Achieving the Asymptotically Optimal Sample Complexity of Offline Reinforcement Learning: A DRO-Based Approach

Anonymous authors
Paper under double-blind review

Abstract

Offline reinforcement learning aims to learn from pre-collected datasets without active exploration. This problem faces significant challenges, including limited data availability and distributional shifts. Existing approaches adopt a pessimistic stance towards uncertainty by penalizing rewards of under-explored state-action pairs to estimate value functions conservatively. In this paper, we show that the distributionally robust optimization (DRO) based approach can also address these challenges and is asymptotically minimax optimal. Specifically, we directly model the uncertainty in the transition kernel and construct an uncertainty set of statistically plausible transition kernels. We then show that the policy that optimizes the worst-case performance over this uncertainty set has a near-optimal performance in the underlying problem. We first design a metric-based Hoeffding-style uncertainty set such that with high probability the true transition kernel is in this set. We prove that to achieve a sub-optimality gap of ϵ, the sample complexity is $\mathcal{O}(S^2C^{\pi^*}\epsilon^{-2}(1-\gamma)^{-4})$, where γ is the discount factor, S is the number of states, and C^{π^*} is the single-policy clipped concentrability coefficient which quantifies the distribution shift. To achieve the optimal sample complexity, we further propose a less conservative Bernstein-style uncertainty set, which, however, does not necessarily include the true transition kernel. We show that an improved sample complexity of $\mathcal{O}(SC^{\pi^*}\epsilon^{-2}(1-\gamma)^{-3})$ can be obtained, which asymptotically matches with the minimax lower bound for offline reinforcement learning, and thus is asymptotically minimax optimal.

1 Introduction

Reinforcement learning (RL) has achieved impressive empirical success in many domains, e.g., (Mnih et al., 2015; Silver et al., 2016). Nonetheless, most of the success stories rely on the premise that the agent can actively explore the environment and receive feedback to further promote policy improvement. This trial-and-error procedure can be costly, unsafe, or even prohibitory in many real-world applications, e.g., autonomous driving (Kiran et al., 2021) and health care (Yu et al., 2021a). To address the challenge, offline (or batch) reinforcement learning (Lange et al., 2012; Levine et al., 2020) was developed, which aims to learn a competing policy from a pre-collected dataset without access to online exploration.

A straightforward idea for offline RL is to use the pre-collected dataset to learn an estimated model of the environment, and then learn an optimal policy for this model. This approach performs well when the dataset sufficiently explored the environment, e.g., (Agarwal et al., 2020a). However, under more general offline settings, the static dataset can be limited, which results in the distribution shift challenge and inaccurate model estimation (Kidambi et al., 2020; Ross & Bagnell, 2012; Li et al., 2022). Namely, the pre-collected dataset is often restricted to a small subset of state-action pairs, and the behavior policy used to collect the dataset induces a state-action visitation distribution that is different from the one induced by the optimal policy. This distribution shift and the limited amount of data lead to uncertainty in the estimation of the model, i.e., transition kernel and/or reward function.

To address the above challenge, one natural approach is to first quantify the uncertainty, and further take a pessimistic (conservative) approach in face of such uncertainty. Despite of the fact that the uncertainty

exists in the transition kernel estimate, existing studies mostly take the approach to penalize the reward function for less-visited state-action pairs to obtain a pessimistic estimation of the value function, known as the Lower Confidence Bound (LCB) approach (Rashidinejad et al., 2021; Li et al., 2022; Shi et al., 2022; Yan et al., 2022). In this paper, we develop a direct approach to analyzing such uncertainty in the transition kernel by constructing a statistically plausible set of transition kernels, i.e., uncertainty set, and optimizing the worst-case performance over this uncertainty set. This principle is referred to as "distributionally robust optimization (DRO)" in the literature (Nilim & El Ghaoui, 2004; Iyengar, 2005). This DRO-based approach directly tackles the uncertainty in the transition kernel. We show that our approach asymptotically achieves the minimax optimal sample complexity (Rashidinejad et al., 2021). We summarize our major contributions as follows.

1.1 Main Contributions

In this work, we focus on the most general partial coverage setting (see Section 2.3 for the definition). We develop a DRO-based framework that efficiently solves the offline RL problem. More importantly, we design a Bernstein-style uncertainty set and show that its sample complexity is minimax optimal.

DRO-based Approach Solves Offline RL. We construct a Hoeffding-style uncertainty set centered at the empirical transition kernel to guarantee that with high probability, the true transition kernel lies within the uncertainty set. Then, optimizing the worst-case performance over the uncertainty set provides a lower bound on the performance under the true environment. Our uncertainty model enables easy solutions using the robust dynamic programming approach developed for robust MDP in (Iyengar, 2005; Nilim & El Ghaoui, 2004) within a polynomial computational complexity. We further show the sample complexity to achieve an ϵ-optimal policy using our approach is $\mathcal{O}\left(\frac{S^2 C^{\pi^*}}{(1-\gamma)^4\epsilon^2}\right)$ (up to a log factor), where γ is the discount factor, and C^{π^*} is the single-policy concentrability for any comparator policy π^* (see Definition 1). This sample complexity matches with the best-known complexity of the Hoeffding-style model-uncertainty method (Rashidinejad et al., 2021; Uehara & Sun, 2021), which demonstrates that our DRO framework can directly tackle the model uncertainty and effectively solve offline RL.

Achieving Asymptotic Minimax Optimality via Design of Bernstein-style Uncertainty Set. While the approach described above is effective in achieving an ϵ-optimal policy with relatively low sample complexity, it tends to exhibit excessive conservatism as its complexity surpasses the minimax lower bound for offline RL algorithms (Rashidinejad et al., 2021) by a factor of $S(1-\gamma)^{-1}$. To close this gap, we discover that demanding the true transition kernel to be within the uncertainty set with high probability, i.e., the true environment and the worst-case one are close, can be overly pessimistic and unnecessary. What is of paramount importance is that the value function under the worst-case transition kernel within the uncertainty set (almost) lower bounds the one under the true transition kernel. Notably, this requirement is considerably less stringent than mandating that the actual kernel be encompassed by the uncertainty set. We then design a less conservative Bernstein-style uncertainty set, which has a smaller radius and thus is a subset of the Hoeffding-style uncertainty set. We prove that to obtain an ϵ-optimal policy, the order of the sample complexity is $\mathcal{O}\left(\frac{SC^{\pi^*}}{(1-\gamma)^3\epsilon^2}\right)$. This complexity indicates the asymptotic minimax optimality of our approach by matching with the minimax lower bound in asymptotic order of the sample complexity for offline RL (Rashidinejad et al., 2021) and the best result from the LCB approach (Li et al., 2022).

1.2 Related Works

There has been a proliferation of works on offline RL. In this section, we mainly discuss works on model-based approaches. There are also model-free approaches, e.g., (Liu et al., 2020; Kumar et al., 2020; Agarwal et al., 2020b; Yan et al., 2022; Shi et al., 2022), which are not the focus here.

Offline RL under global coverage. Existing studies on offline RL often make assumptions on the coverage of the dataset. This can be measured by the distribution shift between the behavior policy and the occupancy measure induced by the target policy, which is referred to as the concentrability coefficient (Munos, 2007; Rashidinejad et al., 2021). Many previous works, e.g., (Scherrer, 2014; Chen & Jiang, 2019; Jiang, 2019; Wang et al., 2019; Liao et al., 2020; Liu et al., 2019; Zhang et al., 2020a; Munos & Szepesvari, 2008; Uehara

et al., 2020; Duan et al., 2020; Xie & Jiang, 2020; Levine et al., 2020; Antos et al., 2007; Farahmand et al., 2010), assume that the density ratio between the above two distributions is finite for all state-action pairs and policies, which is known as global coverage. This assumption essentially requires the behavior policy to be able to visit all possible state-action pairs, which is often violated in practice (Gulcehre et al., 2020; Agarwal et al., 2020b; Fu et al., 2020).

Offline RL under partial coverage. Recent studies relax the above assumption of global coverage to partial coverage or single-policy concentrability. Partial coverage assumes that the density ratio between the distributions induced by a single target policy and the behavior policy is bounded for all state-action pairs. Therefore, this assumption does not require the behavior policy to be able to visit all possible state-action pairs, as long as it can visit those state-actions pairs that the target policy will visit. This partial coverage assumption is more feasible and applicable in real-world scenarios. In this paper, we focus on this practical partial coverage setting. Existing approaches under the partial coverage assumption can be divided into three categories as follows.

Regularized Policy Search. The first approach regularizes the policy so that the learned policy is close to the behavior policy (Fujimoto et al., 2019b; Wu et al., 2019; Jaques et al., 2019; Peng et al., 2019; Siegel et al., 2020; Wang et al., 2020; Kumar et al., 2019; Fujimoto et al., 2019a; Ghasemipour et al., 2020; Nachum et al., 2019; Zhang et al., 2020b; 2023). Thus, the learned policy is similar to the behavior policy which generates the dataset, hence this approach works well when the dataset is collected from experts (Wu et al., 2019; Fu et al., 2020).

Reward Penalization or LCB Approaches. One of the most widely used approaches is to penalize the reward in face of uncertainty to obtain an pessimistic estimation that lower bounds the real value function, e.g., (Kidambi et al., 2020; Yu et al., 2020; 2021b; Buckman et al., 2020; Jin et al., 2021; Xie et al., 2021b; Yin & Wang, 2021; Liu et al., 2020; Cui & Du, 2022; Chen et al., 2021; Zhong et al., 2022). The most popular approach VI-LCB (Rashidinejad et al., 2021; Li et al., 2022) penalizes the reward with a bonus term that is inversely proportional to the number of samples. The tightest sample complexity is obtained in (Li et al., 2022) by designing a Bernstein-style penalty term, which matches the minimax lower bound in (Rashidinejad et al., 2021).

DRO-based Approaches. Another approach is to first construct a set of "statistically plausible" MDP models based on the empirical transition kernel, and then find the policy that optimizes the worst-case performance over this set (Zanette et al., 2021; Uehara & Sun, 2021; Rigter et al., 2022; Bhardwaj et al., 2023; Guo et al., 2022; Hong et al., 2023; Chang et al., 2021; Blanchet et al., 2023). However, finding such a policy under the models proposed in these works can be NP-hard, hence some heuristic approximations without theoretical optimality guarantee are used to deploy their approaches. Our work falls into this category, but the computational complexity is polynomial, and the sample complexity of our approach is minimax optimal. A recent work (Panaganti et al., 2023) also proposes a similar Hoeffding-style DRO framework as ours, and their sample complexity results match ours in the first part, but fails to obtain the minimax optimality as our second part.

Offline RL with function approximation. There is another substantial interest in combining offline reinforcement learning with function approximation (e.g., deep neural networks) in order to encode inductive biases and enable generalization across large, potentially continuous state spaces, with recent progress on both model-free and model-based approaches (Ross & Bagnell, 2012; Laroche et al., 2019; Fujimoto et al., 2019b; Kumar et al., 2019; Agarwal et al., 2020b). Besides the aforementioned challenges, an additional issue regrading the representational conditions of the function class is introduced, which assert that the function approximator is flexible enough to represent value functions induced by certain policies. A huge body of recent works aim to understanding the condition for the function classes, including Bellman completeness condition (Liu et al., 2020; Jin et al., 2021; Xie et al., 2021a; Yin et al., 2021; Rashidinejad et al., 2021), or Bellman realizability (Xie & Jiang, 2021; Chen & Jiang, 2022; Rashidinejad et al., 2022; Jiang & Huang, 2020). However, in this paper, we mainly consider the tabular setting, where no requirement on the function class is needed.

Robust RL with distributional uncertainty. In this paper, our algorithm is based on the framework of robust MDP (Iyengar, 2005; Nilim & El Ghaoui, 2004; Bagnell et al., 2001; Satia & Lave Jr, 1973; Wiesemann et al., 2013), which finds the policy with the best worst-case performance over an uncertainty

set of transition dynamics. When the uncertainty set is fully known, the problem can be solved by robust dynamic programming. The sample complexity of model-based approaches without full knowledge of the uncertainty sets were studied in, e.g., (Yang et al., 2021; Xu et al., 2023; Panaganti & Kalathil, 2022; Shi et al., 2023; Panaganti & Kalathil, 2022), where a generative model is typically assumed. This model-based approach is further adapted to the robust offline setting in (Panaganti et al., 2022; Shi & Chi, 2022). Yet in these works, the challenge of partial coverage is addressed using the LCB aproach, i.e., penalizing the reward functions, whereas we show that the DRO framework itself can also address the challenge of partial coverage in the offline setting.

2 Preliminaries

2.1 Markov Decision Process (MDP)

An MDP can be characterized by a tuple $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $\mathsf{P} = \{\mathsf{P}_s^a \in \Delta(\mathcal{S}), a \in \mathcal{A}, s \in \mathcal{S}\}$[1] is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function[2], and $\gamma \in [0, 1)$ is the discount factor. Specifically, $\mathsf{P}_s^a = (p_{s,s'}^a)_{s' \in \mathcal{S}}$, where $p_{s,s'}^a$ denotes the probability that the environment transits to state s' if taking action a at state s. The reward of taking action a at state s is given by $r(s, a)$. A stationary policy π is a mapping from $\mathcal{S}$ to a distribution over $\mathcal{A}$, which indicates the probabilities of the agent taking actions at each state. At each time t, an agent takes an action $a_t \sim \pi(s_t)$ at state s_t, the environment then transits to the next state s_{t+1} with probability $p_{s_t,s_{t+1}}^{a_t}$, and the agent receives reward $r(s_t, a_t)$.

The value function of a policy π starting from any initial state $s \in \mathcal{S}$ is defined as the expected accumulated discounted reward by following π: $V_{\mathsf{P}}^{\pi}(s) \triangleq \mathbb{E}_{\mathsf{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) | S_0 = s, \pi\right]$, where $\mathbb{E}_{\mathsf{P}}$ denotes the expectation when the state transits according to P. Let ρ denote the initial state distribution, and denote the value function under the initial distribution ρ by $V_{\mathsf{P}}^{\pi}(\rho) \triangleq \mathbb{E}_{s\sim\rho}[V_{\mathsf{P}}^{\pi}(s)]$.

2.2 Robust Markov Decision Process

In the robust MDP, the transition kernel is not fixed and lies in some uncertainty set $\mathcal{P}$. Define the robust value function of a policy π as the worst-case expected accumulated discounted reward over the uncertainty set: $V_{\mathcal{P}}^{\pi}(s) \triangleq \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) | S_0 = s, \pi\right]$. Similarly, the robust action-value function for a policy π is defined as $Q_{\mathcal{P}}^{\pi}(s, a) = \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) | S_0 = s, A_0 = a, \pi\right]$. The goal of robust RL is to find the optimal robust policy that maximizes the worst-case accumulated discounted reward, i.e., $\pi_r = \arg\max_{\pi} V_{\mathcal{P}}^{\pi}(s), \forall s \in \mathcal{S}$. It is shown in (Iyengar, 2005; Nilim & El Ghaoui, 2004; Wiesemann et al., 2013) that the optimal robust value function is the unique solution to the optimal robust Bellman equation $V_{\mathcal{P}}^{\pi_r}(s) = \max_a\{r(s, a) + \gamma\sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^{\pi_r})\}$, where $\sigma_{\mathcal{P}_s^a}(V) \triangleq \min_{p\in\mathcal{P}_s^a} p^\top V$ denotes the support function of V on a set $\mathcal{P}_s^a$ and the corresponding robust Bellman operator is a γ-contraction.

2.3 Offline Reinforcement Learning

Under the offline setting, the agent cannot interact with the MDP and instead is given a pre-collected dataset $\mathcal{D}$ consisting of N tuples $\{(s_i, a_i, s_i', r_i) : i = 1, ..., N\}$, where $r_i = r(s_i, a_i)$ is the deterministic reward, and $s_i' \sim \mathsf{P}_{s_i}^{a_i}$ follows the transition kernel P of the MDP. The (s_i, a_i) pairs in $\mathcal{D}$ are generated i.i.d. according to an unknown data distribution μ over the state-action space. In this paper, we consider the setting where the reward functions r is deterministic but unknown. We denote the number of samples transitions from (s, a) in $\mathcal{D}$ by $N(s, a)$, i.e., $N(s, a) = \sum_{i=1}^{N} \mathbf{1}_{(s_i,a_i)=(s,a)}$ and $\mathbf{1}_{X=x}$ is the indicator function.

The goal of offline RL is to find a policy π which optimizes the value function V_{P}^{π} based on the offline dataset $\mathcal{D}$. Let d^{π} denote the discounted occupancy distribution associated with π: $d^{\pi}(s) = (1 - \gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(S_t = s | S_0 \sim \rho, \pi, \mathsf{P})$. In this paper, we focus on the partial coverage setting and adopt the following definition from

[1] $\Delta(\mathcal{S})$ denotes the probability simplex defined on $\mathcal{S}$.

[2] To streamline our presentation and emphasize our novelty in solving offline RL through DRO formulation, we assume the reward function is deterministic. However, our DRO approach directly extends to the stochastic reward setting by constructing a similar uncertainty set for the reward estimation.

(Li et al., 2022) to measure the distribution shift between the dataset distribution and the occupancy measure induced by a single policy π^*:

Definition 1. *(Single-policy clipped concentrability) The single-policy clipped concentrability coefficient of a policy π^* is defined as*

$$C^{\pi^*} \triangleq \max_{s,a} \frac{\min\{d^{\pi^*}(s,a), \frac{1}{S}\}}{\mu(s,a)}, \tag{1}$$

where $S \triangleq |\mathcal{S}|$ denotes the number of states.

