}}| = K/3$. Denote $N_h^{\text{main}}(s)$ (resp. $N_h^{\text{aux}}(s)$ or $N_h^{\text{var}}(s)$) as the number of sample transitions from state s at step h in $\mathcal{D}^{\text{main}}$ (resp. $\mathcal{D}^{\text{aux}}$ or $\mathcal{D}^{\text{var}}$).

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

$$N_h^{\text{trim}}(s) = \max\{N_h^{\text{aux}}(s) - 6\sqrt{N_h^{\text{aux}}(s)\log\frac{KH}{\delta}}, 0\}. \tag{76}$$

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

Ouput: $\mathcal{D}^{\text{main,sub}}, \mathcal{D}^{\text{var,sub}}$.

In addition, with probability at least $1 - 4\delta$, the following bound also holds:

$$N_h^{\text{trim}}(s,a) \geq \frac{Kd_h^b(s,a)}{12} - \sqrt{6Kd_h^{\text{b}}(s,a)\log\frac{KH}{\delta}}, \forall(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{78}$$

Proof. We begin with proving the first claim (77). Let $\mathcal{S}_{\mathcal{D}^{\mathrm{aux}}} \subset \mathcal{S}$ be the collection of all the states appearing for the dataset $\mathcal{D}^{\mathrm{aux}}$, where $|\mathcal{S}_{\mathcal{D}^{\mathrm{aux}}}| \leq K/3$. Without loss of generality, we assume that $\mathcal{D}^{\mathrm{aux}}$ contains the first $K/3$ trajectories and satisfies

$$N_h^{\mathrm{aux}}(s) = \sum_{k=1}^{K/3} \mathbb{1}(s_h^k = s), \quad \forall (s,h) \in \mathcal{S} \times [H],$$

which can be viewed as the sum of $K/3$ independent Bernoulli random variables. By the union bound and the Bernstein inequality,

$$P\left(\exists (s,h) \in \mathcal{S}_{\mathcal{D}^{\mathrm{aux}}} \times [H] : \left|N_h^{\mathrm{aux}}(s) - \frac{K}{3} d_h^b(s)\right| \geq t\right) \leq \sum_{s \in \mathcal{S}_{\mathcal{D}^{\mathrm{aux}}}, h \in [H]} P\left(\left|N_h^{\mathrm{aux}}(s) - \frac{K}{3} d_h^b(s)\right| \geq t\right)$$

$$\leq \frac{2KH}{3} \exp\left(-\frac{t^2/2}{v_{s,h} + t/3}\right),$$

for any $t \geq 0$, where

$$v_{s,h} = \frac{K}{3} \mathrm{Var}[\mathbb{1}(s_h^k = s)] \leq \frac{K d_h^{\mathrm{b}}(s)}{3}.$$

Here, we abuse the notation Var to represent the variance of the Bernoulli distributed $\mathbb{1}(s_h^k = s)$. Then, with probability at least $1 - 2\delta/3$, we have

$$\left|N_h^{\mathrm{aux}}(s) - \frac{K}{3} d_h^b(s)\right| \leq \sqrt{2 v_{s,h} \log(\frac{KH}{\delta})} + \frac{2}{3} \log(\frac{KH}{\delta})$$

$$\leq \sqrt{K d_h^{\mathrm{b}}(s) \log(\frac{KH}{\delta})} + \log(\frac{KH}{\delta}), \ \forall (s,h) \in \mathcal{S} \times [H]. \tag{79}$$

Similarly, with probability at least $1 - 2\delta/3$, we have

$$\left|N_h^{\mathrm{main}}(s) - \frac{K}{3} d_h^b(s)\right| \leq \sqrt{K d_h^{\mathrm{b}}(s) \log(\frac{KH}{\delta})} + \log(\frac{KH}{\delta}), \quad \forall (s,h) \in \mathcal{S} \times [H]. \tag{80}$$

Therefore, combining (79) and (80) leads to

$$\left|N_h^{\mathrm{main}}(s) - N_h^{\mathrm{aux}}(s)\right| \leq 2\sqrt{K d_h^{\mathrm{b}}(s) \log(\frac{KH}{\delta})} + 2 \log(\frac{KH}{\delta}), \quad \forall (s,h) \in \mathcal{S} \times [H], \tag{81}$$

with probability at least $1 - 4\delta/3$. Then, we consider the following two cases

- Case 1: $N_h^{\mathrm{aux}}(s) \leq 36 \log \frac{KH}{\delta}$. One has

$$N_h^{\mathrm{trim}}(s) = \max\{N_h^{\mathrm{aux}}(s) - 6\sqrt{N_h^{\mathrm{aux}}(s) \log \frac{HK}{\delta}}, 0\} = 0 \leq N_h^{\mathrm{main}}(s).$$

- Case 2: $N_h^{\mathrm{aux}}(s) > 36 \log \frac{KH}{\delta}$. From (79), we have

$$\frac{K}{3} d_h^b(s) + \sqrt{K d_h^{\mathrm{b}}(s) \log(\frac{KH}{\delta})} + \log(\frac{KH}{\delta}) \geq N_h^{\mathrm{aux}}(s) \geq 36 \log \frac{KH}{\delta},$$

implying

$$K d_h^{\mathrm{b}}(s) \geq 72 \log \frac{KH}{\delta}.$$

Also from (79),

$$N_h^{\mathrm{aux}}(s) \geq \frac{K}{3} d_h^b(s) - \sqrt{K d_h^{\mathrm{b}}(s) \log(\frac{KH}{\delta})} - \log(\frac{KH}{\delta}) \geq \frac{K}{6} d_h^b(s).$$

Therefore, with probability exceeding $1 - 4\delta/3$

$$
\begin{aligned}
N_h^{\mathrm{trim}}(s) = N_h^{\mathrm{aux}}(s) - 6\sqrt{N_h^{\mathrm{aux}}(s) \log \frac{HK}{\delta}} &\le N_h^{\mathrm{aux}}(s) - \sqrt{6}\sqrt{Kd_h^{\mathrm{b}}(s) \log \frac{HK}{\delta}} \\
&\le N_h^{\mathrm{aux}}(s) - 2\sqrt{Kd_h^{\mathrm{b}}(s) \log \frac{HK}{\delta}} - \frac{1}{3}\sqrt{Kd_h^{\mathrm{b}}(s) \log \frac{HK}{\delta}} \\
&\le N_h^{\mathrm{aux}}(s) - 2\sqrt{Kd_h^{\mathrm{b}}(s) \log \frac{HK}{\delta}} - 2\log \frac{HK}{\delta} \\
&\le N_h^{\mathrm{main}}(s),
\end{aligned}
$$

where the last inequality is from (81).

Following the same arguments, we also have

$$
|N_h^{\mathrm{var}}(s) - N_h^{\mathrm{aux}}(s)| \le 2\sqrt{Kd_h^{\mathrm{b}}(s) \log(\frac{KH}{\delta})} + 2\log(\frac{KH}{\delta}), \quad \forall (s,h) \in \mathcal{S} \times [H]. \tag{82}
$$

holds, with probability at least $1 - 4\delta/3$. Therefore, we can also guarantee that $N_h^{\mathrm{trim}}(s) \le N_h^{\mathrm{var}}(s)$ with probability at least $1 - 4\delta/3$, for any $(s,h) \in \mathcal{S} \times [H]$.

Putting these two results together, we prove the first claim (77).

Next, we will establish the second claim (78). To begin with, we claim the following statement holds with probability exceeding $1 - 2\delta/3$,

$$
N_h^{\mathrm{trim}}(s,a) \ge N_h^{\mathrm{trim}}(s)\pi_h^{\mathrm{b}}(a|s) - \sqrt{2N_h^{\mathrm{trim}}(s)\pi_h^{\mathrm{b}}(a|s) \log(\frac{KH}{\delta})} - \log \frac{KH}{\delta}, \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{83}
$$

conditioned on the high-probability event that the first part (77) holds. In the sequel, we discuss the following two cases, provided that the inequality (83) holds.

