Near-Optimal Sample Complexity for Iterated CVaR Reinforcement Learning with a Generative Model

Zilong Deng
ECEE, Arizona State University

Simon Khan
Air Force Research Laboratory

Shaofeng Zou
ECEE, Arizona State University

Abstract

In this work, we study the sample complexity problem of risk-sensitive Reinforcement Learning (RL) with a generative model, where we aim to maximize the Conditional Value at Risk (CVaR) with risk tolerance level τ at each step, named Iterated CVaR. We develop nearly matching upper and lower bounds on the sample complexity for this problem. Specifically, we first prove that a value iteration-based algorithm, ICVaR-VI, achieves an ϵ-optimal policy with at most $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^4\tau^2\epsilon^2}\right)$ samples, where γ is the discount factor, and S, A are the sizes of the state and action spaces. Furthermore, if $\tau \geq \gamma$, then the sample complexity can be further improved to $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$. We further show a minimax lower bound of $\tilde{\mathcal{O}}\left(\frac{(1-\gamma\tau)SA}{(1-\gamma)^4\tau^2}\right)$. For a constant risk level $0 < \tau \leq 1$, our upper and lower bounds match with each other, demonstrating the tightness and optimality of our analyses. We also investigate a limiting case with a small risk level τ, called Worst-Path RL, where the objective is to maximize the minimum possible cumulative reward. We develop matching upper and lower bounds of $\tilde{\mathcal{O}}\left(\frac{SA}{p_{\min}}\right)$, where $p_{\min}$ denotes the minimum non-zero reaching probability of the transition kernel.

1 Introduction

Reinforcement learning (RL) (Sutton & Barto, 2018) is a foundational framework for solving sequential decision-making problems, and it finds a wide range of applications in e.g., large language models (Ouyang et al., 2022), robotics (Kober et al., 2013), and finance (Charpentier et al., 2021). Recently there has been a surge of interest in understanding its fundamental sample complexity, e.g., Bhandari et al. (2018); Srikant & Ying (2019); Zou et al. (2019); Agarwal et al. (2021); Bhandari & Russo (2024). However, the main focus was on the risk-neural setting, where the objective is to maximize the expected total rewards accumulated over time. As RL is increasingly applied to real-world sequential decision-making tasks, it often becomes essential to account for risk rather than simply optimizing for the total reward. This is particularly important when safety and worst-case avoidance considerations exist, e.g., in finance and investment, healthcare, autonomous systems, and process engineering. Nevertheless, a fundamental understanding of the sample complexity for risk-sensitive RL remains largely unexplored. In this paper, we focus on one of such problems and aim to understand the fundamental sample complexity for iterated CVaR RL with a generative model.

A widely used risk measure is called coherent risk measure, which satisfies the following four properties: (i) monotonicity;(ii) translation invariance; (iii) sub-additivity; and (iv) positive homogeneity (Artzner et al., 1999). Conditional Value at Risk (CVaR) is a popular coherent risk measure(Rockafellar et al., 2000). There are two types of CVaR RL objectives, Iterated (dynamic) CVaR RL (Hardy & Wirch, 2004) and Static CVaR RL (Wang et al., 2023; Zhao et al., 2023). Iterated CVaR is a special case of Markov coherent risk (Tamar et al., 2015) where the coherent risk measure is CVaR. It has an iterative structure and focuses on the worst portion of the reward at each step. In Iterated CVaR RL problem, the agent aims to maximize the average of the worst portion at every step. Static CVaR RL (Wang et al., 2023; Zhao et al., 2023)) on the other hand, aims to maximize the CVaR of the total reward.

Static CVaR RL and Iterated CVaR RL are quite different. In Static CVaR RL, the optimal policy is

history-dependent and not stationary. In Bäuerle & Ott (2011), it is shown that Static CVaR RL can be optimally solved by resolving to standard, risk-neutral RL in an augmented MDP. In Iterated CVaR RL, the optimal policy is Markovian and stationary if the environment is stationary. Broadly speaking, Static CVaR RL is more concerned with the overall risk of the total reward and may permit the agent to visit catastrophic states as long as the risk of the cumulative reward remains acceptable. In contrast, Iterated CVaR RL assesses risk at each step, offering a more cautious approach by preventing the agent from entering catastrophic states at any point in the trajectory.