We note another unclipped version of C^{π^*} is also commonly used in the literature, e.g., (Rashidinejad et al., 2021; Uehara & Sun, 2021), defined as $\tilde{C}^{\pi^*} \triangleq \max_{s,a} \frac{d^{\pi^*}(s,a)}{\mu(s,a)}$. It is straightforward to verify that $C^{\pi^*} \leq \tilde{C}^{\pi^*}$, and all of our results remain valid if C^{π^*} is replaced by $\tilde{C}^{\pi^*}$. A more detailed discussion can be found in (Li et al., 2022; Shi & Chi, 2022).

The goal of this paper is to find a policy π which minimizes the sub-optimality gap compared to a comparator policy π^* under some initial state distribution ρ: $V_{\mathsf{P}}^{\pi^*}(\rho) - V_{\mathsf{P}}^{\pi}(\rho)$.

3 Offline RL via Distributionally Robust Optimization

Model-based methods usually commence by estimating the transition kernel employing its maximum likelihood estimate. Nevertheless, owing to the inherent challenges associated with distribution shift and limited data in the offline setting, uncertainties can arise in these estimations. For instance, the dataset may not encompass every state-action pair, and the sample size may be insufficient to yield a precise estimate of the transition kernel. In this paper, we directly quantify the uncertainty in the empirical estimation of the transition kernel, and construct a set of "statistically possible" transition kernels, referred to as uncertainty set, so that it encompasses the actual environment. We then employ the DRO approach to optimize the worst-case performance over the uncertainty set. Notably, this formulation essentially transforms the problem into a robust Markov Decision Process (MDP), as discussed in Section 2.2.

In Section 3.1,we first introduce a direct metric-based Hoeffding-style approach to construct the uncertainty set such that the true transition kernel is in the uncertainty set with high probability. We then present the robust value iteration algorithm to solve the DRO problem. We further theoretically characterize the bound on the sub-optimality gap and show that the sample complexity to achieve an ϵ-optimality gap is $\mathcal{O}((1-\gamma)^{-4}\epsilon^{-2}SC^{\pi^*})$. This gap matches with the best-known sample complexity for the LCB method using the Hoeffding-style bonus term (Rashidinejad et al., 2021). This result shows the effectiveness of our DRO-based approach in solving the offline RL problem.

We then design a less conservative Bernstein-style uncertainty set aiming to achieve the minimax optimal sample complexity in Section 3.2. We theoretically establish that our approach attains an enhanced and asymptotically minimax optimal sample complexity of $\mathcal{O}\left((1-\gamma)^{-3}\epsilon^{-2}SC^{\pi^*}\right)$. Notably, this sample complexity matches with the minimax lower bound (Rashidinejad et al., 2021) and stands on par with the best results achieved using the LCB approach (Li et al., 2022) in asymptotic order.

Our approach starts with learning an empirical model of the transition kernel and reward from the dataset as follows. These empirical transition kernels will be used as the centroid of the uncertainty set. For any (s,a), if $N(s,a) > 0$, set

$$\hat{\mathsf{P}}^a_{s,s'} = \frac{\sum_{i \leq N} \mathbf{1}_{(s_i,a_i,s_i')=(s,a,s')}}{N(s,a)}, \hat{r}(s,a) = r_i(s,a); \tag{2}$$

And if $N(s,a) = 0$, set

$$\hat{\mathsf{P}}^a_{s,s'} = \mathbf{1}_{s'=s}, \hat{r}(s,a) = 0. \tag{3}$$

The MDP $\hat{\mathsf{M}} = (\mathcal{S}, \mathcal{A}, \hat{\mathsf{P}}, \hat{r})$ with the empirical transition kernel $\hat{\mathsf{P}}$ and empirical reward $\hat{r}$ is referred to as the empirical MDP. For unseen state-action pairs (s,a) in the offline dataset, we take a conservative approach

and let the estimated $\hat{r}(s,a)=0$, and set s as an absorbing state if taking action a. Then the action value function at (s,a) for the empirical MDP shall be zero, which discourages the choice of action a at state s.

For each state-action pair (s,a), we construct an uncertainty set centered at the empirical transition kernel $\hat{\mathsf{P}}_s^a$, with a radius inversely proportional to the number of samples in the dataset. Specifically, set the uncertainty set as $\hat{\mathcal{P}}=\bigotimes_{s,a}\hat{\mathcal{P}}_s^a$[3] and

$$\hat{\mathcal{P}}_s^a=\left\{q\in\Delta(\mathcal{S}):D(q,\hat{\mathsf{P}}_s^a)\leq R_s^a\right\}, \tag{4}$$

where $D(\cdot,\cdot)$ is some function that measures the difference between two probability distributions, e.g., total variation, Chi-square divergence, R_s^a is the radius ensuring that the uncertainty set adapts to the dataset size and the degree of confidence, which will be determined later. We then construct the robust MDP as $\hat{\mathcal{M}}=(\mathcal{S},\mathcal{A},\hat{\mathcal{P}},\hat{r})$.

As we shall show later, the optimal robust policy

$$\pi_r=\arg\max_{\pi}\min_{\mathsf{P}\in\hat{\mathcal{P}}}V_{\mathsf{P}}^{\pi}(s),\forall s\in\mathcal{S}. \tag{5}$$

w.r.t. $\hat{\mathcal{M}}$ performs well in the real environment and reaches a small sub-optimality gap. In our construction in eq. (4), the uncertainty set is (s,a)-rectangular, i.e., for different state-action pairs, the corresponding uncertainty sets are independent. With this rectangular structure, the optimal robust policy can be found by utilizing the robust value iteration algorithm or robust dynamic programming (Algorithm 1), and the corresponding robust value iteration at each step can be solved in polynomial time (Wiesemann et al., 2013). In contrast, the uncertainty set constructed in (Uehara & Sun, 2021; Bhardwaj et al., 2023), defined as $\mathcal{T}=\{\mathsf{Q}:\mathbb{E}_{\mathcal{D}}[\|\hat{\mathsf{P}}_s^a-\mathsf{Q}_s^a\|^2]\leq\zeta\}$, does not enjoy such a rectangularity. Solving a robust MDP with such an uncertainty set can be, however, NP-hard (Wiesemann et al., 2013). The approach developed in (Rigter et al., 2022) to solve it is based on the heuristic approach of adversarial training, and therefore is lack of theoretical guarantee.

Algorithm 1 Robust Value Iteration (Nilim & El Ghaoui, 2004; Iyengar, 2005)

INPUT: $\hat{r},\hat{\mathcal{P}},V,\mathcal{D}$

1: **while** TRUE **do**
2: **for** $s\in\mathcal{S}$ **do**
3: $V(s)\leftarrow\max_a\{\hat{r}(s,a)+\gamma\sigma_{\hat{\mathcal{P}}_s^a}(V)\}$
4: **end for**
5: **end while**
6: **for** $s\in\mathcal{S}$ **do**
7: $\pi_r(s)\in\arg\max_{a\in\mathcal{A}}\{\hat{r}(s,a)+\gamma\sigma_{\hat{\mathcal{P}}_s^a}(V)\}\}$
8: **end for**

Output: π_r

The algorithm converges to the optimal robust policy linearly since the robust Bellman operator is a γ-contraction (Iyengar, 2005). The computational complexity of the support function $\sigma_{\hat{\mathcal{P}}_s^a}(V)$ in Lines 3 and 7 w.r.t. the uncertainty sets we constructed matches the ones of the LCB approaches (Rashidinejad et al., 2021; Li et al., 2022).

In the following two sections, we specify the constructions of the uncertainty sets.

3.1 Hoeffding-style Radius

We first employ the total variation to construct this uncertainty set. Specifically, we let D be the total variation distance and $R_s^a\triangleq\min\left\{1,\sqrt{\frac{S\log\frac{SA}{\delta}}{8N(s,a)}}\right\}$. With our design, fewer samples result in a larger uncertainty

[3]Here, $\bigotimes_{s,a}$ denotes the Cartesian product of all state-action pairs, i.e., the uncertainty set $\hat{\mathcal{P}}$ is independently defined for each state-action pairs. This structure is also known as the (s,a)-rectangular uncertainty set.

set and imply that we should be more conservative in estimating the transition dynamics at this state-action pair. Other distance function of D can also be used, contingent upon the concentration inequality being applied.

In Algorithm 1, $\sigma_{\hat{\mathcal{P}}_s^a}(V) = \min_{q \in \hat{\mathcal{P}}_s^a}\{q^\top V\}$ can be equivalently solved by solving its dual form (Iyengar, 2005), which is a convex optimization problem: $\max_{0 \le \mu \le V}\{\hat{\mathsf{P}}_s^a(V-\mu) - R_s^a \mathrm{Span}(V-\mu)\}$, and $\mathrm{Span}(X) = \max_i X(i) - \min_i X(i)$ is the span semi-norm of vector X. The computational complexity associated with solving it is $\mathcal{O}(S\log(S))$. Notably, this polynomial computational complexity is on par with the complexity of the VI-LCB approach (Li et al., 2022).

We then show that with this Hoeffding-style radius, the true transition kernel falls into the uncertainty set with high probability.

Lemma 1. *With probability at least $1-\delta$, it holds that for any s, a, $\mathsf{P}_s^a \in \hat{\mathcal{P}}_s^a$, i.e., $\|\mathsf{P}_s^a - \hat{\mathsf{P}}_s^a\| \le 2R_s^a$*[4]*.*

This result implies that the real environment P falls into the uncertainty set $\hat{\mathcal{P}}$ with high probability, and hence finding the optimal robust policy of $\hat{\mathcal{M}}$ provides a worst-case performance guarantee. We further present our result of the sub-optimality gap in the following theorem.

Theorem 1. *Consider an arbitrary deterministic comparator policy π^*. With probability at least $1-2\delta$, the output policy π_r of Algorithm 1 satisfies*

$$V_{\mathsf{P}}^{\pi^*}(\rho) - V_{\mathsf{P}}^{\pi_r}(\rho) \le \frac{16SC^{\pi^*}\log\frac{NS}{\delta}}{(1-\gamma)^2 N} + \sqrt{\frac{96S^2C^{\pi^*}\log\frac{SA}{\delta}}{(1-\gamma)^4 N}}. \tag{6}$$

To achieve an ϵ-optimality gap, a dataset of size $N = \mathcal{O}\left((1-\gamma)^{-4}\epsilon^{-2}S^2C^{\pi^*}\right)$ is required for a Hoeffding-style uncertainty model. This sample complexity matches with the best-known sample complexity for LCB methods with Hoeffding-style bonus term (Rashidinejad et al., 2021) and model uncertainty type approach (Panaganti et al., 2023). It suggests that our DRO-based approach can effectively address the offline RL problem.

However, there is still a gap between this sample complexity and the minimax lower bound in (Rashidinejad et al., 2021) and the best-known sample complexity of LCB-based method (Li et al., 2022), which is $\mathcal{O}\left((1-\gamma)^{-3}\epsilon^{-2}SC^{\pi^*}\right)$. We will address this problem via a Bernstein-style uncertainty set design in the next subsection.

Remark 1. *The choice of total variation and radius when construct the uncertainty set and obtain the results are not essential. We can also use alternative distance functions or divergence, and set radius according to the corresponding concentration inequalities, e.g., including Chi-square divergence, KL-divergence and Wasserstein distance (Canonne, 2020; Bhandari & Russo, 2021; Arora et al., 2023). Our results and methods can be further generalized to large-scale problems when a low-dimensional latent representation is presented, e.g., linear MDPs or low-rank MDPs (Panaganti et al., 2023).*

3.2 Bernstein-style Radius

As discussed above, using a Hoeffding-style radius is able to achieve an ϵ-optimal policy, however, with an unnecessarily large sample complexity. Compared with the minimax lower bound and the tightest result obtained in (Li et al., 2022), there exists a gap of order $\mathcal{O}(S(1-\gamma)^{-1})$. This gap is mainly because the Hoeffding-style radius is overly conservative and the bound is loose. Specifically, Hoeffding-style approach can be viewed as *distribution based*. That is, to construct the uncertainty set $\hat{\mathcal{P}}$ centered at $\hat{\mathsf{P}}$ large enough such that the true transition kernel P falls into $\hat{\mathcal{P}}$ with high probability (Lemma 1). Therefore, it holds that $V_{\hat{\mathcal{P}}}^{\pi_r} \le V_{\mathsf{P}}^{\pi_r}$ and the sub-optimality gap can be bounded as

$$V_{\mathsf{P}}^{\pi^*} - V_{\mathsf{P}}^{\pi_r} = \underbrace{V_{\mathsf{P}}^{\pi^*} - V_{\hat{\mathcal{P}}}^{\pi_r}}_{\Delta_1} + \underbrace{V_{\hat{\mathcal{P}}}^{\pi_r} - V_{\mathsf{P}}^{\pi_r}}_{\Delta_2 \le 0} \le \Delta_1. \tag{7}$$

[4] In this paper, unless stated otherwise, we denote the l_1-norm by $\|\cdot\|$.

To further bound Δ_1, we utilize the distance between the two transition kernels which is upper bounded by the radius R of the uncertainty set, and obtain a sub-optimal bound of order $\mathcal{O}\left(\frac{1}{\sqrt{N(1-\gamma)^4}}\right)$.

This result can be improved from two aspects. Firstly, we note that the uncertainty set under the Hoeffding-style construction is too large to include P with high probability. Although this construction implies $\Delta_2 \leq 0$, as the price of it, the large radius implies a loose bound on Δ_1. Another observation is that both two terms are in fact the differences between the *expectations* under two different distributions. Instead of merely using the distance of the two distributions to bound them, we can utilize other tighter concentration inequalities like Bernstein's inequality, to obtain an involved but tighter bound.

Toward this goal, we construct a smaller and less conservative uncertainty set such that: (1). It implies a tighter bound on Δ_1 combining with Bernstein's inequality; And (2). Although non-zero, the term Δ_2 can also be tightly bounded. Specifically, note that $\Delta_2 = \underbrace{V_{\hat{\mathcal{P}}}^{\pi_r} - V_{\hat{\mathsf{P}}}^{\pi_r}}_{(a)} + \underbrace{V_{\hat{\mathsf{P}}}^{\pi_r} - V_{\mathsf{P}}^{\pi_r}}_{(b)}$. Term (b) can be viewed as an estimation error which is from the inaccurate estimation from the dataset; And Term (a) is the difference between robust value function and the value function under the centroid transition kernel $\hat{\mathsf{P}}$, which is always negative and can be bounded using the dual-form solutions for specific uncertainty sets (Iyengar, 2005). We hence choose a radius such that: (1). the negative bound on (a) cancels with the higher-order terms in the bound on (b) and further implies a tighter bound on Δ_2; And (2). the bound on Δ_1 is also tight by utilizing Bernstein's inequality.

We then construct the Bernstein-style uncertainty set as follows. Instead of the total variation, we construct the uncertainty set using the Chi-square divergence, i.e., $D(p,q) = \chi^2(p||q) = \sum_s q(s)\left(1 - \frac{p(s)}{q(s)}\right)^2$. The reason why we adopt the Chi-square divergence instead of the total variation will be discussed later. We design our uncertainty set for each state-action pair as follows. For (s,a) with $N(s,a) > 0$, we set the radius as $R_s^a \triangleq \frac{48 \log \frac{4N}{\delta}}{N(s,a)}$, and set $\hat{\mathcal{P}}_s^a \triangleq \{q \in \Delta(\mathcal{S}) : \chi^2(q||\hat{\mathsf{P}}_s^a) \leq R_s^a\}$; And for (s,a) pairs that are not covered by the dataset, we set $\hat{\mathcal{P}}_s^a \triangleq \Delta(\mathcal{S})$. We then construct the robust MDP as $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{\mathcal{P}} = \bigotimes_{s,a} \hat{\mathcal{P}}_s^a, \hat{r})$.

Remark 2. *From Pinsker's inequality and the fact that $D_{KL}(p||q) \leq \chi^2(p||q)$ (Nishiyama & Sason, 2020), it holds that $\|p - q\| \leq \sqrt{2\chi^2(p||q)}$. Hence the Bernstein-style uncertainty set is a subset of the Hoeffding-style uncertainty set in Section 3.1, and is less conservative.*

Similarly, we find the optimal robust policy w.r.t. the corresponding robust MDP $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, \hat{r})$ using the robust value iteration with a slight modification, which is presented in Algorithm 2. Specifically, the output

Algorithm 2 Robust Value Iteration

INPUT: $\hat{r}, \hat{\mathcal{P}}, V, \mathcal{D}$

```
while TRUE do
    for s ∈ S do
        N(s) ← Σ_{i=1}^{N} 1_{(s_i)=s}
        V(s) ← max_a{r̂(s,a) + γσ_{P̂_s^a}(V)}
    end for
end while
for s ∈ S do
    if N(s) > 0 then
        π_r(s) ∈ {arg max_{a∈A}{r̂(s,a) + γσ_{P̂_s^a}(V)}} ∩ {a : N(s,a) > 0}
    else
        π_r(s) ∈ arg max_{a∈A}{r̂(s,a) + γσ_{P̂_s^a}(V)}}
    end
end for
```

Output: π_r

policy π_r in Algorithm 2 is set to be the greedy policy satisfying $N(s,a) > 0$ if $N(s) > 0$. The existence of

such a policy is proved in Lemma 2 in the appendix. This is to guarantee that when there is a tie of taking greedy actions, we will take an action that has appeared in the pre-collected dataset $\mathcal{D}$.