- Case 1: $Kd_h^{\mathrm{b}}(s,a) = Kd_h^{\mathrm{b}}(s)\pi_h^{\mathrm{b}}(a|s) > 864 \log \frac{KH}{\delta}$. From (79), with probability exceeding $1 - 2\delta/3$, one has

$$
N_h^{\mathrm{aux}}(s) \ge \frac{K}{3}d_h^b(s) - \sqrt{Kd_h^{\mathrm{b}}(s) \log(\frac{KH}{\delta})} - \log(\frac{KH}{\delta}) \ge \frac{K}{6}d_h^b(s) \ge 144 \log \frac{KH}{\delta}.
$$

Together with the definition (76), we have

$$
N_h^{\mathrm{trim}}(s) \ge N_h^{\mathrm{aux}}(s) - 6\sqrt{N_h^{\mathrm{aux}}(s) \log \frac{KH}{\delta}} \ge \frac{1}{2}N_h^{\mathrm{aux}}(s) \ge \frac{K}{12}d_h^b(s).
$$

Therefore,

$$
N_h^{\mathrm{trim}}(s)\pi_h^{\mathrm{b}}(a|s) \ge \frac{K}{12}d_h^b(s)\pi_h^{\mathrm{b}}(a|s) \ge 72 \log \frac{KH}{\delta}.
$$

Combining with (83), one can derive

$$
\begin{aligned}
N_h^{\mathrm{trim}}(s,a) &\ge \frac{Kd_h^b(s,a)}{12} - \sqrt{\frac{1}{6}Kd_h^b(s,a) \log(\frac{KH}{\delta})} - \log \frac{KH}{\delta} \\
&\ge \frac{Kd_h^b(s,a)}{12} - \sqrt{6Kd_h^{\mathrm{b}}(s,a) \log \frac{KH}{\delta}}.
\end{aligned}
$$

with probability exceeding $1 - 4\delta/3$

- Case 2: $Kd_h^{\mathrm{b}}(s,a) \le 864 \log \frac{KH}{\delta}$. From (76), one has

$$
N_h^{\mathrm{trim}}(s,a) \ge 0 \ge \frac{Kd_h^b(s,a)}{12} - \sqrt{6Kd_h^{\mathrm{b}}(s,a) \log \frac{KH}{\delta}}.
$$

By integrating these two cases, we can claim (78) is valid with probability exceeding $1 - 4\delta/3$, as long as the inequality (83) holds under the condition of the high-probability event described in the the first part (77). Thus, the second claim (78) holds with probability at least $1 - 4\delta$.

Proof of inequality (83). First, we can observe that the inequality (83) holds if $N_h^{\text{trim}}(s)\pi_h^{\text{b}}(s,a) \leq 2\log\frac{KH}{\delta}$. Thus, we focus on the other case that $N_h^{\text{trim}}(s)\pi_h^{\text{b}}(s,a) > 2\log\frac{KH}{\delta}$. Denote that

$$\mathcal{E} = \{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H] | \; N_h^{\text{trim}}(s)\pi_h^{\text{b}}(a|s) > 2\log(\frac{KH}{\delta})\}.$$

Noticed that from Algorithm 4, one has that $|\mathcal{E}| \leq \frac{KH}{3}$. Supposing that the first claim (77) holds, one has $N_h^{\text{trim}}(s) = \min\{N_h^{\text{trim}}(s), N_h^{\text{main}}(s), N_h^{\text{var}}(s)\}$. Therefore, $N_h^{\text{trim}}(s,a)$ can be viewed as the sum of $N_h^{\text{trim}}(s)$ independent Bernoulli random variables, where each is with the mean $\pi_h^{\text{b}}(a|s)$. Then, by the union bound and the Bernstein inequality,

$$P\left(\exists (s,a,h) \in \mathcal{E} : \left|N_h^{\text{trim}}(s,a) - N_h^{\text{trim}}(s)\pi_h^{\text{b}}(a|s)\right| \geq t\right)$$
$$\leq \sum_{(s,a,h) \in \mathcal{E}} P\left(\left|N_h^{\text{trim}}(s,a) - N_h^{\text{trim}}(s)\pi_h^{\text{b}}(a|s)\right| \geq t\right) \leq \frac{2KH}{3}\exp\left(-\frac{t^2/2}{v_{s,h} + t/3}\right),$$

for any $t \geq 0$, where

$$v_{s,h} = N_h^{\text{trim}}(s)\text{Var}[\mathbb{1}((s_h^k, a_h^k) = (s,a))] \leq N_h^{\text{trim}}(s)\pi_h^{\text{b}}(a|s)$$

A little algebra yields that with probability at least $1 - 2\delta/3$, one can obtain

$$\left|N_h^{\text{trim}}(s,a) - N_h^{\text{trim}}(s)\pi_h^{\text{b}}(a|s)\right| \leq \sqrt{2v_{s,h}\log(\frac{KH}{\delta})} + \frac{2}{3}\log(\frac{KH}{\delta})$$
$$\leq \sqrt{2N_h^{\text{trim}}(s)\pi_h^{\text{b}}(a|s)\log(\frac{KH}{\delta})} + \log(\frac{KH}{\delta}), \; \forall (s,h) \in \mathcal{S} \times [H]. \tag{84}$$

Therefore, with probability $1 - 2\delta/3$, one can obtain

$$N_h^{\text{trim}}(s,a) \geq N_h^{\text{trim}}(s)\pi_h^{\text{b}}(a|s) - \sqrt{2N_h^{\text{trim}}(s)\pi_h^{\text{b}}(a|s)\log(\frac{KH}{\delta})} - \log(\frac{KH}{\delta}),$$

for any $(s,a,h) \in \mathcal{E}$, conditioned on the first claim (77) holds. □

In addition, the following lemma guarantees that the samples in $\mathcal{D}^{\text{main,sub}}$ and $\mathcal{D}^{\text{var,sub}}$ are statistically independent with probability exceeding $1 - 3\delta$. Before continuing, we denote $\mathcal{D}^{\text{i.i.d.}}$ as the dataset containing $N_h^{\text{trim}}(s)$ independent transition-reward sample tuples for every $(s,h) \in \mathcal{S} \times [H]$, following π_h^{b} and P_h^0.

Lemma 15 (Modified Lemma 7, Li et al. (2024)). *With probability exceeding $1 - 3\delta$, $\mathcal{D}^{\text{main,sub}}$ and $\mathcal{D}^{\text{var,sub}}$ generated by Algorithm 4 as well as $\mathcal{D}^{\text{i.i.d.}}$ have the same distributions.*

C.3 Proof of Theorem 2

To show that Theorem 2 holds, we first establish the following theorem that considers the temporally independent dataset, where the proof is deferred to the next subsection.

Theorem 4. *Consider the dataset $\mathcal{D}_0$ and $\tilde{D}^0$ used for constructing the variance estimator in `DROP-V` and $\delta \in (0,1)$. Suppose that both $\mathcal{D}^0$ and $\tilde{D}^0$ contain $N_h < K$ sample tuples at every $h \in [H]$. Assume that conditional on $\{N_h\}_{h \in [H]}$, the sample tuples in $\mathcal{D}^0$ and $\tilde{D}^0$ are statistically independent, where*

$$N_h(s,a) \geq \frac{Kd_h^{\text{b}}(s,a)}{24}, \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

Suppose that Assumption 1, 2, and 3 hold. In `DROP-V`, we set

$$\lambda_1 = 1/H^2, \gamma_1 = \xi_1\sqrt{d}, \quad where \ \xi_1 = 66\log(3HK/\delta). \tag{85}$$

Then, with probability at least $1 - 7\delta$, $\{\widehat{\pi}_h\}_{h=1}^H$ generated by `DROP-V` satisfies

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \le \tilde{O}(\sqrt{d}) \sum_{h=1}^H \sum_{i=1}^d \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{d_h^\star} \left[\|\phi_i(s_h, a_h)\mathbb{1}_i\|_{(\Sigma_h^\star)^{-1}} \right],$$

if $\sqrt{d} \ge H$ and $K \ge \max\{\tilde{O}(H^4/\kappa^2), \tilde{O}(H^6 d/\kappa)\}$, where $\Sigma_h^\star$ is defined in (30).