In this paper, we focus on Iterated CVaR RL, and we aim to theoretically understand the sample complexity with access to a generator (Kearns & Singh, 1998). With access to a generative model, one can draw samples from the transition kernel of a Markov decision process (MDP) conditioned on any arbitrary state-action pair. We then take a model-based approach, where we first construct a maximum likelihood estimate of the transition kernel and find the optimal policy for the estimated MDP model. In risk-neural RL, it was shown that such an approach achieves the minimax optimal sample complexity (Gheshlaghi Azar et al., 2013; Agarwal et al., 2020). In Iterated CVaR RL, the goal is to find the policy that maximizes the worst τ-portion of reward-to-go at each step. Clearly, this ensures that policy avoids getting into catastrophic states, prioritizing risk-sensitive behavior. Since the objective focuses on ignoring certain "good" states and prioritizing the worst-case scenarios, we undoubtedly require more samples to learn a safe policy. This naturally raises the question: how many samples are needed to produce an optimal risk-sensitive policy?

In this paper, we make the connection between Iterated CVaR RL and distributionally robust RL (Iyengar, 2005; Nilim & El Ghaoui, 2004) using the dual form of CVaR. We then decompose the error using approaches in robust RL (Shi et al., 2023) and develop novel analytical techniques to bound the change in CVaR resulting from approximation error. We prove that a value iteration-based algorithm, ICVaR-VI, achieves an ϵ-optimal policy with at most $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^4\tau^2\epsilon^2}\right)$ samples, where γ is the discount factor, τ is the risk tolerance, and S, A are the sizes of the state and action spaces. Moreover, if $\tau \geq \gamma$, we further derive an improved sample complexity of $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$, which actually matches with the minimax optimal sample complexity for risk-neural RL (also see Table 1). We then develop a minimax lower bound that for any τ and γ, there always exists an

MDP, for which at least $\tilde{\mathcal{O}}\left(\frac{(1-\gamma\tau)SA}{(1-\gamma)^4\tau\epsilon^2}\right)$ are needed. Comparing our upper and lower bounds, they match with each other in the order of sizes of state and action spaces S, A and effective horizon $(1-\gamma)^{-1}$ when the risk level is a constant in $(0, 1]$.

Finally, we investigate a limiting case, named Worst-Path RL (Du et al., 2022), where we consider a small risk level τ smaller than the minimum non-zero reaching probability of the transition kernel $p_{\min}$. In this case, the CVaR risk measure actually tries to find the worst-case state. This problem cannot be directly solved by taking the limit $\tau \to 0$, and the previous lower and upper bounds will also go to infinity as they depend on τ^{-1}. To tackle this problem, we design a new algorithm based on the reduced Bellman operator and develop matching upper and lower bounds of $\tilde{\mathcal{O}}\left(\frac{SA}{p_{\min}}\right)$, where $p_{\min}$ denotes the minimum non-zero reaching probability of the transition kernel.

2 Related Work

Static CVaR RL: There is a long line of works focusing on static CVaR RL, which refers to the CVaR (i.e. the worst τ portion) of the accumulated total reward. Bastani et al. (2022) proved the first regret bound, and Wang et al. (2023) improved the results to be minimax optimal. Zhao et al. (2023) introduced function approximation to the MDP structure and studied static CVaR in low-rank MDPs. Additionally, Ni et al. (2024) developed the sample complexity of reward-free exploration in static CVaR. However, static CVaR is intrinsically different from the iterated CVaR RL studied in this paper. Iterated CVaR RL concerns the worst τ-percent of the reward-to-go at each step. Intuitively, static CVaR takes more cumulative reward into account and prefers actions that have better performance in general, while iterated CVaR prevents the agent from getting into catastrophic states (Du et al., 2022). Therefore, the algorithm designs and analysis techniques introduced above can not be applied to our problem.

Iterated CVaR RL: Hardy & Wirch (2004) first introduced the Iterated CVaR and showed that it is a coherent risk measure. Osogami (2012), Chu & Zhang (2014) and Bäuerle & Glauner (2022) studied the iterated coherent risk measures (iterated CVaR included) and proved the existence of a Markovian deterministic optimal policy. Bäuerle & Glauner (2022) also established a connection between iterated coherent risk measures and distributional robust MDPs. Chen et al. (2023) studied iterated CVaR with function approximation and human feedback.