The support function $\sigma_{\hat{\mathcal{P}}_s^a}(V)$ w.r.t. the Chi-square divergence uncertainty set can also be computed using its dual form (Iyengar, 2005): $\sigma_{\hat{\mathcal{P}}_s^a}(V) = \max_{\alpha\in[V_{\min},V_{\max}]}\{\hat{\mathsf{P}}_s^a V_\alpha - \sqrt{R_s^a \mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_\alpha)}\}$, where $V_\alpha(s) = \min\{\alpha, V(s)\}$. The dual form is also a convex optimization problem and can be solved efficiently within a polynomial time $\mathcal{O}(S\log S)$ (Iyengar, 2005).

Using the Chi-square divergence enables a smaller radius and yields a tighter bound on $\Delta_2 = (a) + (b)$. Namely, (b) can be bounded by a $N^{-0.5}$-order bound according to the Bernstein's inequality (see Lemma 5 in the Appendix). Simultaneously, our goal is to obtain a bound with the same order on (a), which effectively offsets the bound on (b), and yields a tighter bound on Δ_2. The robust value function w.r.t. the total variation uncertainty set, however, depends on R_s^a linearly (see the dual form we discussed above); On the other hand, the solution to the Chi-square divergence uncertainty set incorporates a term of $\sqrt{R_s^a}$ which enables us to set a lower-order radius (i.e., set $R_s^a = (N(s,a))^{-1}$) to offset the $N^{-0.5}$-order bound on (b).

We then characterize the optimality gap obtained from Algorithm 2 in the following theorem.

Theorem 2. *If $N \geq \frac{1}{(1-\gamma)KSC^{\pi^*}\mu_{\min}^2}$, then the output policy π_r of Algorithm 2 satisfies*

$$V_{\mathsf{P}}^{\pi^*}(\rho) - V_{\mathsf{P}}^{\pi_r}(\rho) \leq \sqrt{\frac{KSC^{\pi^*}\log\frac{4N}{\delta}}{(1-\gamma)^3 N}}, \tag{8}$$

with probability at least $1-4\delta$, where $\mu_{\min} = \min\{\mu(s,a) : \mu(s,a) > 0\}$ denotes the minimal non-zero probability of μ, and K is some universal constant that independent with S, γ, C^{π^*} and N.

Theorem 2 implies that our DRO approach can achieve an ϵ-optimality gap, as long as the size of the dataset exceeds the order of

$$\mathcal{O}\Bigg(\underbrace{\frac{SC^{\pi^*}}{(1-\gamma)^3\epsilon^2}}_{\epsilon\text{-dependent}} + \underbrace{\frac{1}{(1-\gamma)SC^{\pi^*}\mu_{\min}^2}}_{\text{burn-in cost}}\Bigg). \tag{9}$$

The burn-in cost term indicates that the asymptotic bound of the sample complexity becomes relevant after the dataset size surpasses the burn-in cost. It represents the minimal requirement for the amount of data. In fact, if the dataset is too small, we should not expect to learn a well-performed policy from it. In our case, if the dataset is generated under a generative model Panaganti et al. (2022); Yang et al. (2021); Shi et al. (2023) or uniform distribution, the burn-in cost term is in order of $\frac{SA^2}{1-\gamma}$. Burn-in cost also widely exists in the sample complexity studies of RL, e.g., $H^8SC^{\pi^*}$ in (Xie et al., 2021b), $\frac{S^3A^2}{(1-\gamma)^4}$ in (He et al., 2021), and $\frac{SC^{\pi^*}}{(1-\gamma)^5}$ in (Yan et al., 2022), $\frac{H}{\mu_{\min}p_{\min}}$ in (Shi & Chi, 2022). Note that the burn-in cost term is independent of the accuracy level ϵ, which implies the sample complexity is less than $\mathcal{O}\left(\frac{SC^{\pi^*}}{(1-\gamma)^3\epsilon^2}\right)$, as long as ϵ is small. This result matches the optimal asymptotic order of the complexity according to the minimax lower bound in (Rashidinejad et al., 2021), and also matches the tightest bound obtained using the LCB approach (Li et al., 2022) asymptotically. This suggests that our DRO approach can effectively address offline RL while imposing minimal demands on the dataset, thus optimizing the sample complexity associated with offline RL.

Remark 3. *We further discuss the major differences in our approach and the LCB approaches (Li et al., 2022; Rashidinejad et al., 2021). Firstly, the motivations are different. The LCB approach can be viewed as **'value function-based'**, which aims to obtain a pessimistic estimation of the value function by subtracting a penalty term from the reward, e.g., eq (83) in (Li et al., 2022). Our DRO approach aims to construct an uncertainty set that contains the statistically plausible transition dynamics, and optimize the worst-case performance among this uncertainty set. Our approach can be viewed as **'transition model-based'**, meaning that we directly tackle the uncertainty from the model estimation, without using the value function as an intermediate step.*

Our proof techniques are also different from the ones in LCB. To clarify the difference, we first rewrite our update rule using the LCB fashion. The update in robust value iteration Algorithm 1 can be written as

$$V(s) \leftarrow \max_a \{r(s,a) + \gamma \sigma_{\hat{\mathcal{P}}_s^a}(V)\} = \max_a \{r(s,a) + \gamma \hat{\mathsf{P}}_s^a V - b(s,a,V)\}, \tag{10}$$

where $b(s,a,V) \triangleq \gamma \hat{\mathsf{P}}_s^a V - \gamma \sigma_{\hat{\mathcal{P}}_s^a}(V)$, and hence, our algorithm bears formal resemblance to an LCB approach.

However, this formal similarity does not imply that our approach and results can be derived from those of LCB methods. Specifically, in LCB approaches, a crucial step involves the meticulous design of the penalty term $b(s,a,V)$ to satisfy the condition $\gamma|\hat{\mathsf{P}}_s^a V - \mathsf{P}_s^a V| \leq b(s,a,V)$, ensuring that the obtained estimation V remains pessimistic compared to the true value function. This often leads to the development of a complex penalty term, which may incorporate variance (e.g., see equation (28) in (Li et al., 2022)).

In contrast, within the DRO setting, designing an uncertainty set to uphold the aforementioned inequality either yields a loose bound on sub-optimality or results in a highly intricate uncertainty set. On one hand, enlarging the uncertainty set sufficiently to accommodate $\mathsf{P} \in \hat{\mathcal{P}}$ (to ensure $\mathsf{P}_s^a V \geq \min_{q \in \hat{\mathcal{P}}_s^a} qV = \sigma_{\hat{\mathcal{P}}_s^a}(V)$ and the inequality above) falls into the Hoeffding-style construction, leading to sub-optimal sample complexity. On the other hand, adopting a similar approach involving Bernstein inequality introduces dependence on V into the radius, resulting in a time-varying and intricate uncertainty set akin to the one in the LCB method. Our method circumvents these issues by not mandating the above inequality for leveraging the pessimism principle. Instead, we harness the inherent pessimism of the DRO setting and devise a straightforward yet effective uncertainty set. Although the above inequality may not hold, the robust value function we derive remains a pessimistic estimation of the true value function, demonstrating its effectiveness and efficiency.

This underscores the fundamental disparity in motivation and design between our method and LCB approaches, highlighting the novelty of our approach.

We present a comparative analysis of our results alongside those of the most closely related works in Table 1. As evidenced by the comparison, our approach stands out as the first model-uncertainty-based method to achieve the minimax optimal sample complexity in offline RL. Notably, we observe variations in the assumptions made across these works, particularly concerning the adoption of the clipped version of the single-policy concentrability. We meticulously specify these variations in our comparison. However, as previously discussed, the single-policy clipped concentrability is inherently smaller than its unclipped counterpart. Consequently, our assumption is comparatively weaker, enabling our results to directly follow if we adopt the unclipped assumption.

	Approach Type	Assumption	Asymptotic Sample Complexity	Computational Complexity
Our Approach	DRO	Single-policy, clipped	$\mathcal{O}\left(\frac{SC^{\pi^*}}{\epsilon^2(1-\gamma)^3}\right)$	Polynomial
(Rashidinejad et al., 2021)	LCB	Single-policy	$\mathcal{O}\left(\frac{SC^{\pi^*}}{\epsilon^2(1-\gamma)^5}\right)$	Polynomial
(Uehara & Sun, 2021)	DRO	Single-policy	$\mathcal{O}\left(\frac{S^2C^{\pi^*}}{\epsilon^2(1-\gamma)^4}\right)$	NP-Hard
(Panaganti et al., 2023)	DRO	Single-policy	$\mathcal{O}\left(\frac{S^2C^{\pi^*}}{\epsilon^2(1-\gamma)^4}\right)$	Polynomial
(Li et al., 2022)	LCB	Single-policy, clipped	$\mathcal{O}\left(\frac{SC^{\pi^*}}{\epsilon^2(1-\gamma)^3}\right)$	Polynomial
(Rashidinejad et al., 2021)	Minimax Lower bound	Single-policy	$\mathcal{O}\left(\frac{SC^{\pi^*}}{\epsilon^2(1-\gamma)^3}\right)$	-

Table 1: Comparison with related works.

4 Experiments

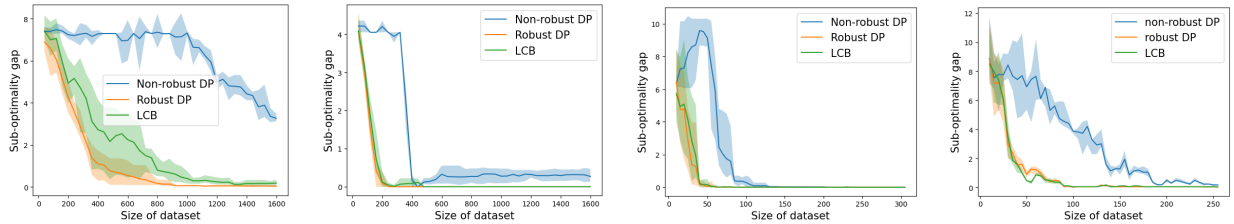
We adapt our DRO framework under two problems, the Garnet problem $\mathcal{G}(30, 20)$ (Archibald et al., 1995), and the Frozen-Lake problem (Brockman et al., 2016) to numerically verify our results.

In the Garnet problem, $|\mathcal{S}| = 30$ and $|\mathcal{A}| = 20$. The transition kernel $\mathsf{P} = \{\mathsf{P}_s^a, s \in \mathcal{S}, a \in \mathcal{A}\}$ is randomly generated following a normal distribution: $\mathsf{P}_s^a \sim \mathcal{N}(\omega_s^a, \sigma_s^a)$ and then normalized, and the reward function $r(s, a) \sim \mathcal{N}(\nu_s^a, \psi_s^a)$, where $\omega_s^a, \sigma_s^a, \nu_s^a, \psi_s^a \sim \mathbf{Uniform}[0, 100]$.

In the Frozen-Lake problem, an agent aim to cross a 4×4 frozen lake from Start to Goal without falling into any Holes by walking over the frozen lake.

In both problems, we deploy our approach under both global coverage and partial coverage conditions. Specifically, under the global coverage setting, the dataset is generated by the uniform policy $\pi(a|s) = \frac{1}{|\mathcal{A}|}$; And under the partial coverage condition, the dataset is generated according to $\mu(s, a) = \frac{\mathbf{1}_{a=\pi^*(s)}}{2} + \frac{\mathbf{1}_{a=\eta}}{2}$, where η is an action randomly chosen from the action space $\mathcal{A}$.

At each time step, we generate 40 new samples and add them to the offline dataset and deploy our DRO approach on it. We also deploy the LCB approach (Li et al., 2022) and non-robust model-based dynamic programming as the baselines. We run the algorithms independently 10 times and plot the average value of the sub-optimality gaps over all 10 trajectories. We also plot the 95th and 5th percentiles of the 10 curves as the upper and lower envelopes of the curves. The results are presented in Figure 1. It can be seen from the results that our DRO approach finds the optimal policy with relatively less data; The LCB approach has a similar convergence rate to the optimal policy, which verifies our theoretical results; The non-robust DP converges much slower, and can even converge to a sub-optimal policy. The results hence demonstrate the effectiveness and efficiency of our DRO approach.



(a) Garnet Problem under global coverage (b) Garnet Problem under partial coverage (c) Frozen-Lake under global coverage (d) Frozen-Lake under partial coverage

Figure 1: Sub-optimality gaps of Robust DP, LCB approach, and Non-robust DP.

5 Conclusion

In this paper, we revisit the problem of offline reinforcement learning from a novel angle of distributional robustness. We develop a DRO-based approach to solve offline reinforcement learning. Our approach directly incorporates conservatism in estimating the transition dynamics instead of penalizing the reward of less-visited state-action pairs. Our algorithms are based on the robust dynamic programming approach, which is computationally efficient. We focus on the challenging partial coverage setting, and develop two uncertainty sets: the Hoeffding-style and the less conservative Bernstein-style. For the Hoeffding-style uncertainty set, we theoretically characterize its sample complexity, and show it matches with the best one of the LCB-based approaches using the Hoeffding-style bonus term. For the Bernstein-style uncertainty set, we show its sample complexity is asymptotically minimax optimal, with a constant-level burn-in cost. Our results provide a DRO-based framework, an alternative approach to efficiently and effectively solve the problem of offline reinforcement learning.

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020a.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020b.

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted Q-iteration in continuous action-space MDPs. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 20, pp. 9–16, 2007.

TW Archibald, KIM McKinnon, and LC Thomas. On the generation of Markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

Vipul Arora, Arnab Bhattacharyya, Clément L Canonne, and Joy Qiping Yang. Near-optimal degree testing for bayes nets. *arXiv preprint arXiv:2304.06733*, 2023.

J Andrew Bagnell, Andrew Y Ng, and Jeff G Schneider. Solving uncertain Markov decision processes. 09 2001.

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite MDPs. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pp. 2386–2394. PMLR, 2021.

Mohak Bhardwaj, Tengyang Xie, Byron Boots, Nan Jiang, and Ching-An Cheng. Adversarial model for offline reinforcement learning. *arXiv preprint arXiv:2302.11048*, 2023.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659*, 2023.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

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

Clément L Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

Jonathan D Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. Mitigating covariate shift in imitation learning via offline data without great coverage. *arXiv preprint arXiv:2106.03207*, 2021.

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

Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. In *Uncertainty in Artificial Intelligence*, pp. 378–388. PMLR, 2022.

Minshuo Chen, Yan Li, Ethan Wang, Zhuoran Yang, Zhaoran Wang, and Tuo Zhao. Pessimism meets invariance: Provably efficient offline mean-field multi-agent RL. *Advances in Neural Information Processing Systems*, 34, 2021.

Qiwen Cui and Simon S Du. When is offline two-player zero-sum Markov game solvable? *arXiv preprint arXiv:2201.03522*, 2022.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 2701–2709. PMLR, 2020.

Amir Massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto, Edoardo Conti, Mohammad Ghavamzadeh, and Joelle Pineau. Benchmarking batch deep reinforcement learning algorithms. *arXiv preprint arXiv:1910.01708*, 2019a.

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

Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, and Shixiang Shane Gu. EMaQ: Expected-max Q-learning operator for simple yet effective offline and online RL. *arXiv preprint arXiv:2007.11091*, 2020.

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gómez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, et al. RL unplugged: Benchmarks for offline reinforcement learning. *arXiv preprint arXiv:2006.13888*, 2020.

Kaiyang Guo, Shao Yunfeng, and Yanhui Geng. Model-based offline reinforcement learning with pessimism-modulated dynamics belief. *Advances in Neural Information Processing Systems*, 35:449–461, 2022.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted MDPs. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021.

Mao Hong, Yue Wu, and Yanxun Xu. Pessimistic model-based actor-critic for offline reinforcement learning: Theory and algorithms, 2023. URL `https://openreview.net/forum?id=VQfSsOTrLIy`.

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

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Nan Jiang. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758, 2020.

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

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

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*, 2019.

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

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pp. 3652–3661. PMLR, 2019.

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

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *in preparation*, 2022.

Peng Liao, Zhengling Qi, and Susan Murphy. Batch policy learning in average reward Markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

R. Munos and C. Szepesvari. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, May 2008.

Rémi Munos. Performance bounds in ℓ_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

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

Arnab Nilim and Laurent El Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 839–846, 2004.

Tomohiro Nishiyama and Igal Sason. On relations between the relative entropy and χ 2-divergence, generalizations and applications. *Entropy*, 22(5):563, 2020.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*, 2022.

Kishan Panaganti, Xu Zaiyan, Kalathil Dileep, and Ghavamzadeh Mohammad. Bridging distributionally robust learning and offline rl: An approach to mitigate distribution shift and partial data coverage. *arXiv preprint arXiv:2310.18434*, 2023.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

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

Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*, 2022.

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

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

Jay K Satia and Roy E Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1314–1322, 2014.

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

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. *International Conference on Machine Learning*, pp. 19967–20025, 2022.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

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

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

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S Merel, Jost Tobias Springenberg, Scott E Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. *Advances in Neural Information Processing Systems*, 33, 2020.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

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

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

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

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

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34: 27395–27407, 2021b.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 9728–9754. PMLR, 2023.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*, 2022.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1567–1575. PMLR, 2021.

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

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

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. COMBO: Conservative offline model-based policy optimization. *arXiv preprint arXiv:2102.08363*, 2021b.