As the construction in Algorithm 4, $\{N_h^{\mathrm{trim}}(s)\}_{s\in\mathcal{S}, h\in[H]}$ is computed using $\mathcal{D}^{\mathrm{aux}}$ that is independent of $\mathcal{D}^0 := \mathcal{D}^{\mathrm{main,sub}}$ and $\tilde{\mathcal{D}}^0 := \mathcal{D}^{\mathrm{var,sub}}$. Moreover, from Lemma 14 and Lemma 15 in the Section C.2, $\{N_h^{\mathrm{trim}}(s)\}_{s\in\mathcal{S}}$ is a valid sampling number and $\mathcal{D}_h^0$ and $\tilde{\mathcal{D}}_h^0$ can be treated as being temporally statistically independent samples and

$$\sum_{s\in\mathcal{S}} N_h^{\mathrm{trim}}(s) \ge K/24,$$

with probability exceeding $1 - 4\delta$, as long as $K \ge c_1 \log \frac{KH}{\delta}/d_{\min}^{\mathrm{b}}$ for some sufficiently large c_1.

Therefore, by invoking Theorem 3 with $N_h := \sum_{s\in\mathcal{S}} N^{\mathrm{trim}}(s)$, we have

$$\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) \le \tilde{O}(\sqrt{d}) \sum_{h=1}^H \sum_{i=1}^d \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{d_h^\star} \left[\|\phi_i(s_h, a_h)\mathbb{1}_i\|_{(\Sigma_h^\star)^{-1}} \right],$$

with probability exceeding $1 - 11\delta$, if $\sqrt{d} \ge H$ and $K \ge \max\{\tilde{O}(H^4/\kappa^2), \tilde{O}(H^6 d/\kappa), \tilde{O}(1/d_{\min}^{\mathrm{b}})\}$.

C.4 Proof of Theorem 4

Before starting, we first introduce some notations that will be used in the following analysis. First, we use

$$\iota_h^\sigma(s, a) = \mathbb{B}_h^{\rho, \sigma}\widehat{V}_{h+1}(s, a) - \widehat{Q}_h(s, a), \quad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{86}$$

to represent the model evaluation error at the h-th step of our proposed Algorithm 3. In addition, For any $h \in [H]$, we let $\Gamma_h^{\star,\sigma} : \mathcal{S} \to \mathbb{R}$ satisfy

$$\Gamma_h^{\star,\sigma}(s) = \Gamma_h^\sigma(s, \pi_h^\star(s)), \quad \forall s \in \mathcal{S}. \tag{87}$$

Also, denote $\mathbb{V}_{P_h^0}V(s, a) = \max\{1, \mathrm{Var}_{P_h^0}[V](s, a)\}$ for any $V : \mathcal{S} \to [0, H]$ and any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Similar to Lemma 10, we have the following key lemma, where the proof can be found in Appendix C.5.

Lemma 16. *Suppose all the assumptions in Theorem 4 hold and follow all the parameters setting in (85). In addition, suppose that the number of trajectories $K \ge \max\{\tilde{O}(H^4/\kappa^2), \tilde{O}(H^6 d/\kappa)\}$. Then for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, with probability at least $1 - 7\delta$, one has*

$$|(\widehat{\mathbb{B}}_h^{\rho,\sigma}\widehat{V}_{h+1})(s, a) - (\mathbb{B}_h^\rho\widehat{V}_{h+1})(s, a)| \le \Gamma_h^\sigma(s, a) := \gamma_1 \sum_{i=1}^d \|\phi_i(s, a)\mathbb{1}_i\|_{\Sigma_h^{-1}}. \tag{88}$$

In addition,

$$\gamma_1 \sum_{i=1}^d \|\phi_i(s, a)\mathbb{1}_i\|_{\Sigma_h^{-1}} \le 2\gamma_1 \sum_{i=1}^d \|\phi_i(s, a)\mathbb{1}_i\|_{(\Sigma_h^\star)^{-1}}.$$

Next, following the same steps in Appendix B.3.2, with probability exceeding $1 - 7\delta$, one has

$$
\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}; \zeta, \mathcal{P}^\rho) &\leq 2\gamma_1 \sum_{h=1}^{H} \sum_{i=1}^{d} \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{d_h^\star} \left[\|\phi_i(s_h, a_h) \mathbb{1}_i\|_{\Sigma_h^{-1}} \right] \\
&\leq 4\gamma_1 \sum_{h=1}^{H} \sum_{i=1}^{d} \max_{d_h^\star \in \mathcal{D}_h^\star} \mathbb{E}_{d_h^\star} \left[\|\phi_i(s_h, a_h) \mathbb{1}_i\|_{(\Sigma_h^\star)^{-1}} \right].
\end{aligned}
$$

C.5 Proof of Lemma 16

Similar to Lemma 11, we first establish the following lemma, which proof is postponed to Appendix C.5.1.

Lemma 17. *Suppose the Assumption 1 and 2 hold. Then, for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and any $V_{h+1} : \mathcal{S} \to [0, H]$, we have*

$$
\begin{aligned}
&|(\widehat{\mathbb{B}}_h^{\rho,\sigma} V_{h+1})(s, a) - (\mathbb{B}_h^\rho V_{h+1})(s, a)| \\
&\leq \left(2\sqrt{\lambda_1 d} H + \max_{\alpha \in [\min_s V_{h+1}(s), \max_s V_{h+1}(s)]} \| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha, V_{h+1}) \|_{\Sigma_h^{-1}} \right) \sum_{i=1}^{d} \|\phi_i(s, a) \mathbb{1}_i\|_{\Sigma_h^{-1}},
\end{aligned}
\tag{89}
$$

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

Letting $\lambda_1 = 1/H^2$ in (89), then Lemma 17 becomes

$$
\begin{aligned}
&|(\widehat{\mathbb{B}}_h^{\rho,\sigma} \widehat{V}_{h+1})(s, a) - (\mathbb{B}_h^{\rho,\sigma} \widehat{V}_{h+1})(s, a)| \\
&\leq \left(2\sqrt{d} + \underbrace{\max_{\alpha \in [\min_s \widehat{V}_{h+1}(s), \max_s \widehat{V}_{h+1}(s)]} \| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha, \widehat{V}_{h+1}) \|_{\Sigma_h^{-1}}}_{M_{1,h}} \right) \sum_{i=1}^{d} \|\phi_i(s, a) \mathbb{1}_i\|_{\Sigma_h^{-1}}.
\end{aligned}
\tag{90}
$$

Due to the correlation between α and $\widehat{V}_{h+1}$, we also apply the uniform concentration with the minimal ϵ_1-covering set $\mathcal{N}(\epsilon_1, H)$ for α defined in (52). Similar to (53), there exists $\alpha^\dagger \in \mathcal{N}(\epsilon_1, H)$ s.t.

$$
M_{1,h}^2 \leq 8\epsilon_1^2 H^2 K^2 + 2 \underbrace{\left\| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha^\dagger, \widehat{V}_{h+1}) \right\|_{\Sigma_h^{-1}}^2}_{M_{2,h}} \leq 8 + 2M_{2,h},
\tag{91}
$$

where the second inequality holds if $\epsilon_1 \leq \frac{1}{HK}$. Without the loss of generality, we let $\epsilon_1 = \frac{1}{HK}$ in the following analysis. The detailed proof of (91) is postponed to Appendix C.5.2.