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

Junjie Zhang, Jiafei Lyu, Xiaoteng Ma, Jiangpeng Yan, Jun Yang, Le Wan, and Xiu Li. Uncertainty-driven trajectory truncation for model-based offline reinforcement learning. *arXiv preprint arXiv:2304.04660*, 2023.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4572–4583, 2020a.

Shantong Zhang, Bo Liu, and Shimon Whiteson. GradientDICE: Rethinking generalized offline estimation of stationary values. *arXiv preprint arXiv:2001.11113*, 2020b.

Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. *arXiv preprint arXiv:2202.07511*, 2022.

A Notations

We first introduce some notations that are used in our proofs. We denote the numbers of states and actions by S, A, i.e., $|\mathcal{S}| = S$, $|\mathcal{A}| = A$. For a transition kernel P and a policy π, P^π denotes the transition matrix induced by them, i.e., $\mathsf{P}^\pi(s) = \sum_a \pi(a|s)\mathsf{P}_s^a \in \Delta(\mathcal{S})$.

For any vector $V \in \mathbb{R}^S$, $V \circ V \in \mathbb{R}^S$ denotes the entry-wise multiplication, i.e., $V \circ V(s) = V(s) * V(s)$. For a distribution $q \in \Delta(\mathcal{S})$, it is straightforward to verify that the variance of V w.r.t. q can be rewritten as $\mathbf{Var}_q(V) = q(V \circ V) - (qV)^2$.

B A Straightforward Analysis: Hoeffding's Inequality

In this section, we present our proofs of the results under the Hoeffding-style construction.

We first note that Lemma 1 can be directly obtained from the existing results of concentration inequality, e.g., Theorem 1 from (Canonne, 2020) or Theorem 2.2 from (Weissman et al., 2003). Hence the proof is omitted here.

Theorem 3. *(Restatement of Theorem 1) With probability at least $1 - 2\delta$, it holds that*

$$V_{\mathsf{P}}^{\pi^*}(\rho) - V_{\mathsf{P}}^{\pi_r}(\rho) \leq \frac{2}{(1-\gamma)^2}\frac{8SC^{\pi^*}\log\frac{NS}{\delta}}{N} + \frac{2}{(1-\gamma)^2}\sqrt{\frac{24S^2C^{\pi^*}\log\frac{SA}{\delta}}{N}}, \tag{11}$$

To obtain an ϵ-optimal policy, a dataset of size

$$N = \mathcal{O}\left(\frac{SC^{\pi^*}}{(1-\gamma)^4\epsilon^2}\right) \tag{12}$$

is required.

Proof. In the following proof, we only focus on the case when

$$N > \frac{8SC^{\pi^*}\log\frac{NS}{\delta}}{1-\gamma}; \tag{13}$$

Otherwise, eq. (6) follows directly from the trivial bound $V_{\mathsf{P}}^{\pi^*}(\rho) - V_{\mathsf{P}}^{\pi_r}(\rho) \leq \frac{1}{1-\gamma}$.

According to Lemma 1, with probability at least $1-\delta$, $\mathsf{P} \in \hat{\mathcal{P}}$. Moreover, due to the fact $r(s,a) \geq \hat{r}(s,a)$, hence

$$V_{\mathsf{P}}^{\pi}(s) \geq V_{\hat{r},\mathsf{P}}^{\pi}(s) \geq V_{\hat{r},\hat{\mathcal{P}}}^{\pi}(s) = V_{\hat{\mathcal{P}}}^{\pi}(s) \tag{14}$$

for any π and $s \in \mathcal{S}$, where $V_{\hat{r},\mathsf{P}}^{\pi}$ denotes the value function w.r.t. P and reward $\hat{r}$. Thus $V_{\mathsf{P}}^{\pi} \geq V_{\hat{\mathcal{P}}}^{\pi}$ for any policy π.

Therefore,

$$\begin{aligned}
&V_{\mathsf{P}}^{\pi^*}(s) - V_{\mathsf{P}}^{\pi_r}(s)\\
&= V_{\mathsf{P}}^{\pi^*}(s) - V_{\hat{\mathcal{P}}}^{\pi_r}(s) + V_{\hat{\mathcal{P}}}^{\pi_r}(s) - V_{\mathsf{P}}^{\pi_r}(s)\\
&\leq V_{\mathsf{P}}^{\pi^*}(s) - V_{\hat{\mathcal{P}}}^{\pi_r}(s)\\
&= r(s,\pi^*(s)) + \gamma\mathsf{P}_s^{\pi^*(s)}V_{\mathsf{P}}^{\pi^*} - V_{\hat{\mathcal{P}}}^{\pi_r}(s)\\
&\overset{(a)}{\leq} r(s,\pi^*(s)) - \hat{r}(s,\pi^*(s)) + \gamma\mathsf{P}_s^{\pi^*(s)}V_{\mathsf{P}}^{\pi^*} - \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})\\
&= r(s,\pi^*(s)) - \hat{r}(s,\pi^*(s)) + \gamma\mathsf{P}_s^{\pi^*(s)}V_{\mathsf{P}}^{\pi^*} - \gamma\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} + \gamma\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} - \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})\\
&= r(s,\pi^*(s)) - \hat{r}(s,\pi^*(s)) + \gamma\mathsf{P}_s^{\pi^*(s)}(V_{\mathsf{P}}^{\pi^*} - V_{\hat{\mathcal{P}}}^{\pi_r}) + \gamma(\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} - \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r}))
\end{aligned}$$

$$\triangleq \gamma \mathsf{P}_s^{\pi^*(s)}(V_{\mathsf{P}}^{\pi^*} - V_{\hat{\mathcal{P}}}^{\pi_r}) + b^*(V_{\hat{\mathcal{P}}}^{\pi_r}), \tag{15}$$

where (a) is from $V_{\hat{\mathcal{P}}}^{\pi_r}(s) = \max_a Q_{\hat{\mathcal{P}}}^{\pi_r}(s,a) \geq Q_{\hat{\mathcal{P}}}^{\pi_r}(s,\pi^*(s)) = \hat{r}(s,\pi^*(s)) + \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})$, and $b^*(V)(s) \triangleq r(s,\pi^*(s)) - \hat{r}(s,\pi^*(s)) + \gamma \mathsf{P}_s^{\pi^*(s)}V - \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V)$.

Recursively applying this inequality further implies

$$V_{\mathsf{P}}^{\pi^*}(\rho) - V_{\mathsf{P}}^{\pi_r}(\rho) \leq \frac{1}{1-\gamma}\left\langle d^{\pi^*}, b^*(V_{\hat{\mathcal{P}}}^{\pi_r})\right\rangle, \tag{16}$$

where $d^{\pi^*}(\cdot) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(S_t = \cdot | S_0 \sim \rho, \pi^*, \mathsf{P})$ is the discounted visitation distribution induced by π^* and P.

To bound the term in eq. (15), we introduce the following notations.

$$\mathcal{S}_s \triangleq \left\{ s : N\mu(s,\pi^*(s)) \leq 8\log\frac{NS}{\delta}\right\}, \tag{17}$$

$$\mathcal{S}_l \triangleq \left\{ s : N\mu(s,\pi^*(s)) > 8\log\frac{NS}{\delta}\right\}. \tag{18}$$

The two sets divide the state space into two sub-spaces. $\mathcal{S}_s$ contains the states that have less probability in the behavior distribution μ on its corresponding optimal action, and are less-frequently covered by the dataset; And $\mathcal{S}_l$ are the ones that have larger probability, and are more frequently covered.

For $s \in \mathcal{S}_s$, from eq. (13), we have that

$$\min\left\{d^{\pi^*}(s), \frac{1}{S}\right\} \leq C^{\pi^*}\mu(s,\pi^*(s)) \leq \frac{8C^{\pi^*}\log\frac{NS}{\delta}}{N} < \frac{1}{S}, \tag{19}$$

which further implies that $d^{\pi^*}(s) \leq \frac{8C^{\pi^*}\log\frac{NS}{\delta}}{N}$. Hence

$$\sum_{s\in\mathcal{S}_s} d^{\pi^*}(s) b^*(V_{\hat{\mathcal{P}}}^{\pi_r})(s) \leq \frac{1}{1-\gamma}\frac{8SC^{\pi^*}\log\frac{NS}{\delta}}{N}, \tag{20}$$

which is due to $b^*(V)(s) \triangleq r(s,\pi^*(s)) - \hat{r}(s,\pi^*(s)) + \gamma\mathsf{P}_s^{\pi^*(s)}V - \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V) \leq 1 + \frac{\gamma}{1-\gamma} = \frac{1}{1-\gamma}$.

We then consider $s \in \mathcal{S}_l$. From the definition, $N\mu(s,\pi^*(s)) > 8\log\frac{NS}{\delta}$. According to Lemma 4, with probability $1-\delta$,

$$\max\{12N(s,\pi^*(s)), 8\log\frac{NS}{\delta}\} \geq N\mu(s,\pi^*(s)) > 8\log\frac{NS}{\delta}, \tag{21}$$

hence $\max\{12N(s,\pi^*(s)), 8\log\frac{NS}{\delta}\} = 12N(s,\pi^*(s))$ and $N(s,\pi^*(s)) \geq \frac{2}{3}\log\frac{NS}{\delta} > 0$. This hence implies that for any $s \in \mathcal{S}_l$, $N(s,\pi^*(s)) > 0$ and $\hat{r}(s,\pi^*(s)) = r(s,\pi^*(s))$. Thus

$$\begin{aligned}
|b^*(V_{\hat{\mathcal{P}}}^{\pi_r})(s)| &= \gamma|\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} - \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})| \\
&\leq \|\mathsf{P}_s^{\pi^*(s)} - \mathsf{Q}_s^{\pi^*(s)}\|_1 \|V_{\hat{\mathcal{P}}}^{\pi_r}\|_\infty \\
&\leq \frac{2}{1-\gamma}\min\left\{2, \sqrt{\frac{S\log\frac{SA}{\delta}}{2N(s,\pi^*(s))}}\right\} \\
&\leq \frac{1}{1-\gamma}\sqrt{\frac{2S\log\frac{SA}{\delta}}{N(s,\pi^*(s))}}.
\end{aligned} \tag{22}$$

Moreover, from eq. (21),

$$\frac{1}{N(s,\pi^*(s))} \leq \frac{12}{N\mu(s,\pi^*(s))} \leq \frac{12C^{\pi^*}}{N\min\{d^{\pi^*}(s), \frac{1}{S}\}} \leq \frac{12C^{\pi^*}}{N}\left(\frac{1}{d^{\pi^*}(s)} + S\right). \tag{23}$$

Combining with eq. (22) further implies

$$
\begin{aligned}
|b^*(V_{\hat{\mathcal{P}}}^{\pi_r})(s)| &\leq \frac{1}{1-\gamma}\sqrt{2S\log\frac{SA}{\delta}}\sqrt{\frac{12C^{\pi^*}}{N}\left(\frac{1}{d^{\pi^*}(s)}+S\right)} \\
&\leq \frac{1}{1-\gamma}\sqrt{2S\log\frac{SA}{\delta}}\sqrt{\frac{12C^{\pi^*}}{N}\frac{1}{d^{\pi^*}(s)}}+\frac{1}{1-\gamma}\sqrt{2S\log\frac{SA}{\delta}}\sqrt{\frac{12C^{\pi^*}}{N}S}. \qquad (24)
\end{aligned}
$$

Thus

$$
\begin{aligned}
\sum_{s\in\mathcal{S}_l} d^{\pi^*}(s)b^*(V_{\hat{\mathcal{P}}}^{\pi_r})(s) &\leq \frac{1}{1-\gamma}\sqrt{\frac{24SC^{\pi^*}\log\frac{SA}{\delta}}{N}}\sum_s\sqrt{d^{\pi^*}(s)}+\frac{1}{1-\gamma}\sqrt{\frac{24S^2C^{\pi^*}\log\frac{SA}{\delta}}{N}} \\
&\leq \frac{1}{1-\gamma}\sqrt{\frac{24S^2C^{\pi^*}\log\frac{2}{\delta}}{N}}+\frac{1}{1-\gamma}\sqrt{\frac{24S^2C^{\pi^*}\log\frac{SA}{\delta}}{N}} \\
&= \frac{2}{1-\gamma}\sqrt{\frac{24S^2C^{\pi^*}\log\frac{SA}{\delta}}{N}}. \qquad (25)
\end{aligned}
$$

Thus combining eq. (20) and eq. (25) implies

$$
\begin{aligned}
&V_{\mathsf{P}}^{\pi^*}(\rho)-V_{\mathsf{P}}^{\pi_r}(\rho) \\
&\leq \frac{1}{1-\gamma}\left\langle d^{\pi^*}, b^*(V_{\hat{\mathcal{P}}}^{\pi_r})\right\rangle \\
&\leq \frac{2}{(1-\gamma)^2}\frac{8SC^{\pi^*}\log\frac{NS}{\delta}}{N}+\frac{1}{(1-\gamma)^2}\sqrt{\frac{24S^2C^{\pi^*}\log\frac{SA}{\delta}}{N}}, \qquad (26)
\end{aligned}
$$

which completes the proof. □

C A Refined Analysis: Bernstein's Inequality

In this section, we provide a refined analysis of the sub-optimality gap using Bernstein's Inequality.

Theorem 4. *(Restatement of Thm 2) Consider the robust MDP $\hat{\mathcal{M}}$, there exist an optimal policy π_r such that there exists some universal constant K, such that with probability at least $1-4\delta$, if $N \geq \frac{1}{(1-\gamma)KSC^{\pi^*}\mu_{\min}^2}$, it holds that*

$$
V_{\mathsf{P}}^{\pi^*}(\rho)-V_{\mathsf{P}}^{\pi_r}(\rho) \leq \sqrt{\frac{KSC^{\pi^*}\log\frac{4N}{\delta}}{(1-\gamma)^3N}}+\frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)^2N\mu_{\min}}, \qquad (27)
$$

where $K = 442368$.

Proof. We first have that

$$
V_{\mathsf{P}}^{\pi^*}-V_{\mathsf{P}}^{\pi_r} = \underbrace{V_{\mathsf{P}}^{\pi^*}-V_{\hat{\mathcal{P}}}^{\pi_r}}_{\Delta_1}+\underbrace{V_{\hat{\mathcal{P}}}^{\pi_r}-V_{\mathsf{P}}^{\pi_r}}_{\Delta_2}. \qquad (28)
$$

In the following proof, we only focus on the case when

$$
N > \max\left\{\frac{SKC^{\pi^*}\log\frac{NS}{\delta}}{1-\gamma}, \frac{1}{(1-\gamma)KSC^{\pi^*}\mu_{\min}^2}\right\}; \qquad (29)
$$

Otherwise, eq. (8) follows directly from the trivial bound $V_{\mathsf{P}}^{\pi^*}(\rho)-V_{\mathsf{P}}^{\pi_r}(\rho) \leq \frac{1}{1-\gamma}$.

We note that Lemma 8 of (Shi & Chi, 2022) states that under eq. (29), with probability $1-\delta$, for any (s,a) pair,

$$N(s,a) \geq \frac{N\mu(s,a)}{8\log\frac{4SA}{\delta}}. \tag{30}$$

This moreover implies that with probability $1-\delta$, if $\mu(s,a)>0$, then $N(s,a)>0$. We hence focus on the case when this event holds.

The remaining proof can be completed by combining the following two theorems. □

Theorem 5. *With probability at least $1-2\delta$, it holds that*

$$\rho^\top \Delta_1 \leq \frac{2c_2 SC^{\pi^*}\log\frac{4N}{\delta}}{(1-\gamma)^2 N} + \frac{80Sc_1C^{\pi^*}\log\frac{NS}{\delta}}{(1-\gamma)^2 N} + \frac{96}{\gamma}\sqrt{\frac{48SC^{\pi^*}\log\frac{4N}{\delta}}{(1-\gamma)^3 N}}. \tag{31}$$

Proof. We first define the following set:

$$\mathcal{S}_0 \triangleq \{s : d^{\pi^*}(s) = 0\}. \tag{32}$$

And for $s \notin \mathcal{S}_0$, it holds that $d^{\pi^*}(s) > 0$.

We first consider $s \notin \mathcal{S}_0$. Due to the fact $d^{\pi^*}(s)>0$, hence $d^{\pi^*}(s,\pi^*(s)) = d^{\pi^*}(s)\pi^*(\pi^*(s)|s) > 0$. Thus it implies that $\mu(s,\pi^*(s))>0$, and equation 30 further implies $N(s,\pi^*(s))>0$ and $\hat{r}(s,\pi^*(s)) = r(s,\pi^*(s))$. Hence we have that

$$\begin{aligned}
V_{\mathsf{P}}^{\pi^*}(s) - V_{\hat{\mathcal{P}}}^{\pi_r}(s) &= r(s,\pi^*(s)) + \gamma \mathsf{P}_s^{\pi^*(s)} V_{\mathsf{P}}^{\pi^*} - V_{\hat{\mathcal{P}}}^{\pi_r}(s) \\
&\overset{(a)}{\leq} \gamma \mathsf{P}_s^{\pi^*(s)} V_{\mathsf{P}}^{\pi^*} - \gamma \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) \\
&= \gamma \mathsf{P}_s^{\pi^*(s)} V_{\mathsf{P}}^{\pi^*} - \gamma \mathsf{P}_s^{\pi^*(s)} V_{\hat{\mathcal{P}}}^{\pi_r} + \gamma \mathsf{P}_s^{\pi^*(s)} V_{\hat{\mathcal{P}}}^{\pi_r} - \gamma \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) \\
&= \gamma \mathsf{P}_s^{\pi^*(s)}(V_{\mathsf{P}}^{\pi^*} - V_{\hat{\mathcal{P}}}^{\pi_r}) + \gamma(\mathsf{P}_s^{\pi^*(s)} V_{\hat{\mathcal{P}}}^{\pi_r} - \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})),
\end{aligned} \tag{33}$$

where (a) is from $V_{\hat{\mathcal{P}}}^{\pi_r}(s) = \max_a Q_{\hat{\mathcal{P}}}^{\pi_r}(s,a) \geq Q_{\hat{\mathcal{P}}}^{\pi_r}(s,\pi^*(s)) = \hat{r}(s,\pi^*(s)) + \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) = r(s,\pi^*(s)) + \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})$.