Next, we will focus on bound the term $M_{2,h}$. Before proceeding, we first define the σ-algebra

$$
\mathcal{F}_{h,\tau} = \sigma(\{(s_h^j, a_h^j)\}_{j=1}^{(\tau+1) \wedge N_h}, \{r_h^j, s_{h+1}^j\}_{j=1}^{\tau}),
$$

for any fixed $h \in [H]$ and $\tau \in \mathcal{D}_h^0$. Noted that the samples in $\mathcal{D}^0$ are temporally statistically independent, i.e., $\widehat{V}_{h+1}$ is independent of $\mathcal{D}_h^0$ for any $h \in [H]$. In addition, $\{\widehat{\sigma}_h^2\}_{h \in [H]}$ is constructed using an additional dataset $\tilde{\mathcal{D}}^0$, which is also independent of $\mathcal{D}^0$. Thus, for any $h \in [H]$ and $\tau \in \mathcal{D}_h^0$, we have $\frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)}$ is $\mathcal{F}_{h,\tau-1}$-measurable and $|\frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)}| \leq 1$. Also, $\epsilon_h^{\tau,\sigma}(\alpha^\dagger, \widehat{V}_{h+1})$ is $\mathcal{F}_{h,\tau}$-measurable,

$$
\mathbb{E}[\epsilon_h^{\tau,\sigma}(\alpha^\dagger, \widehat{V}_{h+1}) | \mathcal{F}_{h,\tau-1}] = 0, \qquad |\epsilon_h^{\tau,\sigma}(\alpha^\dagger, \widehat{V}_{h+1})| \leq H,
$$

It follows the independence between $\widehat{\sigma}_h^2$ and $\mathcal{D}_h^0$ that

$$\mathrm{Var}_{P_h^0}\left[\frac{\int_{\mathcal{S}} P_h^0(s'|s_h^\tau,a_h^\tau)[\widehat{V}_{h+1}]_\alpha(s')\mathrm{d}s' - [\widehat{V}_{h+1}]_\alpha(s)}{\widehat{\sigma}_h(s_h^\tau,a_h^\tau)}\right](s_h^\tau,a_h^\tau) = \frac{\mathrm{Var}_{P_h^0}[\widehat{V}_{h+1}]_\alpha(s_h^\tau,a_h^\tau)}{\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)}$$
$$\leq \frac{\mathbb{V}_{P_h^0}\widehat{V}_{h+1}(s_h^\tau,a_h^\tau)}{\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)}, \tag{92}$$

for any $h \in [H]$ and $\tau \in \mathcal{D}_h^0$, where the inequality is from $\mathbb{V}_{P_h^0}\widehat{V}_{h+1}(s_h^\tau,a_h^\tau) = \max\{1, \mathrm{Var}_{P_h^0}[\widehat{V}_{h+1}](s_h^\tau,a_h^\tau)\}$.

The analysis of the improvement on sample complexity heavily relies on the following lemma about the variance estimation error, where the proof is deferred to Appendix C.5.3.

Lemma 18. *Suppose that $\mathcal{D}^0$ and $\tilde{\mathcal{D}}^0$ satisfy all the conditions imposed in Theorem 4. Assume the Assumption 1, 2 and 3 hold . For any $h \in [H]$ and given the nominal transition kernel $P_h^0 : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, the $\widehat{V}_{h+1}$ generated by the* DROP-V *on $\mathcal{D}^0$ and $\widehat{\sigma}_h^2$ generated by* DROP *on $\tilde{\mathcal{D}}^0$ satisfies*

$$\left|\mathbb{V}_{P_h^0}V_{h+1}^{\star,\rho}(s_h^\tau,a_h^\tau) - \widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right| \leq \frac{70c_b H^3\sqrt{d}}{\sqrt{K\kappa}}, \quad \forall\tau \in \mathcal{D}_h^0, \tag{93}$$

$$\left|\mathbb{V}_{P_h^0}\widehat{V}_{h+1}(s_h^\tau,a_h^\tau) - \mathbb{V}_{P_h^0}V_{h+1}^{\star,\rho}(s_h^\tau,a_h^\tau)\right| \leq \frac{320c_b H^3\sqrt{d}}{\sqrt{K\kappa}}, \quad \forall\tau \in \mathcal{D}_h^0, \tag{94}$$

where $c_b = 12\log(3HK/\delta)$ and $K \geq c_1\log(2Hd/\delta)H^4/\kappa^2$ for some sufficiently large universal constant c_1, with probability at least $1 - 6\delta$.

Notice that $1 \leq \widehat{\sigma}_h^2(s,a) \leq H^2$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Invoking the Lemma 18, we have

$$\frac{\mathbb{V}_{P_h^0}\widehat{V}_{h+1}(s_h^\tau,a_h^\tau)}{\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)} = \frac{\mathbb{V}_{P_h^0}V_{h+1}^{\star,\rho}(s_h^\tau,a_h^\tau)}{\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)} + \frac{\left|\mathbb{V}_{P_h^0}\widehat{V}_{h+1}(s_h^\tau,a_h^\tau) - \mathbb{V}_{P_h^0}V_{h+1}^{\star,\rho}(s_h^\tau,a_h^\tau)\right|}{\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)}$$
$$\leq 1 + \frac{\frac{70c_b H^3\sqrt{d}}{\sqrt{K\kappa}}}{\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)} + \frac{320c_b H^3\sqrt{d}}{\sqrt{K\kappa}\cdot\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)} \leq 1 + \frac{400c_b H^3\sqrt{d}}{\sqrt{K\kappa}} \leq 2$$

where the penultimate inequality uses $1 \leq \widehat{\sigma}_h^2(s,a)$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and the last inequality holds as long as $K \geq c_k H^6 d/\kappa$ for some sufficiently large universal constant c_k. Therefore, combining with (92) leads to $\mathrm{Var}[\epsilon_h^{\tau,\sigma}(\alpha^\dagger, V_{h+1})|\mathcal{F}_{h,\tau-1}] \leq 2$.

Suppose that $\sqrt{d} \geq H$. From Vershynin (2018), one has $|\mathcal{N}(\epsilon_1, H)| \leq \frac{3H}{\epsilon_1} = 3H^2K$. By the union bound and invoking Lemma 3, we have

$$\sup_{\alpha\in\mathcal{N}(\epsilon_1,H)}\left\|\sum_{\tau\in\mathcal{D}_h^0}\frac{\phi(s_h^\tau,a_h^\tau)}{\widehat{\sigma}(s_h^\tau,a_h^\tau)}\epsilon_h^{\tau,\sigma}(\alpha,\widehat{V}_{h+1})\right\|_{\Sigma_h^{-1}} \leq 16\sqrt{d\log(1+H^2K/d)\log(12H^3K^3/\delta)} + 4H\log(12H^3K^3/\delta)$$
$$\leq c_1\sqrt{d}$$

with probability $1 - 7\delta$ and for a fixed $\alpha \in \mathcal{N}(\epsilon_1, \widehat{V}_{h+1})$, for $c_1 = 40\log(3HK/\delta)$. Then, the equation (89) becomes

$$|(\widehat{\mathbb{B}}_h^{\rho,\sigma}\widehat{V}_{h+1})(s,a) - (\mathbb{B}_h^\rho\widehat{V}_{h+1})(s,a)| \leq (2\sqrt{d} + 2\sqrt{2} + \sqrt{2}c_1\sqrt{d})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Sigma_h^{-1}}$$
$$= \gamma_1\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Sigma_h^{-1}} := \Gamma_h^\sigma(s,a).$$

Moreover, from (93), we have

$$
\sum_{\tau \in \mathcal{D}_h^0} \frac{\phi_h^\tau(s_h^\tau, a_h^\tau)\phi_h^\tau(s_h^\tau, a_h^\tau)^\top}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} \succeq \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi_h^\tau(s_h^\tau, a_h^\tau)\phi_h^\tau(s_h^\tau, a_h^\tau)^\top}{\mathbb{V}_{P_h^0} V_{h+1}^{\star,\rho}(s_h^\tau, a_h^\tau) + \frac{70 c_b H^3 d}{\sqrt{K j i \kappa}}}
$$
$$
\succeq \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi_h^\tau(s_h^\tau, a_h^\tau)\phi_h^\tau(s_h^\tau, a_h^\tau)^\top}{2\mathbb{V}_{P_h^0} V_{h+1}^{\star,\rho}(s_h^\tau, a_h^\tau)}
$$

where the last inequality is from $\frac{70 c_b H^3 \sqrt{d}}{\sqrt{K\kappa}} \le \frac{1}{2}$. Then, we obtain $\Sigma_h \succeq \frac{1}{2}\Sigma_h^\star$ for any $h \in [H]$, which completes the proof.