For $s \in \mathcal{S}_0$, it holds that

$$\begin{aligned}
V_{\mathsf{P}}^{\pi^*}(s) - V_{\hat{\mathcal{P}}}^{\pi_r}(s) &= r(s,\pi^*(s)) + \gamma \mathsf{P}_s^{\pi^*(s)} V_{\mathsf{P}}^{\pi^*} - V_{\hat{\mathcal{P}}}^{\pi_r}(s) \\
&\overset{(a)}{\leq} \gamma \mathsf{P}_s^{\pi^*(s)} V_{\mathsf{P}}^{\pi^*} - \gamma \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) + r(s,\pi^*(s)) - \hat{r}(s,\pi^*(s)) \\
&\leq \gamma \mathsf{P}_s^{\pi^*(s)}(V_{\mathsf{P}}^{\pi^*} - V_{\hat{\mathcal{P}}}^{\pi_r}) + \gamma(\mathsf{P}_s^{\pi^*(s)} V_{\hat{\mathcal{P}}}^{\pi_r} - \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})) + 1,
\end{aligned} \tag{34}$$

where (a) follows similarly from eq. (33), and the last inequality is from $r(s,\pi^*(s)) \leq 1$.

Hence combining eq. (33) and eq. (34) implies

$$V_{\mathsf{P}}^{\pi^*}(s) - V_{\hat{\mathcal{P}}}^{\pi_r}(s) \leq \gamma \mathsf{P}_s^{\pi^*(s)}(V_{\mathsf{P}}^{\pi^*} - V_{\hat{\mathcal{P}}}^{\pi_r}) + b^*(V)(s), \tag{35}$$

where $b^*(V)(s) \triangleq \gamma \mathsf{P}_s^{\pi^*(s)} V - \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V)$ if $s \notin \mathcal{S}_0$, and $b^*(V)(s) \triangleq 1 + \gamma \mathsf{P}_s^{\pi^*(s)} V - \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V)$ for $s \in \mathcal{S}_0$.

Moreover we set

$$\tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})(s) = \max\{0, b^*(V_{\hat{\mathcal{P}}}^{\pi_r})(s)\}, \tag{36}$$

then it holds that $b^*(V_{\hat{\mathcal{P}}}^{\pi_r}) \leq \tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})$.

Then apply eq. (35) recursively and we have that

$$\rho^\top \Delta_1 \leq \frac{1}{1-\gamma}\left\langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})\right\rangle, \tag{37}$$

Not that for $s \in \mathcal{S}_0$, it holds that $d^{\pi^*}(s) = 0$, and

$$\left\langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})\right\rangle = \sum_{s \notin \mathcal{S}_0} d^{\pi^*}(s)\tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})(s). \tag{38}$$

This implies that we only need to focus on $s \notin \mathcal{S}_0$. We further defined the following sets:

$$\mathcal{S}_s \triangleq \left\{ s \notin \mathcal{S}_0 : N\mu(s, \pi^*(s)) \leq 8 \log \frac{NS}{\delta} \right\}, \tag{39}$$

$$\mathcal{S}_l \triangleq \left\{ s \notin \mathcal{S}_0 : N\mu(s, \pi^*(s)) > 8 \log \frac{NS}{\delta} \right\}. \tag{40}$$

For $s \in \mathcal{S}_s$, we have that

$$\min\left\{ d^{\pi^*}(s), \frac{1}{S} \right\} \leq C^{\pi^*}\mu(s, \pi^*(s)) \leq \frac{8C^{\pi^*} \log \frac{NS}{\delta}}{N} \overset{(a)}{<} \frac{1}{S}, \tag{41}$$

where (a) is due to the fact eq. (29).

This further implies that $d^{\pi^*}(s) \leq \frac{8C^{\pi^*} \log \frac{NS}{\delta}}{N}$. Hence

$$\sum_{s \in \mathcal{S}_s} d^{\pi^*}(s)\tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})(s) \leq \frac{2}{1-\gamma}\frac{8SC^{\pi^*} \log \frac{NS}{\delta}}{N}, \tag{42}$$

which is due to $\|\mathsf{P}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}} - \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V^{\pi_r}_{\hat{\mathcal{P}}})\| \leq \frac{2}{1-\gamma}$.

We then consider $s \in \mathcal{S}_l$. Note that from the definition and equation 30, it holds that $N(s, \pi^*(s)) > 0$ for $s \in \mathcal{S}_l$.

Therefore it holds that

$$\begin{aligned}
\tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})(s) &\leq |\gamma \mathsf{P}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}} - \gamma \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V^{\pi_r}_{\hat{\mathcal{P}}})| \\
&\leq |\gamma \mathsf{P}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}} - \gamma \hat{\mathsf{P}}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}}| + |\gamma \hat{\mathsf{P}}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}} - \gamma \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V^{\pi_r}_{\hat{\mathcal{P}}})| \\
&\leq \gamma \left| \mathsf{P}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}} - \hat{\mathsf{P}}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}} \right| + \sqrt{\frac{\log \frac{4N}{\delta} \mathbf{Var}_{\hat{\mathsf{P}}_s^{\pi^*(s)}}(V^{\pi_r}_{\hat{\mathcal{P}}})}{N(s, \pi^*(s))}},
\end{aligned} \tag{43}$$

where the last inequality is shown as follows.

Since $\hat{\mathsf{P}}_s^{\pi^*(s)} \in \hat{\mathcal{P}}_s^a$, we have that $\hat{\mathsf{P}}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}} \geq \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V^{\pi_r}_{\hat{\mathcal{P}}})$. Now note that it is shown in (Iyengar, 2005; Shi et al., 2023) that

$$\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V^{\pi_r}_{\hat{\mathcal{P}}}) = \max_{\mu \in [0, V^{\pi_r}_{\hat{\mathcal{P}}}]} \left\{ \hat{\mathsf{P}}_s^{\pi^*(s)}(V^{\pi_r}_{\hat{\mathcal{P}}} - \mu) - \sqrt{R_s^{\pi^*(s)} \mathbf{Var}_{\hat{\mathsf{P}}_s^{\pi^*(s)}}(V^{\pi_r}_{\hat{\mathcal{P}}} - \mu)} \right\}. \tag{44}$$

Thus

$$\begin{aligned}
&|\gamma \hat{\mathsf{P}}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}} - \gamma \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V^{\pi_r}_{\hat{\mathcal{P}}})| \\
&= \gamma \hat{\mathsf{P}}_s^{\pi^*(s)} V^{\pi_r}_{\hat{\mathcal{P}}} - \gamma \sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V^{\pi_r}_{\hat{\mathcal{P}}})
\end{aligned}$$

$$
\begin{aligned}
&= \gamma \hat{\mathsf{P}}_s^{\pi^*(s)} V_{\hat{\mathcal{P}}}^{\pi_r} - \gamma \max_{\mu \in [0, V_{\hat{\mathcal{P}}}^{\pi_r}]} \left\{ \hat{\mathsf{P}}_s^{\pi^*(s)} (V_{\hat{\mathcal{P}}}^{\pi_r} - \mu) - \sqrt{R_s^{\pi^*(s)} \mathbf{Var}_{\hat{\mathsf{P}}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r} - \mu)} \right\} \\
&\overset{(a)}{\leq} \gamma \hat{\mathsf{P}}_s^{\pi^*(s)} V_{\hat{\mathcal{P}}}^{\pi_r} - \gamma \left(\hat{\mathsf{P}}_s^{\pi^*(s)} (V_{\hat{\mathcal{P}}}^{\pi_r}) - \sqrt{R_s^{\pi^*(s)} \mathbf{Var}_{\hat{\mathsf{P}}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} \right) \\
&= \sqrt{R_s^{\pi^*(s)} \mathbf{Var}_{\hat{\mathsf{P}}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})}, \qquad (45)
\end{aligned}
$$

where (a) is due to the maximum term is larger than the function value at $\mu = 0$, and this inequality completes the proof of Equation (43).

To further bound eq. (43), we invoke Lemma 6 and have that

$$
\begin{aligned}
&\left| \mathsf{P}_s^{\pi^*(s)} V_{\hat{\mathcal{P}}}^{\pi_r} - \hat{\mathsf{P}}_s^{\pi^*(s)} V_{\hat{\mathcal{P}}}^{\pi_r} \right| \\
&\leq 12 \sqrt{\frac{\mathbf{Var}_{\hat{\mathsf{P}}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r}) \log \frac{4N}{\delta}}{N(s, \pi^*(s))}} + \frac{74 \log \frac{4N}{\delta}}{(1-\gamma) N(s, \pi^*(s))} \\
&\leq 12 \sqrt{\frac{\log \frac{4N}{\delta}}{N(s, \pi^*(s))} \left(2 \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r}) + \frac{41 \log \frac{4N}{\delta}}{(1-\gamma)^2 N(s, \pi^*(s))} \right)} + \frac{74 \log \frac{4N}{\delta}}{(1-\gamma) N(s, \pi^*(s))} \\
&\leq 12 \sqrt{\frac{2 \log \frac{4N}{\delta}}{N(s, \pi^*(s))} \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} + \frac{(74 + 12\sqrt{41}) \log \frac{4N}{\delta}}{(1-\gamma) N(s, \pi^*(s))}, \qquad (46)
\end{aligned}
$$

where the last inequality is from $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$.

Combine eq. (43) and eq. (46), we further have that

$$
\begin{aligned}
\tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})(s) &\leq 24 \sqrt{\frac{2 \log \frac{4N}{\delta}}{N(s, \pi^*(s))} \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} + \frac{(74 + 12\sqrt{41}) \log \frac{4N}{\delta}}{(1-\gamma) N(s, \pi^*(s))} + \frac{\log \frac{N}{\delta}}{(1-\gamma) N(s, \pi^*(s))} \\
&\leq 24 \sqrt{\frac{2 \log \frac{4N}{\delta}}{N(s, \pi^*(s))} \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} + \frac{c_1 \log \frac{4N}{\delta}}{(1-\gamma) N(s, \pi^*(s))}, \qquad (47)
\end{aligned}
$$

where $c_1 = 75 + 12\sqrt{41}$.

Note that in eq. (23), we showed that $\frac{1}{N(s, \pi^*(s))} \leq \frac{12 C^{\pi^*}}{N} (S + \frac{1}{d^{\pi^*}(s)})$. Hence plugging in eq. (47) implies that

$$
\begin{aligned}
\tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})(s) &\leq 24 \sqrt{\frac{24 C^{\pi^*} \log \frac{4N}{\delta}}{N} \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} \left(\sqrt{S} + \frac{1}{\sqrt{d^{\pi^*}(s)}} \right) \\
&\quad + \frac{12 c_1 C^{\pi^*} \log \frac{4N}{\delta}}{(1-\gamma) N} \left(S + \frac{1}{d^{\pi^*}(s)} \right). \qquad (48)
\end{aligned}
$$

Firstly we have that

$$
\begin{aligned}
&\sum_{s \in \mathcal{S}_l} 24 d^{\pi^*}(s) \sqrt{\frac{24 C^{\pi^*} \log \frac{4N}{\delta}}{N} \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} \left(\sqrt{S} + \frac{1}{\sqrt{d^{\pi^*}(s)}} \right) \\
&= \sum_{s \in \mathcal{S}_l} 24 d^{\pi^*}(s) \sqrt{\frac{24 S C^{\pi^*} \log \frac{4N}{\delta}}{N} \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} + \sum_{s \in \mathcal{S}_l} 12 \sqrt{d^{\pi^*}(s)} \sqrt{\frac{24 C^{\pi^*} \log \frac{4N}{\delta}}{N} \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} \\
&= 24 \sqrt{\frac{24 C^{\pi^*} \log \frac{4N}{\delta}}{N}} \left(\sum_{s \in \mathcal{S}_l} \sqrt{d^{\pi^*}(s) \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} + \sum_{s \in \mathcal{S}_l} \sqrt{d^{\pi^*}(s)} \sqrt{S d^{\pi^*}(s) \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} \right) \\
&\overset{(a)}{\leq} 24 \sqrt{\frac{24 C^{\pi^*} \log \frac{4N}{\delta}}{N}} \left(\sqrt{S} \sqrt{\sum_{s \in \mathcal{S}_l} d^{\pi^*}(s) \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} + \sqrt{\sum_{s \in \mathcal{S}_l} S d^{\pi^*}(s) \mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}} (V_{\hat{\mathcal{P}}}^{\pi_r})} \right)
\end{aligned}
$$

$$= 48\sqrt{\frac{24SC^{\pi^*}\log\frac{4N}{\delta}}{N}}\sqrt{\sum_{s\in\mathcal{S}_l} d^{\pi^*}(s)\mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})}, \tag{49}$$

where (a) is from Cauchy's inequality and the fact $\sum_{s\in\mathcal{S}_l} d^{\pi^*}(s) \leq 1$.

In addition, we have that

$$\sum_{s\in\mathcal{S}_l} d^{\pi^*}(s)\frac{12c_1C^{\pi^*}\log\frac{4N}{\delta}}{(1-\gamma)N}\left(S+\frac{1}{d^{\pi^*}(s)}\right) \leq \frac{24c_1SC^{\pi^*}\log\frac{4N}{\delta}}{(1-\gamma)N}. \tag{50}$$

Combine the two inequalities above and we have that

$$\begin{aligned}
&\sum_{s\in\mathcal{S}_l} d^{\pi^*}(s)\tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})(s) \\
&\leq 48\sqrt{\frac{24SC^{\pi^*}\log\frac{4N}{\delta}}{N}}\sqrt{\sum_{s\in\mathcal{S}_l} d^{\pi^*}(s)\mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})} + \frac{24Sc_1C^{\pi^*}\log\frac{4N}{\delta}}{(1-\gamma)N}.
\end{aligned} \tag{51}$$

Then we combine eq. (42) and eq. (51), and it implies that

$$\begin{aligned}
&\langle d^{\pi^*}, \tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})\rangle \\
&= \sum_{s\in\mathcal{S}_s} d^{\pi^*}(s)\tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})(s) + \sum_{s\in\mathcal{S}_l} d^{\pi^*}(s)\tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})(s) \\
&\leq \frac{16SC^{\pi^*}\log\frac{NS}{\delta}}{(1-\gamma)N} + 48\sqrt{\frac{24SC^{\pi^*}\log\frac{4N}{\delta}}{N}}\sqrt{\sum_{s\in\mathcal{S}_l} d^{\pi^*}(s)\mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})} + \frac{24Sc_1C^{\pi^*}\log\frac{4N}{\delta}}{(1-\gamma)N} \\
&\leq \frac{40c_1SC^{\pi^*}\log\frac{NS}{\delta}}{(1-\gamma)N} + 48\sqrt{\frac{24SC^{\pi^*}\log\frac{4N}{\delta}}{N}}\sqrt{\sum_{s\in\mathcal{S}} d^{\pi^*}(s)\mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})}.
\end{aligned} \tag{52}$$

We then bound the term $\sum_{s\in\mathcal{S}} d^{\pi^*}(s)\mathbf{Var}_{\mathsf{P}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})$. We first claim the following inequality:

$$V_{\hat{\mathcal{P}}}^{\pi_r} - \gamma\mathsf{P}^{\pi^*}V_{\hat{\mathcal{P}}}^{\pi_r} + 2\tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r}) \geq 0. \tag{53}$$

To prove eq. (53), we note that

$$\begin{aligned}
V_{\hat{\mathcal{P}}}^{\pi_r}(s) &= \max_a Q_{\hat{\mathcal{P}}}^{\pi_r}(s,a) \\
&\geq Q_{\hat{\mathcal{P}}}^{\pi_r}(s,\pi^*(s)) \\
&= \hat{r}(s,\pi^*(s)) + \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) \\
&= \hat{r}(s,\pi^*(s)) + \gamma\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} - \gamma\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} + \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) \\
&\overset{(a)}{\geq} \hat{r}(s,\pi^*(s)) + \gamma\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} - b^*(V_{\hat{\mathcal{P}}}^{\pi_r})(s) \\
&\geq \hat{r}(s,\pi^*(s)) + \gamma\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} - 2\tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})(s),
\end{aligned} \tag{54}$$

where (a) is from $b^*(V_{\hat{\mathcal{P}}}^{\pi_r})(s) \geq \gamma\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} - \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi^*(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r})$.

Hence for any $s \in \mathcal{S}$,

$$V_{\hat{\mathcal{P}}}^{\pi_r}(s) - \gamma\mathsf{P}_s^{\pi^*(s)}V_{\hat{\mathcal{P}}}^{\pi_r} + 2\tilde{b}(V_{\hat{\mathcal{P}}}^{\pi_r})(s) \geq \hat{r}(s,\pi^*(s)) \geq 0, \tag{55}$$

which proves eq. (53).