C.5.1 Proof of Lemma 17

Following (56)-(58), we have

$$
|(\widehat{\mathbb{B}}_h^{\rho,\sigma} V_{h+1})(s, a) - (\mathbb{B}_h^\rho V_{h+1})(s, a)|
$$
$$
\le \sqrt{d\lambda_1} \sum_{i=1}^d \|\phi_i(s, a)\mathbb{1}_i\|_{\Sigma_h^{-1}} + \underbrace{\sum_{i=1}^d \max_{\alpha \in [\min_s V_{h+1}(s), \max_s V_{h+1}(s)]} \left| \phi_i(s, a) \int_{\mathcal{S}} (\widehat{\mu}_{h,i}^\sigma(s') - \mu_{h,i}(s'))[V_{h+1}]_\alpha(s')\mathrm{d}s' \right|}_{\text{(i)}},
$$

for $\forall (s, a, h) \times \mathcal{S} \times \mathcal{A} \times [H]$, where $\widehat{\mu}_{h,i}^\sigma(s)$ is the i-th coordinate of

$$
\widehat{\mu}_h^\sigma(s) = \Sigma_h^{-1} \left(\sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)\mathbb{1}(s = s_{h+1}^\tau)}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} \right) \in \mathbb{R}^d
$$

such that $\bar{\nu}_h^{\widehat{V}}(\alpha) = \int_{\mathcal{S}} \widehat{\mu}_h^\sigma(s')[\widehat{V}_{h+1}(s')]_\alpha \mathrm{d}s'$ defined in the update (26). Similar to (59), by letting $\epsilon_h^{\tau,\sigma}(\alpha, V) = \frac{\int_{\mathcal{S}} P_h^0(s'|s_h^\tau, a_h^\tau)[V]_\alpha(s')\mathrm{d}s' - [V]_\alpha(s_{h+1}^\tau)}{\widehat{\sigma}(s_h^\tau, a_h^\tau)}$ for any $V : \mathcal{S} \to [0, H]$, any $\tau \in \mathcal{D}_h^0$ and $\alpha \in [\min_s V(s), \max_s V(s)]$, we have

$$
\left| \int_{\mathcal{S}} \mu_{h,i}^0(s')[V_{h+1}]_\alpha(s')\mathrm{d}s' - \int_{\mathcal{S}} \widehat{\mu}_{h,i}^\sigma(s')[V_{h+1}]_\alpha(s')\mathrm{d}s' \right|
$$
$$
= \left| \mathbb{1}_i^\top \Sigma_h^{-1} \left[\lambda_1 \int_{\mathcal{S}} \mu_h^0(s')[V_{h+1}]_\alpha(s')\mathrm{d}s' + \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} \left(\int_{\mathcal{S}} P_h^0(s'|s_h^\tau, a_h^\tau)[V_{h+1}]_\alpha(s')\mathrm{d}s' - [V_{h+1}]_\alpha(s_{h+1}^\tau) \right) \right] \right|
$$
$$
= \left| \mathbb{1}_i^\top \Sigma_h^{-1} \left[\lambda_1 \int_{\mathcal{S}} \mu_h^0(s')[V_{h+1}]_\alpha(s')\mathrm{d}s' + \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha, V_{h+1}) \right] \right|
$$

Then, we obtain

$$
\left| \phi_i(s, a) \int_{\mathcal{S}} (\widehat{\mu}_{h,i}^\sigma(s') - \mu_{h,i}(s'))[V_{h+1}]_\alpha(s')\mathrm{d}s' \right|
$$
$$
\le \left| \phi_i(s, a)\mathbb{1}_i^\top \Sigma_h^{-1} \left(\lambda_1 \int_{\mathcal{S}} \mu_h^0(s')[V_{h+1}]_\alpha(s')\mathrm{d}s' + \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha, V_{h+1}) \right) \right|
$$
$$
\le \|\phi_i(s, a)\mathbb{1}_i\|_{\Sigma_h^{-1}} \left(\lambda_1 \underbrace{\| \int_{\mathcal{S}} \mu_h^0(s')[V_{h+1}]_\alpha(s')\mathrm{d}s' \|_{\Sigma_h^{-1}}}_{\text{(ii)}} + \| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha, V_{h+1}) \|_{\Sigma_h^{-1}} \right), \quad (95)
$$

where the last inequality follows Cauchy-Schwarz inequality. Moreover, the term (ii) in (95) can be further simplified

$$\text{(ii)} \leq \lambda_1 \|\Sigma_h^{-1}\|^{\frac{1}{2}} \| \int_{\mathcal{S}} \mu_h^0(s')[V_{h+1}]_\alpha(s')\mathrm{d}s' \| \leq \sqrt{\lambda_1} H,$$

since $V(s) \leq H$ for any $s \in \mathcal{S}$ and $\|\Sigma_h^{-1}\| \leq 1/\lambda_1$. Then we have

$$\text{(i)} \leq \left(\sqrt{\lambda_1} H + \max_{\alpha \in [\min_s V_{h+1}(s), \max_s V_{h+1}(s)]} \| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha, V_{h+1}) \|_{\Sigma_h^{-1}} \right) \sum_{i=1}^{d} \|\phi_i(s, a) \mathbb{1}_i \|_{\Sigma_h^{-1}}, \tag{96}$$

which concludes our proof of (89).

C.5.2 Proof of (91)

Due to the semi-positiveness of Σ_h^{-1}, one can control $M_{1,h}^2$ for any $h \in [H]$ as

$$\max_{\alpha \in [\min_s \widehat{V}_{h+1}(s), \max_s \widehat{V}_{h+1}(s)]} \| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha, \widehat{V}_{h+1}) \|_{\Sigma_h^{-1}}^2$$

$$\leq \max_{\alpha \in [0,H]} 2\| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \left(\epsilon_h^{\tau,\sigma}(\alpha, \widehat{V}_{h+1}) - \epsilon_h^{\tau,\sigma}(\alpha^\dagger, \widehat{V}_{h+1}) \right) \|_{\Sigma_h^{-1}}^2 + 2\| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha^\dagger, \widehat{V}_{h+1}) \|_{\Sigma_h^{-1}}^2,$$

for some $\alpha^\dagger \in \mathcal{N}(\epsilon_1, H)$. Note that $\epsilon_h^{\tau,\sigma}(\alpha, V)$ is 2-Lipschitz w.r.t. α for any $V : \mathcal{S} \to [0, H]$, i.e.,

$$|\epsilon_h^{\tau,\sigma}(\alpha, V) - \epsilon_h^{\tau,\sigma}(\alpha^\dagger, V)| \leq 2|\alpha - \alpha^\dagger| \leq 2\epsilon_1.$$

Therefore, for any $\alpha \in [0, H]$, we have

$$\| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \left(\epsilon_h^{\tau,\sigma}(\alpha, V) - \epsilon_h^{\tau,\sigma}(\alpha^\dagger, V) \right) \|_{\Sigma_h^{-1}}^2$$

$$= \sum_{\tau, \tau' \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)}^\top \Sigma_h^{-1} \frac{\phi(s_h^{\tau'}, a_h^{\tau'})}{\widehat{\sigma}_h(s_h^{\tau'}, a_h^{\tau'})} \left[\left(\epsilon_h^{\tau,\sigma}(\alpha, V) - \epsilon_h^{\tau,\sigma}(\alpha^\dagger, V) \right) \left(\epsilon_h^{\tau',\sigma}(\alpha, V) - \epsilon_h^{\tau',\sigma}(\alpha^\dagger, V) \right) \right]$$

$$\leq \sum_{\tau, \tau' \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)}^\top \Sigma_h^{-1} \frac{\phi(s_h^{\tau'}, a_h^{\tau'})}{\widehat{\sigma}_h(s_h^{\tau'}, a_h^{\tau'})} \cdot 4\epsilon_1^2$$

$$\leq 4\epsilon_1^2 N_h^2 / \lambda_1,$$

where the last inequality is based on $\|\phi(s, a)\|_2 \leq 1$, $\widehat{\sigma}_h(s, a) \geq 1$ for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $\lambda_{\min}(\Sigma_h) \geq \lambda_1 = \frac{1}{H^2}$ for any $h \in [H]$ such that

$$\sum_{\tau, \tau' \in \mathcal{D}_h^0} \phi(s_h^\tau, a_h^\tau)^\top \Sigma_h^{-1} \phi(s_h^{\tau'}, a_h^{\tau'}) = \sum_{\tau, \tau' \in \mathcal{D}_h^0} \|\phi(s_h^\tau, a_h^\tau)\|_2 \cdot \|\phi(s_h^{\tau'}, a_h^{\tau'})\|_2 \cdot \|\Sigma_h^{-1}\| \leq N_h^2 / \lambda_1. \tag{97}$$

Due to the fact $N_h \leq K$ for nay $h \in [H]$, we conclude that

$$M_{1,h}^2 \leq 8\epsilon_1^2 H^2 K^2 + 2 \underbrace{\left\| \sum_{\tau \in \mathcal{D}_h^0} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha^\dagger, \widehat{V}_{h+1}) \right\|_{\Sigma_h^{-1}}^2}_{M_{2,h}} \leq 8 + 2M_{2,h}, \tag{98}$$

where the second inequality holds if $\epsilon_1 \leq \frac{1}{HK}$.

C.5.3 Proof of Lemma 18

Recall that in Section 4.1, $\{\widehat{\sigma}_h^2\}_{h=1}^H$ is constructed via $\{\tilde{V}_{h+1}\}_{h=1}^H$ generated by DROP on $\tilde{\mathcal{D}}^0$. Before starting, we define

$$\widehat{\mathrm{Var}}_h\tilde{V}_{h+1}(s,a) = [\phi(s,a)^\top\nu_{h,1}]_{[0,H^2]} - \left([\phi(s,a)^\top\nu_{h,2}]_{[0,H]}\right)^2, \quad \forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H],$$

such that $\widehat{\sigma}_h^2(s,a) = \max\{1,\widehat{\mathrm{Var}}_h\tilde{V}_{h+1}(s,a)\}$. In addition, recall that $\mathbb{V}_{P_h^0}V(s,a) = \max\{1,\mathrm{Var}_{P_h^0}V(s,a)\}$ for any value function $V:\mathcal{S}\to[0,H]$ and any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$. Then, we can decompose the target terms by

$$
\begin{aligned}
&\left|\mathbb{V}_{P_h^0}V_{h+1}^{\star,\rho}(s_h^\tau,a_h^\tau) - \widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right| \\
&\leq \left|\mathbb{V}_{P_h^0}\tilde{V}_{h+1}(s_h^\tau,a_h^\tau) - \widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right| + \left|\mathbb{V}_{P_h^0}V_{h+1}^{\star,\rho}(s_h^\tau,a_h^\tau) - \mathbb{V}_{P_h^0}\tilde{V}_{h+1}(s_h^\tau,a_h^\tau)\right| \\
&\leq \underbrace{\left|\mathrm{Var}_{P_h^0}\tilde{V}_{h+1}(s_h^\tau,a_h^\tau) - \widehat{\mathrm{Var}}_h\tilde{V}_{h+1}(s_h^\tau,a_h^\tau)\right|}_{(a)} + \underbrace{\left|\mathrm{Var}_{P_h^0}V_{h+1}^{\star,\rho}(s_h^\tau,a_h^\tau) - \mathrm{Var}_{P_h^0}\tilde{V}_{h+1}(s_h^\tau,a_h^\tau)\right|}_{(b)},
\end{aligned}
$$

for every $h\in[H]$, where the last inequality is based on the non-expansiveness of $\max\{1,\cdot\}$. Similarly,

$$\left|\mathbb{V}_{P_h^0}\widehat{V}_{h+1}(s_h^\tau,a_h^\tau) - \mathbb{V}_{P_h^0}V_{h+1}^{\star,\rho}(s_h^\tau,a_h^\tau)\right| \leq \underbrace{\left|\mathrm{Var}_{P_h^0}V_{h+1}^{\star,\rho}(s_h^\tau,a_h^\tau) - \mathrm{Var}_{P_h^0}\widehat{V}_{h+1}(s_h^\tau,a_h^\tau)\right|}_{(c)}.$$

In the sequel, we will control (a), (b) and (c) respectively.

Step 1: controlling (a). For each $\tau\in\mathcal{D}_h^0$, we decompose the term (a) by

$$
\begin{aligned}
(a) &= \left|\widehat{\mathrm{Var}}_h\tilde{V}_{h+1}(s_h^\tau,a_h^\tau) - \mathrm{Var}_{P_h^0}\tilde{V}_{h+1}(s_h^\tau,a_h^\tau)\right| \\
&\leq \left|[\phi(s_h^\tau,a_h^\tau)^\top\nu_{h,1}]_{[0,H^2]} - \int_{\mathcal{S}}P_{h,s_h^\tau,a_h^\tau}^0(s')\tilde{V}_{h+1}^2(s')\mathrm{d}s'\right| \\
&\quad + \left|\left([\phi(s,a)^\top\nu_{h,2}]_{[0,H]}\right)^2 - \left[\int_{\mathcal{S}}P_{h,s_h^\tau,a_h^\tau}^0(s')\tilde{V}_{h+1}(s')\mathrm{d}s'\right]^2\right| \\
&\leq \underbrace{\left|\phi(s_h^\tau,a_h^\tau)^\top\left(\nu_{h,1} - \int_{\mathcal{S}}\mu_h^0(s')\tilde{V}_{h+1}^2(s')\mathrm{d}s'\right)\right|}_{(a_1)} + 2H\underbrace{\left|\phi(s_h^\tau,a_h^\tau)^\top\left(\nu_{h,2} - \int_{\mathcal{S}}\mu_h^0(s')\tilde{V}_{h+1}(s')\mathrm{d}s'\right)\right|}_{(a_2)},
\end{aligned}
$$

where the last inequality is based on $a^2 - b^2 = (a+b)(a-b)$ for any $a,b\in\mathbb{R}$. In the sequel, we control (a_1) and (a_2), respectively. Before continuing, we first define $\tilde{\mu}_{h,i}:\mathcal{S}\to\mathbb{R}$ is the i-th coordinate of

$$\tilde{\mu}_h(s) = (\tilde{\Lambda}_h)^{-1}\left(\sum_{\tau'\in\tilde{\mathcal{D}}_h^0}\phi(s_h^{\tau'},a_h^{\tau'})\mathbb{1}(s = s_{h+1}^{\tau'})\right) \in \mathbb{R}^d$$

such that $\nu_{h,1} = \int_{\mathcal{S}}\tilde{\mu}_h(s')\tilde{V}_{h+1}^2(s')\mathrm{d}s'\in\mathbb{R}^d$ and $\nu_{h,2} = \int_{\mathcal{S}}\tilde{\mu}_h(s')\tilde{V}_{h+1}(s')\mathrm{d}s'$. With this new notation, we reformulate (a_1) as

$$(a_1) = |\phi(s_h^\tau,a_h^\tau)\int_{\mathcal{S}}(\tilde{\mu}_h(s') - \mu_h^0(s'))\tilde{V}_{h+1}^2(s')\mathrm{d}s'|.$$

Following the steps in Lemma 11, i.e., the equations (59)-(61) with $\lambda_0 = 1$, we can obtain

$$
(a_1) \le \left(H + \max_{\alpha \in [\min_s \tilde{V}_{h+1}(s), \max_s \tilde{V}_{h+1}(s)]} \| \sum_{\tau' \in \tilde{\mathcal{D}}_h^0} \phi(s_h^{\tau'}, a_h^{\tau'}) \epsilon_h^{\tau'}(\alpha, \tilde{V}_{h+1}^2) \|_{(\tilde{\Lambda}_h)^{-1}} \right) \sum_{i=1}^{d} \| \phi_i(s_h^\tau, a_h^\tau) \mathbb{1}_i \|_{\tilde{\Lambda}_h^{-1}}
$$

$$
\le \left(H + 2\sqrt{2}H + \sqrt{2} \sup_{\alpha \in \mathcal{N}(\epsilon_0, H)} \| \sum_{\tau' \in \tilde{\mathcal{D}}_h^0} \phi(s_h^{\tau'}, a_h^{\tau'}) \epsilon_h^{\tau'}(\alpha, \tilde{V}_{h+1}^2) \|_{(\tilde{\Lambda}_h)^{-1}} \right) \sum_{i=1}^{d} \| \phi_i(s_h^\tau, a_h^\tau) \mathbb{1}_i \|_{\tilde{\Lambda}_h^{-1}},
$$