Now with eq. (53), we first note that

$$
\begin{aligned}
&(V^{\pi_r}_{\hat{\mathcal{P}}} \circ V^{\pi_r}_{\hat{\mathcal{P}}}) - (\gamma \mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}}) \circ (\gamma \mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}}) \\
&= (V^{\pi_r}_{\hat{\mathcal{P}}} - \gamma \mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}}) \circ (V^{\pi_r}_{\hat{\mathcal{P}}} + \gamma \mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}}) \\
&\leq (V^{\pi_r}_{\hat{\mathcal{P}}} - \gamma \mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}} + 2\tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})) \circ (V^{\pi_r}_{\hat{\mathcal{P}}} + \gamma \mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}}) \\
&\leq \frac{2}{1-\gamma}(V^{\pi_r}_{\hat{\mathcal{P}}} - \gamma \mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}} + 2\tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})),
\end{aligned} \tag{56}
$$

where the last inequality is due to the fact $\|V^{\pi_r}_{\hat{\mathcal{P}}} + \gamma \mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}}\| \leq \frac{2}{1-\gamma}$ and eq. (53).

We then have that

$$
\begin{aligned}
&\sum_{s\in\mathcal{S}} d^{\pi^*}(s) \mathbf{Var}_{\mathsf{P}^{\pi^*(s)}_s}(V^{\pi_r}_{\hat{\mathcal{P}}}) \\
&= \langle d^{\pi^*}, \mathsf{P}^{\pi^*}(V^{\pi_r}_{\hat{\mathcal{P}}} \circ V^{\pi_r}_{\hat{\mathcal{P}}}) - (\mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}}) \circ (\mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle \\
&\overset{(a)}{\leq} \left\langle d^{\pi^*}, \mathsf{P}^{\pi^*}(V^{\pi_r}_{\hat{\mathcal{P}}} \circ V^{\pi_r}_{\hat{\mathcal{P}}}) - \frac{1}{\gamma^2}(V^{\pi_r}_{\hat{\mathcal{P}}} \circ V^{\pi_r}_{\hat{\mathcal{P}}}) + \frac{2}{\gamma^2(1-\gamma)}(V^{\pi_r}_{\hat{\mathcal{P}}} - \gamma \mathsf{P}^{\pi^*} V^{\pi_r}_{\hat{\mathcal{P}}} + 2\tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})) \right\rangle \\
&\overset{(b)}{\leq} \left\langle d^{\pi^*}, \mathsf{P}^{\pi^*}(V^{\pi_r}_{\hat{\mathcal{P}}} \circ V^{\pi_r}_{\hat{\mathcal{P}}}) - \frac{1}{\gamma}(V^{\pi_r}_{\hat{\mathcal{P}}} \circ V^{\pi_r}_{\hat{\mathcal{P}}}) + \frac{2}{\gamma^2(1-\gamma)}(I - \gamma \mathsf{P}^{\pi^*}) V^{\pi_r}_{\hat{\mathcal{P}}} + \frac{4}{\gamma^2(1-\gamma)}\tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})) \right\rangle \\
&= \left\langle d^{\pi^*}, \frac{1}{\gamma}(\gamma \mathsf{P}^{\pi^*} - I)(V^{\pi_r}_{\hat{\mathcal{P}}} \circ V^{\pi_r}_{\hat{\mathcal{P}}}) + \frac{2}{\gamma^2(1-\gamma)}(I - \gamma \mathsf{P}^{\pi^*}) V^{\pi_r}_{\hat{\mathcal{P}}} + \frac{4}{\gamma^2(1-\gamma)}\tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}})) \right\rangle \\
&= (d^{\pi^*})^\top (I - \gamma \mathsf{P}^{\pi^*}) \left(-\frac{1}{\gamma}(V^{\pi_r}_{\hat{\mathcal{P}}} \circ V^{\pi_r}_{\hat{\mathcal{P}}}) + \frac{2}{\gamma^2(1-\gamma)} V^{\pi_r}_{\hat{\mathcal{P}}} \right) + \frac{4}{\gamma^2(1-\gamma)} \langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle \\
&\overset{(c)}{=} (1-\gamma)\rho^\top \left(-\frac{1}{\gamma}(V^{\pi_r}_{\hat{\mathcal{P}}} \circ V^{\pi_r}_{\hat{\mathcal{P}}}) + \frac{2}{\gamma^2(1-\gamma)} V^{\pi_r}_{\hat{\mathcal{P}}} \right) + \frac{4}{\gamma^2(1-\gamma)} \langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle \\
&\leq \frac{2}{\gamma^2} \rho^\top V^{\pi_r}_{\hat{\mathcal{P}}} + \frac{4}{\gamma^2(1-\gamma)} \langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle \\
&\leq \frac{2}{\gamma^2(1-\gamma)} + \frac{4}{\gamma^2(1-\gamma)} \langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle,
\end{aligned} \tag{57}
$$

where (a) is from eq. (56), (b) is due to $\gamma < 1$, (c) is from the definition of visitation distribution.

Hence by plugging eq. (57) in eq. (52), we have that

$$
\begin{aligned}
&\langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle \\
&\leq \frac{40 c_1 S C^{\pi^*} \log \frac{NS}{\delta}}{(1-\gamma)N} + 48 \sqrt{\frac{24 S C^{\pi^*} \log \frac{4N}{\delta}}{N}} \sqrt{\sum_{s\in\mathcal{S}} d^{\pi^*}(s) \mathbf{Var}_{\mathsf{P}^{\pi^*(s)}_s}(V^{\pi_r}_{\hat{\mathcal{P}}})} \\
&\leq \frac{40 c_1 S C^{\pi^*} \log \frac{NS}{\delta}}{(1-\gamma)N} + 48 \sqrt{\frac{24 S C^{\pi^*} \log \frac{4N}{\delta}}{N}} \sqrt{\frac{2}{\gamma^2(1-\gamma)} + \frac{4}{\gamma^2(1-\gamma)} \langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle} \\
&\leq \frac{40 c_1 S C^{\pi^*} \log \frac{NS}{\delta}}{(1-\gamma)N} + \frac{24}{\gamma} \sqrt{\frac{48 S C^{\pi^*} \log \frac{4N}{\delta}}{(1-\gamma)N}} + \frac{48}{\gamma} \sqrt{\frac{96 S C^{\pi^*} \log \frac{4N}{\delta}}{(1-\gamma)N}} \sqrt{\langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle} \\
&\overset{(a)}{\leq} \frac{1}{2} \langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle + c_2 \frac{S C^{\pi^*} \log \frac{4N}{\delta}}{(1-\gamma)N} + \frac{40 S c_1 C^{\pi^*} \log \frac{NS}{\delta}}{(1-\gamma)N} + \frac{48}{\gamma} \sqrt{\frac{48 S C^{\pi^*} \log \frac{4N}{\delta}}{(1-\gamma)N}},
\end{aligned} \tag{58}
$$

where (a) is from $x + y \geq 2\sqrt{xy}$ and $c_2 = 8 * 24^3 = 110592$. This inequality moreover implies that

$$
\langle d^{\pi^*}, \tilde{b}(V^{\pi_r}_{\hat{\mathcal{P}}}) \rangle \leq 2 c_2 \frac{S C^{\pi^*} \log \frac{4N}{\delta}}{(1-\gamma)N} + \frac{80 S c_1 C^{\pi^*} \log \frac{NS}{\delta}}{(1-\gamma)N} + \frac{96}{\gamma} \sqrt{\frac{48 S C^{\pi^*} \log \frac{4N}{\delta}}{(1-\gamma)N}}. \tag{59}
$$

Recall the definition of Δ_1, we hence have that

$$\rho^\top \Delta_1 \leq \frac{2c_2 SC^{\pi^*} \log \frac{4N}{\delta}}{(1-\gamma)^2 N} + \frac{80Sc_1 C^{\pi^*} \log \frac{NS}{\delta}}{(1-\gamma)^2 N} + \frac{96}{\gamma}\sqrt{\frac{48SC^{\pi^*} \log \frac{4N}{\delta}}{(1-\gamma)^3 N}}. \tag{60}$$

This hence completes the proof of the lemma. □

Theorem 6. *With probability at least $1-2\delta$, it holds that*

$$\rho^\top \Delta_2 \leq 2\sqrt{\frac{384 \log^2 \frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)^3 N^2 \mu_{\min}}} + 2\sqrt{\frac{384 \log^2 \frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)^3 N^3 \mu_{\min}}} + \frac{384 \log^2 \frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)^2 N \mu_{\min}}. \tag{61}$$

Proof. We first define the following set:

$$\mathcal{S}^0 \triangleq \{s \in \mathcal{S} : N(s) = 0\}. \tag{62}$$

Note that $N(s) = \sum_a N(s,a)$, hence it holds that $N(s,a) = 0$ for any $s \in \mathcal{S}^0$, $a \in \mathcal{A}$.

We moreover construct an absorbing MDP $\bar{M} = (\mathcal{S}, \mathcal{A}, \hat{r}, \bar{\mathsf{P}})$ as follows. For $s \in \mathcal{S}^0$, $\bar{\mathsf{P}}^a_{s,x} = \mathbf{1}_{x=s}$; And for $s \notin \mathcal{S}^0$, set $\bar{\mathsf{P}}^a_{s,x} = \mathsf{P}^a_{s,x}$.

Then for any $s \in \mathcal{S}^0$, from Lemma 2, it holds that $V^{\pi_r}_{\hat{\mathcal{P}}}(s) = 0$, and hence

$$V^{\pi_r}_{\hat{\mathcal{P}}}(s) - V^{\pi_r}_{\mathsf{P}}(s) \leq 0. \tag{63}$$

It further implies that

$$V^{\pi_r}_{\hat{\mathcal{P}}}(s) - V^{\pi_r}_{\mathsf{P}}(s) = \bar{\mathsf{P}}^{\pi_r(s)}_s (V^{\pi_r}_{\hat{\mathcal{P}}} - V^{\pi_r}_{\mathsf{P}}) \leq \gamma \bar{\mathsf{P}}^{\pi_r(s)}_s (V^{\pi_r}_{\hat{\mathcal{P}}} - V^{\pi_r}_{\mathsf{P}}). \tag{64}$$

On the other hand, for $s \notin \mathcal{S}^0$, Lemma 2 implies that $N(s, \pi_r(s)) > 0$, hence equation 30 implies $N(s,\pi_r(s)) > 0$, $\hat{r}(s,\pi_r(s)) = r(s,\pi_r(s))$. Hence

$$\begin{aligned}
V^{\pi_r}_{\hat{\mathcal{P}}}(s) - V^{\pi_r}_{\mathsf{P}}(s) &\overset{(a)}{=} \gamma \sigma_{\hat{\mathcal{P}}^{\pi_r(s)}_s}(V^{\pi_r}_{\hat{\mathcal{P}}}) - \gamma \mathsf{P}^{\pi_r(s)}_s V^{\pi_r}_{\mathsf{P}} \\
&= \gamma(\sigma_{\hat{\mathcal{P}}^{\pi_r(s)}_s}(V^{\pi_r}_{\hat{\mathcal{P}}}) - \mathsf{P}^{\pi_r(s)}_s V^{\pi_r}_{\hat{\mathcal{P}}} + \mathsf{P}^{\pi_r(s)}_s V^{\pi_r}_{\hat{\mathcal{P}}} - \mathsf{P}^{\pi_r(s)}_s V^{\pi_r}_{\mathsf{P}}) \\
&= \gamma \mathsf{P}^{\pi_r(s)}_s (V^{\pi_r}_{\hat{\mathcal{P}}} - V^{\pi_r}_{\mathsf{P}}) + \gamma(\sigma_{\hat{\mathcal{P}}^{\pi_r(s)}_s}(V^{\pi_r}_{\hat{\mathcal{P}}}) - \mathsf{P}^{\pi_r(s)}_s V^{\pi_r}_{\hat{\mathcal{P}}}) \\
&\triangleq \gamma \bar{\mathsf{P}}^{\pi_r(s)}_s (V^{\pi_r}_{\hat{\mathcal{P}}} - V^{\pi_r}_{\mathsf{P}}) + c(s),
\end{aligned} \tag{65}$$

where (a) is from $r(s,\pi_r(s)) = \hat{r}(s,\pi_r(s))$, and $c(s) \triangleq \gamma(\sigma_{\hat{\mathcal{P}}^{\pi_r(s)}_s}(V^{\pi_r}_{\hat{\mathcal{P}}}) - \mathsf{P}^{\pi_r(s)}_s V^{\pi_r}_{\hat{\mathcal{P}}})$.

According to the bound we obtained in Lemma 7, it holds that

$$c(s) \leq 2\sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta}\epsilon_1}{(1-\gamma)N(s,\pi_r(s))}} + 2\epsilon_1\sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{N(s,\pi_r(s))}} + \frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,\pi_r(s))}. \tag{66}$$

Combine eq. (64) and eq. (65), then

$$V^{\pi_r}_{\hat{\mathcal{P}}}(s) - V^{\pi_r}_{\mathsf{P}}(s) \leq \gamma \bar{\mathsf{P}}^{\pi_r(s)}_s (V^{\pi_r}_{\hat{\mathcal{P}}} - V^{\pi_r}_{\mathsf{P}}) + \tilde{c}(s), \tag{67}$$

where

$$\tilde{c}(s) = \begin{cases} 2\sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta}\epsilon_1}{(1-\gamma)N(s,\pi_r(s))}} + 2\epsilon_1\sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{N(s,\pi_r(s))}} + \frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,\pi_r(s))}, & s \notin \mathcal{S}^0 \\ 0, & s \in \mathcal{S}^0 \end{cases} \tag{68}$$

Applying eq. (67) recursively further implies

$$\rho^\top \Delta_2 \leq \frac{1}{1-\gamma}\langle \bar{d}^{\pi_r}, \tilde{c}\rangle, \tag{69}$$

where $\bar{d}^{\pi_r}$ is the discounted visitation distribution induced by π_r and $\bar{\mathsf{P}}$.

Note that eq. (29) and Lemma 8 of (Shi & Chi, 2022) state that with probability $1-\delta$, for any (s,a) pair,

$$N(s,a) \geq \frac{N\mu(s,a)}{8\log\frac{4SA}{\delta}}. \tag{70}$$

Hence under this event, $\tilde{c}(s)$ can be bounded as

$$\tilde{c}(s) \leq 2\sqrt{\frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}\epsilon_1}{(1-\gamma)N\mu_{\min}}} + 2\epsilon_1\sqrt{\frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}}{N\mu_{\min}}} + \frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N\mu_{\min}}. \tag{71}$$

Hence we have that

$$\begin{aligned}
\rho^\top \Delta_2 &\leq \frac{1}{1-\gamma}\langle \bar{d}^{\pi_r}, \tilde{c}\rangle \\
&\leq 2\sqrt{\frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}\epsilon_1}{(1-\gamma)^3N\mu_{\min}}} + 2\epsilon_1\sqrt{\frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)^3N\mu_{\min}}} + \frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)^2N\mu_{\min}} \\
&\leq 2\sqrt{\frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)^3N^2\mu_{\min}}} + 2\sqrt{\frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)^3N^3\mu_{\min}}} + \frac{384\log^2\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)^2N\mu_{\min}}.
\end{aligned} \tag{72}$$

□

C.1 Auxiliary Lemmas

Lemma 2. *Recall the set $\mathcal{S}^0 \triangleq \{s \in \mathcal{S} : N(s) = 0\}$. Then*

(1). For any policy π and $s \in \mathcal{S}^0$, $V^\pi_{\hat{\mathcal{P}}}(s) = 0$;

(2). There exists a deterministic robust optimal policy π_r, such that for any $s \notin \mathcal{S}^0$, $N(s,\pi_r(s)) > 0$.

Proof. **Proof of (1).**

For any $s \in \mathcal{S}^0$, it holds that $N(s,a) = 0$ for any $a \in \mathcal{A}$. Hence $\hat{r}(s,a) = 0$ and $\hat{\mathcal{P}}^a_s = \Delta(\mathcal{S})$.

Then for any policy π and $a \in \mathcal{A}$, it holds that

$$Q^\pi_{\hat{\mathcal{P}}}(s,a) = \hat{r}(s,a) + \gamma\sigma_{\hat{\mathcal{P}}^a_s}(V^\pi_{\hat{\mathcal{P}}}) \leq \gamma V^\pi_{\hat{\mathcal{P}}}(s). \tag{73}$$

Thus

$$V^\pi_{\hat{\mathcal{P}}}(s) = \sum_a \pi(a|s)Q^\pi_{\hat{\mathcal{P}}}(s,a) \leq \gamma V^\pi_{\hat{\mathcal{P}}}(s), \tag{74}$$

which implies $V^\pi_{\hat{\mathcal{P}}}(s) = 0$ together with $V^\pi_{\hat{\mathcal{P}}} \geq 0$.

Proof of (2).

We prove Claim (2) by contradiction. Assume that for any optimal policy π_r, there exists $s \notin \mathcal{S}^0$ such that $N(s,\pi_r(s)) = 0$. We then consider a fixed pair (π_r, s).

$N(s,\pi_r(s)) = 0$ further implies $\hat{r}(s,\pi_r(s)) = 0$, $\hat{\mathcal{P}}^{\pi_r(s)}_s = \Delta(\mathcal{S})$, and

$$V^{\pi_r}_{\hat{\mathcal{P}}}(s) = \max_a Q^{\pi_r}_{\hat{\mathcal{P}}}(s,a) = Q^{\pi_r}_{\hat{\mathcal{P}}}(s,\pi_r(s)) = \hat{r}(s,\pi_r(s)) + \gamma\sigma_{\hat{\mathcal{P}}^{\pi_r(s)}_s}(V^{\pi_r}_{\hat{\mathcal{P}}}) \leq \gamma V^{\pi_r}_{\hat{\mathcal{P}}}(s), \tag{75}$$

where the last inequality is from $\hat{\mathcal{P}}_s^{\pi_r(s)} = \Delta(\mathcal{S})$, $\hat{r}(s, \pi_r(s)) = 0$, and $\sigma_{\hat{\mathcal{P}}_s^{\pi_r(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) \leq \mathbf{1}_s V_{\hat{\mathcal{P}}}^{\pi_r} = V_{\hat{\mathcal{P}}}^{\pi_r}(s)$. This further implies that $V_{\hat{\mathcal{P}}}^{\pi_r}(s) = 0$ because $V_{\hat{\mathcal{P}}}^{\pi_r} \geq 0$.

On the other hand, since $s \notin \mathcal{S}^0$, there exists another action $b \neq \pi_r(s)$ such that $N(s, b) > 0$, and hence $\hat{r}(s, b) = r(s, b)$. We consider the following two cases.

(I). If $r(s, b) > 0$, then

$$Q_{\hat{\mathcal{P}}}^{\pi_r}(s, b) = \hat{r}(s, b) + \gamma \sigma_{\hat{\mathcal{P}}_s^b}(V_{\hat{\mathcal{P}}}^{\pi_r}) > 0 = Q_{\hat{\mathcal{P}}}^{\pi_r}(s, \pi_r(s)), \tag{76}$$

which is contradict to $V_{\hat{\mathcal{P}}}^{\pi_r}(s) = \max_a Q_{\hat{\mathcal{P}}}^{\pi_r}(s, a) = Q_{\hat{\mathcal{P}}}^{\pi_r}(s, \pi_r(s))$.

(II). If $r(s, b) = 0$, Lemma 3 then implies the modified policy $f_b^s(\pi_r)$ is also optimal, and satisfies $N(x, f_b^s(\pi_r)(x)) = N(x, \pi_r(x))$ for any $x \neq s$, and $N(s, f_b^s(\pi_r)(s)) > 0$.

Then consider the modified policy $f_b^s(\pi_r)$.

If there still exists $s' \notin \mathcal{S}^0$ such that $N(s', f_b^s(\pi_r)(s')) = 0$, then similarly, there exists another action $b' \neq f_b^s(\pi_r)(s')$ such that $N(s', b') > 0$. Then whether $r(s', b') > 0$, which falls into Case (I) and leads to a contradiction, or applying Lemma 3 again implies another optimal policy $f_{b'}^{s'}(f_b^s(\pi_r))$, such that $N(s, f_{b'}^{s'}(f_b^s(\pi_r))(x)) = N(s, f_b^s(\pi_r)(x)) > 0$ for $x \notin \{s, s'\}$, $N(s, f_{b'}^{s'}(f_b^s(\pi_r))(s)) = N(s, f_b^s(\pi_r)(s)) > 0$ and $N(s', f_{b'}^{s'}(f_b^s(\pi_r))(s')) > 0$.

Repeating this procedure recursively further implies there exists an optimal policy π, such that $N(s, \pi(s)) > 0$ for any $s \notin \mathcal{S}^0$, which is a contraction to our assumption.

Therefore it completes the proof. □

Lemma 3. *For a robust optimal policy π_r, if there exists a state $s \notin \mathcal{S}^0$ and an action b such that $N(s, \pi_r(s)) = 0$, $r(s, b) = 0$ and $N(s, b) > 0$, define a modified policy $f_b^s(\pi_r)$ as*

$$f_b^s(\pi_r)(s) = b, \tag{77}$$

$$f_b^s(\pi_r)(x) = \pi_r(x), \text{ for } x \neq s. \tag{78}$$

Then the modified policy $f_b^s(\pi_r)$ is also optimal, and satisfies $N(s, f_b^s(\pi_r)(s)) > 0$, $N(x, f_b^s(\pi_r)(x)) = N(x, \pi_r(x)), \forall x \neq s$.

Proof. Recall that $\hat{\mathcal{P}}_s^{\pi_r(s)} = \Delta(\mathcal{S})$ and $\hat{\mathcal{P}}_s^b \subset \Delta(\mathcal{S})$, we have that

$$V_{\hat{\mathcal{P}}}^{f_b^s(\pi_r)} \geq V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)}, \tag{79}$$

where $\hat{\mathcal{P}}_s^b$ is a modified uncertainty set defined as

$$(\hat{\mathcal{P}}_s^b)_s^b = \Delta(\mathcal{S}), \tag{80}$$

$$(\hat{\mathcal{P}}_s^b)_x^a = \hat{\mathcal{P}}_x^a, \text{ for } (x, a) \neq (s, b). \tag{81}$$

Now we have that

$$V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)}(s) = Q_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)}(s, b) = r(s, b) + \gamma \sigma_{(\hat{\mathcal{P}}_s^b)_s^b}(V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)}) \leq \gamma V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)}(s), \tag{82}$$

which further implies $V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)}(s) = 0$. Note that in eq. (75), we have shown $V_{\hat{\mathcal{P}}}^{\pi_r}(s) = 0$, hence $V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)}(s) = V_{\hat{\mathcal{P}}}^{\pi_r}(s) = 0$.

Now consider the two robust Bellman operator $\mathbf{T}_b^s V(x) = \sum_a f_b^s(\pi_r)(a|x)(\hat{r}(x, a) + \gamma \sigma_{(\hat{\mathcal{P}}_s^b)_x^a}(V))$ and $\mathbf{T}V(x) = \hat{r}(x, \pi_r(x)) + \gamma \sigma_{\hat{\mathcal{P}}_x^{\pi_r(x)}}(V)$. It is known that $V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)}$ is the unique fixed point of the robust Bellman operator $\mathbf{T}_b^s$ and $V_{\hat{\mathcal{P}}}^{\pi_r}$ is the unique fixed point of $\mathbf{T}$.

When $x \neq s$,

$$\begin{aligned}
\mathbf{T}_b^s V_{\hat{\mathcal{P}}}^{\pi_r}(x) &= \sum_a f_b^s(\pi_r)(a|x)(\hat{r}(x,a) + \gamma\sigma_{(\hat{\mathcal{P}}_s^b)_x^a}(V_{\hat{\mathcal{P}}}^{\pi_r})) \\
&\overset{(a)}{=} \sum_a \pi_r(a|x)(\hat{r}(x,a) + \gamma\sigma_{(\hat{\mathcal{P}}_s^b)_x^a}(V_{\hat{\mathcal{P}}}^{\pi_r})) \\
&= \hat{r}(x,\pi_r(x)) + \gamma\sigma_{(\hat{\mathcal{P}}_s^b)_x^{\pi_r(x)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) \\
&\overset{(b)}{=} \hat{r}(x,\pi_r(x)) + \gamma\sigma_{\hat{\mathcal{P}}_x^{\pi_r(x)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) \\
&= \mathbf{T}V_{\hat{\mathcal{P}}}^{\pi_r}(x) = V_{\hat{\mathcal{P}}}^{\pi_r}(x),
\end{aligned} \tag{83}$$

where (a) is from $f_b^s(\pi_r)(x) = \pi_r(x)$ when $x \neq s$, (b) is from $(\hat{\mathcal{P}}_s^b)_x^{\pi_r(x)} = \hat{\mathcal{P}}_x^{\pi_r(x)}$.

And for s, it holds that

$$\begin{aligned}
\mathbf{T}_b^s V_{\hat{\mathcal{P}}}^{\pi_r}(s) &= \hat{r}(s,b) + \gamma\sigma_{(\hat{\mathcal{P}}_s^b)_s^b}(V_{\hat{\mathcal{P}}}^{\pi_r}) \\
&\overset{(a)}{=} \hat{r}(s,\pi_r(s)) + \gamma\sigma_{\Delta(\mathcal{S})}(V_{\hat{\mathcal{P}}}^{\pi_r}) \\
&\overset{(b)}{=} \hat{r}(s,\pi_r(s)) + \gamma\sigma_{\hat{\mathcal{P}}_s^{\pi_r(s)}}(V_{\hat{\mathcal{P}}}^{\pi_r}) \\
&= \mathbf{T}V_{\hat{\mathcal{P}}}^{\pi_r}(s) \\
&= V_{\hat{\mathcal{P}}}^{\pi_r}(s),
\end{aligned} \tag{84}$$

where (a) is from $(\hat{\mathcal{P}}_s^b)_s^b = \Delta(\mathcal{S})$ and $\hat{r}(s,b) = r(s,b) = 0 = \hat{r}(s,\pi_r(s))$, and (b) follows from the fact $\hat{\mathcal{P}}_s^{\pi_r(s)} = \Delta(\mathcal{S})$.

eq. (83) and eq. (84) further imply that $V_{\hat{\mathcal{P}}}^{\pi_r}$ is also a fixed point of $\mathbf{T}_b^s$. Hence it must be identical to $V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)}$, i.e., $V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)} = V_{\hat{\mathcal{P}}}^{\pi_r}$.

Combine with eq. (79), we have

$$V_{\hat{\mathcal{P}}}^{f_b^s(\pi_r)} \geq V_{\hat{\mathcal{P}}_s^b}^{f_b^s(\pi_r)} = V_{\hat{\mathcal{P}}}^{\pi_r}, \tag{85}$$

which implies that $f_b^s(\pi_r)$ is also optimal with $N(s, f_b^s(\pi_r)(s)) = N(s,b) > 0$. And since $f_b^s(\pi_r)(x) = \pi_r(x)$ for $x \neq s$, then $N(x, f_b^s(\pi_r)(x)) = N(x, \pi_r(x))$. This thus completes the proof. □

Lemma 4 (Lemma 4, (Li et al., 2022))**.** *For any δ, with probability $1-\delta$, $\max\{12N(s,a), 8\log\frac{NS}{\delta}\} \geq N\mu(s,a)$, $\forall s, a$.*

Lemma 5 (Lemma 9, (Li et al., 2022))**.** *For any (s,a) pair with $N(s,a) > 0$, if V is an vector independent of $\hat{\mathsf{P}}_s^a$ obeying $\|V\| \leq \frac{1}{1-\gamma}$, then with probability at least $1-\delta$,*

$$|(\hat{\mathsf{P}}_s^a - \mathsf{P}_s^a)V| \leq \sqrt{\frac{48\,\boldsymbol{Var}_{\hat{\mathsf{P}}_s^a}(V)\log\frac{4N}{\delta}}{N(s,a)}} + \frac{48\log\frac{4N}{\delta}}{(1-\gamma)N(s,a)}, \tag{86}$$

$$\boldsymbol{Var}_{\hat{\mathsf{P}}_s^a}(V) \leq 2\,\boldsymbol{Var}_{\mathsf{P}_s^a}(V) + \frac{5\log\frac{4N}{\delta}}{3(1-\gamma)^2 N(s,a)}. \tag{87}$$

Lemma 6. *Suppose $\gamma \in [0.5, 1)$, with probability at least $1-\delta$, it holds*

$$|(\hat{\mathsf{P}}_s^a - \mathsf{P}_s^a)V_{\hat{\mathcal{P}}}^{\pi_r}| \leq 12\sqrt{\frac{\boldsymbol{Var}_{\hat{\mathsf{P}}_s^a}(V_{\hat{\mathcal{P}}}^{\pi_r})\log\frac{4N}{\delta}}{N(s,a)}} + \frac{74\log\frac{4N}{\delta}}{(1-\gamma)N(s,a)}, \tag{88}$$

$$\boldsymbol{Var}_{\hat{\mathsf{P}}_s^a}(V_{\hat{\mathcal{P}}}^{\pi_r}) \leq 2\,\boldsymbol{Var}_{\mathsf{P}_s^a}(V_{\hat{\mathcal{P}}}^{\pi_r}) + \frac{41\log\frac{4N}{\delta}}{(1-\gamma)^2 N(s,a)}. \tag{89}$$

simultaneously for any pair $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Proof. When $N(s,a)=0$, the results hold naturally. We hence only consider (s,a) with $N(s,a)>0$. **Part 1**. Recall that $\hat{\mathsf{M}}=(\mathcal{S},\mathcal{A},\hat{\mathsf{P}},\hat{r})$ is the estimated MDP. For any state s and positive scalar $u>0$, we first construct an auxiliary state-absorbing MDP $\hat{\mathsf{M}}^{s,u}=(\mathcal{S},\mathcal{A},\mathsf{P}^{s,u},r^{s,u})$ as follows.

For all states except s, the MDP structure of $\hat{\mathsf{M}}^{s,u}$ is identical to $\hat{\mathsf{M}}$, i.e., for any $x\neq s$ and $a\in\mathcal{A}$,

$$(\mathsf{P}^{s,u})^a_{x,\cdot}=\hat{\mathsf{P}}^a_{x,\cdot},r^{s,u}(x,a)=\hat{r}(x,a); \tag{90}$$

State s is an absorbing state in $\hat{\mathsf{M}}^{s,u}$, namely, for any $a\in\mathcal{A}$,

$$(\mathsf{P}^{s,u})^a_{s,x}=\mathbf{1}_{x=s},r^{s,u}(s,a)=u. \tag{91}$$

We then define a robust MDP $\mathcal{M}^{s,u}=(\mathcal{S},\mathcal{A},\mathcal{P}^{s,u},r^{s,u})$ centered at $\hat{\mathsf{M}}^{s,u}$ as following: the uncertainty set $\mathcal{P}^{s,u}$ is defined as $\hat{\mathcal{P}}=\bigotimes_{x,a}(\mathcal{P}^{s,u})^a_x$, where if $x\neq s$,

$$(\mathcal{P}^{s,u})^a_x=\left\{q\in\Delta(\mathcal{S}):\|q-(\mathsf{P}^{s,u})^a_x\|\leq\min\left\{2,\frac{\log\frac{N}{\delta}}{N(x,a)}\right\}\right\} \tag{92}$$

and

$$(\mathcal{P}^{s,u})^a_s=\{\mathbf{1}_s\}. \tag{93}$$

The optimal robust value function of $\mathcal{M}^{s,u}$ is denoted by $V^{s,u}$.

Part 2. We claim that if we choose $u^*=(1-\gamma)V^{\pi_r}_{\hat{\mathcal{P}}}(s)$, then $V^{s,u^*}=V^{\pi_r}_{\hat{\mathcal{P}}}$. We prove this as follows.

Firstly note that the function V^{s,u^*} is the unique fixed point of the operator $\mathbf{T}^{s,u^*}(V)(x)=\max_a\{r^{s,u^*}(x,a)+\gamma\sigma_{(\mathcal{P}^{s,u^*})^a_x}(V)\}$.

For $x\neq s$, we note that

$$\begin{aligned}\mathbf{T}^{s,u^*}(V^{\pi_r}_{\hat{\mathcal{P}}})(x)&=\max_a\{r^{s,u^*}(x,a)+\gamma\sigma_{(\mathcal{P}^{s,u^*})^a_x}(V^{\pi_r}_{\hat{\mathcal{P}}})\}\\&=\max_a\{\hat{r}(x,a)+\gamma\sigma_{\hat{\mathcal{P}}^a_x}(V^{\pi_r}_{\hat{\mathcal{P}}})\}\\&=V^{\pi_r}_{\hat{\mathcal{P}}}(x),\end{aligned} \tag{94}$$

which is because $r^{s,u^*}(x,a)=\hat{r}(x,a)$ and $(\mathcal{P}^{s,u^*})^a_x=\hat{\mathcal{P}}^a_x$ for $x\neq s$.

For s, we have that

$$\begin{aligned}\mathbf{T}^{s,u^*}(V^{\pi_r}_{\hat{\mathcal{P}}})(s)&=\max_a\{r^{s,u^*}(s,a)+\gamma\sigma_{(\mathcal{P}^{s,u^*})^a_s}(V^{\pi_r}_{\hat{\mathcal{P}}})\}\\&=\max_a\{u^*+\gamma\sigma_{(\mathcal{P}^{s,u^*})^a_s}(V^{\pi_r}_{\hat{\mathcal{P}}})\}\\&=\max_a\{(1-\gamma)V^{\pi_r}_{\hat{\mathcal{P}}}(s)+\gamma(V^{\pi_r}_{\hat{\mathcal{P}}})(s)\}\\&=V^{\pi_r}_{\hat{\mathcal{P}}}(s),\end{aligned} \tag{95}$$

which from $(\mathcal{P}^{s,u^*})^a_s=\{\mathbf{1}_s\}$.

Hence combining with eq. (94) implies that $V^{\pi_r}_{\hat{\mathcal{P}}}$ is also a fixed point of $\mathbf{T}^{s,u^*}$, and hence it must be identical to V^{s,u^*}, which proves our claim.

Part 3. Define a set $U_c\triangleq\left\{\frac{i}{N}|i=1,...,N\right\}$. Clearly, U_c is a $\frac{1}{N}$-net (Vershynin, 2018; Li et al., 2022) of the interval $[0,1]$.

Note that for any $u\in U_c$, $\mathcal{P}^{s,u}$ is independent with $\hat{\mathsf{P}}^a_s$, hence $V^{s,u}$ is also independent with $\hat{\mathsf{P}}^a_s$. Also, since $u\leq 1$, $\|V^{s,u}\|\leq\frac{1}{1-\gamma}$.