(99)

where $\epsilon_h^\tau(\alpha, V) = \int_{\mathcal{S}} P_h^0(s'|s_h^\tau, a_h^\tau)[V]_\alpha(s')\mathrm{d}s' - [V]_\alpha(s_{h+1}^\tau)$ for any $V : \mathcal{S} \to [0, H]$, $\tau' \in \tilde{\mathcal{D}}_h^0$ and $\alpha \in [\min_s V(s), \max_s V(s)]$. Since $\tilde{V}_{h+1}$ is independent of $\tilde{\mathcal{D}}_h^0$, we can directly apply Lemma 2 following the same arguments in Lemma 12. Therefore, with probability exceeding $1 - \delta$, we have

$$
\sup_{\alpha \in \mathcal{N}(\epsilon_0, H)} \| \sum_{\tau' \in \tilde{\mathcal{D}}_h^0} \phi(s_h^{\tau'}, a_h^{\tau'}) \epsilon_h^{\tau'}(\alpha, \tilde{V}_{h+1}^2) \|_{(\tilde{\Lambda}_h)^{-1}} \le H^2 \sqrt{2\log(3HK/\delta) + d\log(1 + K)} \le c_a H^2 \sqrt{d},
$$

(100)

where $c_a = 3\log(3HK/\delta)$. Therefore, with probability exceeding $1 - \delta$,

$$
(a_1) \le 6c_a H^2 \sqrt{d} \sum_{i=1}^{d} \| \phi_i(s_h^\tau, a_h^\tau) \mathbb{1}_i \|_{\tilde{\Lambda}_h^{-1}}.
$$

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

$$
(a_2) = |\phi(s_h^\tau, a_h^\tau) \int_{\mathcal{S}} (\tilde{\mu}_h(s') - \mu_h^0(s')) \tilde{V}_{h+1}(s') \mathrm{d}s'| \le 6c_a H \sqrt{d} \sum_{i=1}^{d} \| \phi_i(s_h^\tau, a_h^\tau) \mathbb{1}_i \|_{\tilde{\Lambda}_h^{-1}}.
$$

(101)

Combining (99), (100), and (101), we can obtain

$$
(a) \le (a_1) + 2H(a_2) \le 12c_a H^2 \sqrt{d} \sum_{i=1}^{d} \| \phi_i(s_h^\tau, a_h^\tau) \mathbb{1}_i \|_{\tilde{\Lambda}_h^{-1}}
$$

with probability exceeding $1 - 2\delta$.

Step 2: controlling (b)**.** Then,

$$
(b) = \left| \mathrm{Var}_{P_h^0} V_{h+1}^{\star,\rho}(s_h^\tau, a_h^\tau) - \mathrm{Var}_{P_h^0} \tilde{V}_{h+1}(s_h^\tau, a_h^\tau) \right|
$$

$$
\le \left| \int_{\mathcal{S}} P_{h, s_h^\tau, a_h^\tau}(s') \left(V_{h+1}^{\star,\rho}(s') - \tilde{V}_{h+1}(s') \right) \left(V_{h+1}^{\star,\rho}(s') + \tilde{V}_{h+1}(s') \right) \mathrm{d}s' \right|
$$

$$
+ \left| \int_{\mathcal{S}} P_{h, s_h^\tau, a_h^\tau}(s') \left(V_{h+1}^{\star,\rho}(s') - \tilde{V}_{h+1}(s') \right) \mathrm{d}s' \right| \left| \int_{\mathcal{S}} P_{h, s_h^\tau, a_h^\tau}(s') \left(V_{h+1}^{\star,\rho}(s') + \tilde{V}_{h+1}(s') \right) \mathrm{d}s' \right|
$$

$$
\le 4H \left| \int_{\mathcal{S}} P_{h, s_h^\tau, a_h^\tau}(s') \left(V_{h+1}^{\star,\rho}(s') - \tilde{V}_{h+1}(s') \right) \mathrm{d}s' \right| \le 4H \max_{s \in \mathcal{S}} V_{h+1}^{\star,\rho}(s) - \tilde{V}_{h+1}(s).
$$

Denote $\tilde{\iota}_h(s, a) = \mathbb{B}_h^\rho \tilde{V}_{h+1}(s, a) - \tilde{Q}_h(s, a)$, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and

$$
P_{h, s, \pi_h^\star(s)}^{\inf, \tilde{V}}(\cdot) := \arg\min_{P(\cdot) \in \mathcal{P}^\rho(P_{h, s, \pi_h^\star(s)}^0)} \int_{\mathcal{S}} P(s') \tilde{V}_{h+1}(s') \mathrm{d}s'.
$$

(102)

For any $h \in [H]$, define $\tilde{P}_h^{\inf} : \mathcal{S} \to \mathcal{S}$ and $\tilde{\iota}_h^\star \in \mathcal{S} \to \mathbb{R}$ by

$$
\tilde{P}_h^{\inf}(s) = P_{h, s, \pi_h^\star(s)}^{\inf, \tilde{V}}(\cdot) \quad \text{and} \quad \tilde{\iota}_h^\star(s) := \tilde{\iota}_h(s, \pi^\star(s)), \quad \forall s \in \mathcal{S}.
$$

(103)

Following the step 1 and 2 in Section B.3.2, we have

$$V_h^{\star,\rho}(s) - \tilde{V}_h(s) \le \left(\prod_{t=h}^{H} \tilde{P}_j^{\inf} \right) \left(V_{H+1}^{\star,\rho} - \tilde{V}_{H+1} \right)(s) + \sum_{t=h}^{H} \left(\prod_{j=h}^{t-1} \tilde{P}_j^{\inf} \right) \tilde{\iota}_t^{\star}(s)$$

$$= \sum_{t=h}^{H} \left(\prod_{j=h}^{t-1} \tilde{P}_j^{\inf} \right) \tilde{\iota}_t^{\star}(s)$$

for any $s \in \mathcal{S}$ and $h \in [H]$, where the equality is from $V_{H+1}^{\star,\rho}(s) = \tilde{V}_{H+1}(s) = 0$ for any $s \in \mathcal{S}$ and we denote

$$\left(\prod_{j=t}^{t-1} \tilde{P}_j^{\inf} \right)(s) = \mathbb{1}_s \quad \text{and} \quad \tilde{d}_{h:t}^{\star} = d_h^{\star} \left(\prod_{j=h}^{t-1} \tilde{P}_j^{\inf} \right) \in \mathcal{D}_t^{\star}.$$

for any $d_h^{\star} \in \mathcal{D}_h^{\star}$. Therefore,

$$\max_{s \in \mathcal{S}} V_{h+1}^{\star,\rho}(s) - \tilde{V}_{h+1}(s) \le \sum_{t=h+1}^{H} \max_{s \in \mathcal{S}} \mathbb{E}_{\tilde{d}_{h:t}^{\star}} \tilde{\iota}_t^{\star} \le \sum_{h=1}^{H} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \tilde{\iota}_h(s,a). \tag{104}$$

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

$$|\tilde{\iota}_h(s,a)| \le |(\widehat{\mathbb{B}}_h^{\rho} \tilde{V}_{h+1})(s,a) - (\mathbb{B}_h^{\rho} \tilde{V}_{h+1})(s,a)| + \Gamma_h(s,a) \le 2\Gamma_h(s,a)$$

$$\le c_b H \sqrt{d} \sum_{i=1}^{d} \|\phi_i(s,a) \mathbb{1}_i\|_{\tilde{\Lambda}_h^{-1}}, \tag{105}$$

where $c_b = 12 \log(3HK/\delta)$. Substituting (105) into (104), we have

$$\max_{s \in \mathcal{S}} V_{h+1}^{\star,\rho}(s) - \tilde{V}_{h+1}(s) \le \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} c_b H^2 \sqrt{d} \sum_{i=1}^{d} \|\phi_i(s,a) \mathbb{1}_i\|_{\tilde{\Lambda}_h^{-1}}.$$

Therefore,

$$(b) \le \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} 4 c_b H^3 \sqrt{d} \sum_{i=1}^{d} \|\phi_i(s,a) \mathbb{1}_i\|_{\tilde{\Lambda}_h^{-1}}$$

with probability exceeding $1 - \delta$.