Then invoking Lemma 5 implies that for any $N(s,a) > 0$, with probability at least $1-\delta$, it holds simultaneously for all $u \in U_c$ that

$$|(\hat{\mathsf{P}}_s^a - \mathsf{P}_s^a)V^{s,u}| \leq \sqrt{\frac{48\mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V^{s,u})\log\frac{4N^2}{\delta}}{N(s,a)}} + \frac{48\log\frac{4N^2}{\delta}}{(1-\gamma)N(s,a)}, \tag{96}$$

$$\mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V^{s,u}) \leq 2\mathbf{Var}_{\mathsf{P}_s^a}(V^{s,u}) + \frac{5\log\frac{4N^2}{\delta}}{3(1-\gamma)^2 N(s,a)}. \tag{97}$$

Part 4. Since $u^* = (1-\gamma)V_{\hat{\mathcal{P}}}^{\pi_r} \leq 1$, then there exists $u_0 \in U_c$, such that $|u_0 - u^*| \leq \frac{1}{N}$. Moreover, we claim that

$$\|V^{s,u^*} - V^{s,u_0}\| \leq \frac{1}{N(1-\gamma)}. \tag{98}$$

To prove eq. (98), first note that

$$\begin{aligned}
|V^{s,u^*}(s) - V^{s,u_0}(s)| &\leq \max_a |(u^* - u_0) + \gamma(\sigma_{(\mathcal{P}^{s,u^*})_s^a}(V^{s,u^*}) - \sigma_{(\mathcal{P}^{s,u_0})_s^a}(V^{s,u_0}))| \\
&\overset{(a)}{\leq} |u^* - u_0| + \gamma \max_a |\sigma_{(\mathcal{P}^{s,u^*})_s^a}(V^{s,u^*}) - \sigma_{(\mathcal{P}^{s,u^*})_s^a}(V^{s,u_0})| \\
&\overset{(b)}{\leq} |u^* - u_0| + \gamma\|V^{s,u^*} - V^{s,u_0}\|,
\end{aligned} \tag{99}$$

where (a) is because $(\mathcal{P}^{s,u_0})_s^a = (\mathcal{P}^{s,u^*})_s^a = \{\mathbf{1}_s\}$, and (b) is due to the non-expansion of the support function (Lemma 1, (Panaganti & Kalathil, 2022)).

For $x \neq s$, we have that

$$\begin{aligned}
|V^{s,u^*}(x) - V^{s,u_0}(x)| &\leq \max_a |\hat{r}(x,a) - \hat{r}(x,a) + \gamma(\sigma_{\hat{\mathcal{P}}_x^a}(V^{s,u^*}) - \sigma_{\hat{\mathcal{P}}_x^a}(V^{s,u_0}))| \\
&\leq \gamma\|V^{s,u^*} - V^{s,u_0}\|.
\end{aligned} \tag{100}$$

Thus by combining eq. (99) and eq. (100), we have that

$$\|V^{s,u^*} - V^{s,u_0}\| \leq \frac{1}{N} + \gamma\|V^{s,u^*} - V^{s,u_0}\|, \tag{101}$$

and hence proof the claim eq. (98).

Therefore,

$$\begin{aligned}
&\mathbf{Var}_{\mathsf{P}_s^a}(V^{s,u_0}) - \mathbf{Var}_{\mathsf{P}_s^a}(V^{s,u^*}) \\
&= \mathsf{P}_s^a\Big((V^{s,u_0} - \mathsf{P}_s^a V^{s,u_0}) \circ (V^{s,u_0} - \mathsf{P}_s^a V^{s,u_0}) - (V^{s,u^*} - \mathsf{P}_s^a V^{s,u^*}) \circ (V^{s,u^*} - \mathsf{P}_s^a V^{s,u^*})\Big) \\
&\overset{(a)}{\leq} \mathsf{P}_s^a\Big((V^{s,u_0} - \mathsf{P}_s^a V^{s,u^*}) \circ (V^{s,u_0} - \mathsf{P}_s^a V^{s,u^*}) - (V^{s,u^*} - \mathsf{P}_s^a V^{s,u^*}) \circ (V^{s,u^*} - \mathsf{P}_s^a V^{s,u^*})\Big) \\
&\leq \mathsf{P}_s^a\Big((V^{s,u_0} - \mathsf{P}_s^a V^{s,u^*} + V^{s,u^*} - \mathsf{P}_s^a V^{s,u^*}) \circ (V^{s,u_0} - V^{s,u^*})\Big) \\
&\leq \frac{2}{1-\gamma}|\mathsf{P}_s^a(V^{s,u_0} - V^{s,u^*})| \\
&\leq \frac{2}{N(1-\gamma)^2},
\end{aligned} \tag{102}$$

where (a) is due to the fact $\mathbb{E}[X] = \arg\min_c \mathbb{E}[(X-c)^2]$, and the last inequality is due to $\|V^{s,u_0}\| \leq \frac{1}{1-\gamma}, \|V^{s,u^*}\| \leq \frac{1}{1-\gamma}$.

Similarly, swapping V^{s,u_0} and V^{s,u^*} implies

$$\mathbf{Var}_{\mathsf{P}_s^a}(V^{s,u^*}) - \mathbf{Var}_{\mathsf{P}_s^a}(V^{s,u_0}) \leq \frac{2}{N(1-\gamma)^2}, \tag{103}$$

and further

$$|\mathbf{Var}_{\mathsf{P}_s^a}(V^{s,u^*}) - \mathbf{Var}_{\mathsf{P}_s^a}(V^{s,u_0})| \leq \frac{2}{N(1-\gamma)^2}. \tag{104}$$

We note that eq. (104) is exactly identical to (159) in Section A.4 of (Li et al., 2022), and hence the remaining proof can be obtained by following the proof in Section A.4 in (Li et al., 2022), and are omitted here. □

Lemma 7. *With probability at least $1-\delta$, it holds that for any s, a,*

$$\begin{aligned}
&\hat{\mathsf{P}}_s^a V_{\hat{\mathcal{P}}}^{\pi_r} - \mathsf{P}_s^a V_{\hat{\mathcal{P}}}^{\pi_r} \\
&\leq \hat{\mathsf{P}}_s^a V_{\hat{\mathcal{P}}}^{\pi_r} - \sigma_{\hat{\mathcal{P}}_s^a}(V_{\hat{\mathcal{P}}}^{\pi_r}) + 2\sqrt{\frac{48\log\frac{4SAN}{(1-\gamma)\delta}\epsilon_1}{(1-\gamma)N(s,a)}} + 2\epsilon_1\sqrt{\frac{48\log\frac{4SAN}{(1-\gamma)\delta}}{N(s,a)}} + \frac{96\log\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}.
\end{aligned} \tag{105}$$

Proof. We first show the inequality above holds for any $V \in \left[0, \frac{1}{1-\gamma}\right]$ that is independent with $\hat{\mathsf{P}}_s^a$.

From the duality form of the $\sigma_{\mathcal{P}}(V)$ (Iyengar, 2005), it holds that

$$\begin{aligned}
\hat{\mathsf{P}}_s^a V - \sigma_{\hat{\mathcal{P}}_s^a}(V) &= \hat{\mathsf{P}}_s^a V - \max_{\alpha\in[V_{\min},V_{\max}]}\left\{\hat{\mathsf{P}}_s^a V_\alpha - \sqrt{R_s^a \mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_\alpha)}\right\} \\
&= \min_{\alpha\in[V_{\min},V_{\max}]}\left\{\hat{\mathsf{P}}_s^a (V - V_\alpha) + \sqrt{R_s^a \mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_\alpha)}\right\},
\end{aligned} \tag{106}$$

where $V_\alpha \in \mathbb{R}^S$ and $V_\alpha(s) = \min\{V(s), \alpha\}$.

We denote the optimum of the optimization by α^*, i.e.,

$$\hat{\mathsf{P}}_s^a (V - V_{\alpha^*}) + \sqrt{R_s^a \mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_{\alpha^*})} = \min_{\alpha\in[V_{\min},V_{\max}]}\left\{\hat{\mathsf{P}}_s^a (V - V_\alpha) + \sqrt{R_s^a \mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_\alpha)}\right\}. \tag{107}$$

Then eq. (106) can be further bounded as

$$\begin{aligned}
\hat{\mathsf{P}}_s^a V - \sigma_{\hat{\mathcal{P}}_s^a}(V) &= \hat{\mathsf{P}}_s^a (V - V_{\alpha^*}) + \sqrt{R_s^a \mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_{\alpha^*})} \\
&\geq \hat{\mathsf{P}}_s^a (V - V_{\alpha^*}) - \mathsf{P}_s^a (V - V_{\alpha^*}) + \sqrt{R_s^a \mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_{\alpha^*})},
\end{aligned} \tag{108}$$

where the inequality is due to the fact that $V(s) \geq V_\alpha(s)$.

On the other hand, for any $\alpha \in [V_{\min}, V_{\max}]$ that is fixed and independent with $\hat{\mathsf{P}}_s^a$, we have that

$$\begin{aligned}
\hat{\mathsf{P}}_s^a V - \mathsf{P}_s^a V &= \hat{\mathsf{P}}_s^a V_\alpha - \mathsf{P}_s^a V_\alpha + \hat{\mathsf{P}}_s^a (V - V_\alpha) - \mathsf{P}_s^a (V - V_\alpha) \\
&\leq \hat{\mathsf{P}}_s^a (V - V_\alpha) - \mathsf{P}_s^a (V - V_\alpha) + \frac{48\log\frac{4N}{\delta}}{(1-\gamma)N(s,a)} + \sqrt{\frac{48\log\frac{4N}{\delta}\mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_\alpha)}{N(s,a)}},
\end{aligned} \tag{109}$$

which is due to α is independent from $\hat{\mathsf{P}}_s^a$, and applying the Bernstein's inequality and (102) of (Shi et al., 2023). Moreover, it implies that

$$\begin{aligned}
&\hat{\mathsf{P}}_s^a V - \mathsf{P}_s^a V \\
&\leq \hat{\mathsf{P}}_s^a (V - V_{\alpha^*}) - \mathsf{P}_s^a (V - V_{\alpha^*}) + \frac{48\log\frac{4N}{\delta}}{(1-\gamma)N(s,a)} + \sqrt{\frac{48\log\frac{4N}{\delta}\mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_{\alpha^*})}{N(s,a)}} \\
&\quad + \left(\sqrt{\frac{48\log\frac{4N}{\delta}\mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_\alpha)}{N(s,a)}} - \sqrt{\frac{48\log\frac{4N}{\delta}\mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_{\alpha^*})}{N(s,a)}}\right) + (\hat{\mathsf{P}}_s^a (V_{\alpha^*} - V_\alpha) - \mathsf{P}_s^a (V_{\alpha^*} - V_\alpha)).
\end{aligned} \tag{110}$$

We now construct an ϵ_1-Net (Vershynin, 2018) of $\left[0, \frac{1}{1-\gamma}\right]$ with $\epsilon_1 = \frac{1}{N}$. Specifically, there exists $\mathcal{U} = \left\{\alpha_1, \alpha_2, ..., \alpha_m | \alpha_i \in \left[0, \frac{1}{1-\gamma}\right]\right\}$, such that for any $\alpha \in \left[0, \frac{1}{1-\gamma}\right]$, there exists $\alpha_j \in \mathcal{U}$ with $|\alpha - \alpha_j| \leq \epsilon_1$. Since $\alpha^* \in \left[0, \frac{1}{1-\gamma}\right]$, there exists $\beta \in \mathcal{U}$ with $|\beta - \alpha^*| \leq \epsilon_1$.

It is straightforward to see that

$$|V_{\alpha^*} - V_\beta| \leq |\beta - \alpha^*| \leq \epsilon_1, \tag{111}$$

and similarly following (207) of (Shi et al., 2023) implies that

$$\left|\sqrt{\mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_\beta)} - \sqrt{\mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_{\alpha^*})}\right| \leq 2\sqrt{\frac{\epsilon_1}{1-\gamma}}. \tag{112}$$

Hence we set $\alpha = \beta$ in eq. (110), take the union bound over $\mathcal{S}, \mathcal{A}$ and $\mathcal{U}$, and plug in the two inequalities above, we have that

$$\begin{aligned}
&\hat{\mathsf{P}}_s^a V - \mathsf{P}_s^a V \\
&\leq \hat{\mathsf{P}}_s^a (V - V_{\alpha^*}) - \mathsf{P}_s^a (V - V_{\alpha^*}) + \frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} + \sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta} \mathbf{Var}_{\hat{\mathsf{P}}_s^a}(V_{\alpha^*})}{N(s,a)}} \\
&\quad + 2\sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta} \epsilon_1}{(1-\gamma)N(s,a)}} + 2\epsilon_1 \sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{N(s,a)}}.
\end{aligned} \tag{113}$$

Involving eq. (108) further implies that

$$\begin{aligned}
&\hat{\mathsf{P}}_s^a V - \mathsf{P}_s^a V \\
&\leq \hat{\mathsf{P}}_s^a V - \sigma_{\hat{\mathcal{P}}_s^a}(V) + 2\sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta} \epsilon_1}{(1-\gamma)N(s,a)}} + 2\epsilon_1 \sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{N(s,a)}} + \frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)},
\end{aligned} \tag{114}$$

and hence

$$\sigma_{\hat{\mathcal{P}}_s^a}(V) - \mathsf{P}_s^a V \leq 2\sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta} \epsilon_1}{(1-\gamma)N(s,a)}} + 2\epsilon_1 \sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{N(s,a)}} + \frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}. \tag{115}$$

Now we consider $V_{\hat{\mathcal{P}}}^{\pi_r}$. Following the $\frac{1}{N}$-net $\mathcal{U}_2$ we constructed in Lemma 6, $V_{\hat{\mathcal{P}}}^{\pi_r} = V^{s,u}$ for some $u \in [0,1]$. Hence there exists some $u_i \in \mathcal{U}_2$ with $|u - u_i| \leq \frac{1}{N}$. Note that

$$\begin{aligned}
&\sigma_{\hat{\mathcal{P}}_s^a}(V_{\hat{\mathcal{P}}}^{\pi_r}) - \mathsf{P}_s^a V_{\hat{\mathcal{P}}}^{\pi_r} \\
&= \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u}) - \mathsf{P}_s^a V^{s,u} \\
&= \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u_i}) - \mathsf{P}_s^a V^{s,u_i} \\
&\quad + \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u}) - \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u_i}) + \mathsf{P}_s^a V^{s,u_i} - \mathsf{P}_s^a V^{s,u} \\
&\leq 2\sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta} \epsilon_1}{(1-\gamma)N(s,a)}} + 2\epsilon_1 \sqrt{\frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{N(s,a)}} + \frac{48 \log \frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} \\
&\quad + \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u}) - \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u_i}) + \mathsf{P}_s^a V^{s,u_i} - \mathsf{P}_s^a V^{s,u},
\end{aligned} \tag{116}$$

which is due to $V^{s,u}$ is independent with $\hat{\mathsf{P}}_s^a$. Moreover, note that both $\sigma_{\hat{\mathcal{P}}_s^a}(V)$ and $\hat{\mathsf{P}}_s^a V$ are 1-Lipschitz, thus

$$\sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u}) - \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u_i}) \leq \|V^{s,u} - V^{s,u_i}\|, \tag{117}$$

$$\mathsf{P}_s^a V^{s,u_i} - \mathsf{P}_s^a V^{s,u} \leq \|V^{s,u} - V^{s,u_i}\|. \tag{118}$$

Now from Equation (98), it follows that

$$\sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u}) - \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u_i}) \leq \frac{1}{N(1-\gamma)}, \tag{119}$$

$$\mathsf{P}_s^a V^{s,u_i} - \mathsf{P}_s^a V^{s,u} \leq \frac{1}{N(1-\gamma)}. \tag{120}$$

Hence combine with Equation (116), and we have that

$$\begin{aligned}
&\sigma_{\hat{\mathcal{P}}_s^a}(V_{\hat{\mathcal{P}}}^{\pi_r}) - \mathsf{P}_s^a V_{\hat{\mathcal{P}}}^{\pi_r} \\
&\leq 2\sqrt{\frac{48\log\frac{4SAN}{(1-\gamma)\delta}\epsilon_1}{(1-\gamma)N(s,a)}} + 2\epsilon_1\sqrt{\frac{48\log\frac{4SAN}{(1-\gamma)\delta}}{N(s,a)}} + \frac{48\log\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} \\
&\quad + \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u}) - \sigma_{\hat{\mathcal{P}}_s^a}(V^{s,u_i}) + \mathsf{P}_s^a V^{s,u_i} - \mathsf{P}_s^a V^{s,u} \\
&\leq 2\sqrt{\frac{48\log\frac{4SAN}{(1-\gamma)\delta}\epsilon_1}{(1-\gamma)N(s,a)}} + 2\epsilon_1\sqrt{\frac{48\log\frac{4SAN}{(1-\gamma)\delta}}{N(s,a)}} + \frac{48\log\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} + \frac{2}{N(1-\gamma)} \\
&\leq 2\sqrt{\frac{48\log\frac{4SAN}{(1-\gamma)\delta}\epsilon_1}{(1-\gamma)N(s,a)}} + 2\epsilon_1\sqrt{\frac{48\log\frac{4SAN}{(1-\gamma)\delta}}{N(s,a)}} + \frac{96\log\frac{4SAN}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)},
\end{aligned} \tag{121}$$

which completes the proof. □