Step 3: controlling (c).   Similarly,

$$(c) \le 4H \max_{s \in \mathcal{S}} V_{h+1}^{\star,\rho}(s) - \widehat{V}_{h+1}(s) \le \sum_{t=h+1}^{H} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\iota_t^{\sigma}(s,a)| \tag{106}$$

where

$$|\iota_h^{\sigma}(s,a)| \le |(\widehat{\mathbb{B}}_h^{\rho,\sigma} \widehat{V}_{h+1})(s,a) - (\mathbb{B}_h^{\rho} \widehat{V}_{h+1})(s,a)| + \Gamma_h^{\sigma}(s,a). \tag{107}$$

Following (90) and (91), we have

$$|(\widehat{\mathbb{B}}_h^{\rho,\sigma} \widehat{V}_{h+1})(s,a) - (\mathbb{B}_h^{\rho} \widehat{V}_{h+1})(s,a)| \le \left(5\sqrt{d} + \sqrt{2} M_{3,h} \right) \sum_{i=1}^{d} \|\phi_i(s,a) \mathbb{1}_i\|_{\Sigma_h^{-1}}, \tag{108}$$

where $M_{3,h} := \sup_{\alpha^\dagger \in \mathcal{N}(\epsilon_1, H)} \left\| \sum_{\tau=1}^{N_h} \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \epsilon_h^{\tau,\sigma}(\alpha^\dagger, \widehat{V}_{h+1}) \right\|_{\Sigma_h^{-1}}$. Since $\widehat{\sigma}_h$ and $\widehat{V}_{h+1}$ are independent of $\mathcal{D}_h^0$, then we can directly apply Lemma 2 following the same arguments in Lemma 12. Therefore, with probability exceeding $1 - \delta$, we have

$$M_{3,h} \leq H \sqrt{2 \log(\frac{3H^2}{\epsilon_1 \delta}) + d \log(1 + N_h/\lambda_1)}$$
$$\leq H \sqrt{2 \log(3H^3 K/\delta) + d \log(2H^2 K)} \leq c_b H \sqrt{d}/\sqrt{2}, \tag{109}$$

where $c_b = 12 \log(3HK/\delta)$. Substituting (109) into (108) and combining with (107) result in

$$|\iota_h^\sigma(s,a)| \leq (5 + c_b + \xi_1) H \sqrt{d} \sum_{i=1}^d \|\phi_i(s,a) \mathbb{1}_i\|_{\Sigma_h^{-1}},$$
$$\leq 8 c_b H \sqrt{d} \sum_{i=1}^d \|\phi_i(s,a) \mathbb{1}_i\|_{\Sigma_h^{-1}}, \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

Therefore,

$$(c) \leq 4H \max_{s \in \mathcal{S}} V_{h+1}^{\star,\rho}(s) - \widehat{V}_{h+1}(s) \leq \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} 32 c_b H^3 \sqrt{d} \sum_{i=1}^d \|\phi_i(s,a) \mathbb{1}_i\|_{\Sigma_h^{-1}}$$

with probability at least $1 - \delta$.

Step 4: finishing up. Then, with probability at least $1 - 3\delta$, we have

$$\left| \mathbb{V}_{P_h^0} V_{h+1}^{\star,\rho}(s_h^\tau, a_h^\tau) - \widehat{\sigma}_h^2(s_h^\tau, a_h^\tau) \right| \leq \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} 7 c_b H^3 \sqrt{d} \sum_{i=1}^d \|\phi_i(s,a) \mathbb{1}_i\|_{\tilde{\Lambda}_h^{-1}}, \tag{110}$$

since $c_b = 4 c_a$. With probability at least $1 - \delta$,

$$\left| \mathbb{V}_{P_h^0} \widehat{V}_{h+1}(s_h^\tau, a_h^\tau) - \mathbb{V}_{P_h^0} V_{h+1}^{\star,\rho}(s_h^\tau, a_h^\tau) \right| \leq \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} 32 c_b H^3 \sqrt{d} \sum_{i=1}^d \|\phi_i(s,a) \mathbb{1}_i\|_{\Sigma_h^{-1}}. \tag{111}$$

Recall that from Lemma 13, we can control the term $\sum_{i=1}^d \|\phi_i(s,a) \mathbb{1}_i\|_{\tilde{\Lambda}_h^{-1}}$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, as long as N_h is sufficiently large. Similar to Lemma 13, we also employ Lemma 4 to control the term $\sum_{i=1}^d \|\phi_i(s,a) \mathbb{1}_i\|_{\Sigma_h^{-1}}$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ as follows, where the proof is deferred to Appendix C.5.4.

Lemma 19. *Consider $\delta \in (0,1)$. Suppose Assumption 2, Assumption 4 and all conditions in Lemma 4 hold. For any $h \in [H]$, if $N_h \geq 512 \log(2Hd/\delta) H^4/\kappa^2$, we have*

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

with probability exceeding $1 - \delta$.

From Lemma 13, Lemma 19 and the fact $N_h \geq \frac{K}{24}$, with probability exceeding $1 - 2\delta$, we have

$$\sum_{i=1}^d \|\phi_i(s,a) \mathbb{1}_i\|_{\tilde{\Lambda}_h^{-1}} \leq \frac{2}{\sqrt{N_h \kappa}} \leq \frac{10}{\sqrt{K \kappa}}, \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{112}$$

$$\sum_{i=1}^d \|\phi_i(s,a) \mathbb{1}_i\|_{\Sigma_h^{-1}} \leq \frac{2}{\sqrt{N_h \kappa}} \leq \frac{10}{\sqrt{K \kappa}}, \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{113}$$

as long as $K \geq c_1 \log(2Hd/\delta)H^4/\kappa^2$ for some sufficient large universal constant c_1. Substituting (112) and (113) into (110) and (111) respectively, we finally arrive at

$$\left| \mathbb{V}_{P_h^0} V_{h+1}^{\star,\rho}(s_h^\tau, a_h^\tau) - \widehat{\sigma}_h^2(s_h^\tau, a_h^\tau) \right| \leq \frac{70c_b H^3 \sqrt{d}}{\sqrt{K\kappa}}, \tag{114}$$

$$\left| \mathbb{V}_{P_h^0} \widehat{V}_{h+1}(s_h^\tau, a_h^\tau) - \mathbb{V}_{P_h^0} V_{h+1}^{\star,\rho}(s_h^\tau, a_h^\tau) \right| \leq \frac{320c_b H^3 \sqrt{d}}{\sqrt{K\kappa}}, \tag{115}$$

for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, with probability at least $1 - 6\delta$, which completes the proof as long as $K \geq c_1 \log(2Hd/\delta)H^4/\kappa^2$ for some sufficient large universal constant c_1.

C.5.4 Proof of Lemma 19

From Assumption 4, one has $\lambda_{\min}(\mathbb{E}_{d_h^b}[\frac{\phi(s,a)\phi(s,a)^\top}{\widehat{\sigma}_h^2(s,a)}]) \geq \frac{\kappa}{H^2}$ for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Following Lemma 4, we can obtain

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

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

$$1 = \int_{\mathcal{S}} P_h^0(s'|s,a)\mathrm{d}s' = \int_{\mathcal{S}} \phi(s,a)^\top \mu_h^0(s')\mathrm{d}s' = \sum_{i=1}^d \phi_i(s,a) \int_{\mathcal{S}} \mu_{h,i}^0(s')\mathrm{d}s' = \sum_{i=1}^d \phi_i(s,a), \tag{116}$$

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

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