For the more general iterated coherent risk measures

Zilong Deng, Simon Khan, Shaofeng Zou

Problem	Lower bound	Upper bound
Risk-neutral (Agarwal et al., 2020)	$\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$
Episodic, Iterated CVaR (Du et al., 2022)	$\tilde{\mathcal{O}}\left(\frac{A}{\epsilon^2\tau^{(1-\gamma)^{-1}-1}}\right)$	$\tilde{\mathcal{O}}\left(\frac{S^2A}{(1-\gamma)^4\tau^{(1-\gamma)^{-1}+1}\epsilon^2}\right)$
Episodic, Iterated OCE (Xu et al., 2023)	$\mathcal{O}\left(\frac{SA}{(1-\gamma)^3\tau\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{\left(\tau^{-\frac{1}{2}(1-\gamma)^{-1}}-1-(1-\gamma)^{-1}(\tau^{-1/2}-1)\right)^2 S^2A}{(1-\sqrt{\tau})^4\epsilon^2}\right)$
Infinite horizon, Iterated CVaR (**This work**)	$\tilde{\mathcal{O}}\left(\frac{(1-\gamma\tau)SA}{(1-\gamma)^4\tau\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^4\tau^2\epsilon^2}\right)$ $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$ if $\tau\geq\gamma$

Table 1: Comparison of lower and upper bound for RL with Iterative Risk Measures. Some of the results are presented in terms of regret bound, and we converted them to sample complexity for the ease of comparison.

(Markov coherent risk), Tamar et al. (2015) derived the policy gradient algorithm for both static and dynamic (iterated) coherent risk measures. Huang et al. (2021) then proved that gradient dominant doesn't hold for iterated coherent risk measures and that stationary point is not guaranteed to be globally optimal.

To the best of our knowledge, Du et al. (2022) is the most related work to ours, where the Iterated CVaR objective in the episodic setting was studied. Their sample complexity results are listed in Table 1. In this paper, we obtain tighter upper and lower bounds, which are polynomial in the effective horizon $(1-\gamma)^{-1}$, and have a better dependence on the number of states S. Moreover, our bounds are minimax optimal for almost all choices of risk level. Xu et al. (2023) studied the recursive optimized certainty equivalent (OCE) problem in an episodic setting, where OCE is a more generalized risk measure, including CVaR. For Iterated CVaR RL, the upper and lower bounds in their paper are listed in Table 1. Compared to their results, our upper and lower bounds are much tighter (minimax optimal for almost all choices of risk level) and have a clear and easy-to-understand dependence on relevant factors.

Sample Complexity of Distributionally Robust RL with a Generative Model: The Iterative CVar RL problem can be equivalently written as a distributionally robust RL problem with a certain uncertainty set. The fundamental sample complexity for distributionally robust RL with a generative model has been studied in the literature for uncertainty sets defined by e.g., total variation (Shi et al., 2023; Panaganti & Kalathil, 2022; Yang et al., 2022), χ^2-divergence (Shi et al., 2023; Panaganti & Kalathil, 2022; Yang et al., 2022), Kullback-Leibler (KL) divergence (Shi & Chi, 2024).

3 Preliminaries and Problem Formulation

Notations. We denote by $\Delta(\mathcal{S})$ the probability simplex over a set $\mathcal{S}$. In this work, we use the standard $\mathcal{O}(\cdot)$ notation to hide universal constant factors and use $\tilde{\mathcal{O}}(\cdot)$ to hide logarithmic factors.

Conditional Value-at-Risk (CVaR). We begin by introducing two risk measures: value-at-risk (VaR) and conditional value-at-risk (CVaR). Let Z be a random variable with cumulative distribution function $F_Z(z) = P(Z \leq z)$. The Value-at-risk at risk level $\tau \in (0, 1]$ is defined as

$$\text{VaR}_\tau(Z) = \inf\left\{z : F_Z(z) \geq \tau\right\}. \quad (1)$$

The conditional value-at-risk at risk level $\tau \in (0, 1]$ is defined as

$$\text{CVaR}_\tau(Z) = \sup_{s\in\mathbb{R}}\left\{s - \frac{1}{\tau}\mathbb{E}\left[(s - Z)_+\right]\right\}, \quad (2)$$

where $(x)_+ = \max\{x, 0\}$ for some $x \in \mathbb{R}$. If $F_Z(z)$ is continuous at $\text{VaR}_\tau(Z)$, then CVaR can also be equivalently written as (Shapiro et al., 2021):

$$\text{CVaR}_\tau(Z) = \mathbb{E}\left[Z|Z \leq \text{VaR}_\tau(Z)\right]. \quad (3)$$

From the above equation, CVaR can be viewed as the average of the worst τ-fraction of Z. When $\tau = 1$, $\text{CVaR}_\tau(Z) = \mathbb{E}[Z]$, and when $\tau \to 0$, $\text{CVaR}_\tau(Z) \to \text{ess}\inf(Z)$.

When we need to specify the distribution of random variable Z, we write $\text{CVaR}(Z)$ as

$$\text{CVaR}_{Z\sim P}^\tau(Z) = \sup_{s\in\mathbb{R}}\left\{s - \frac{1}{\tau}\mathbb{E}_{Z\sim P}\left[(s - Z)_+\right]\right\}, \quad (4)$$

where P is the distribution of Z.

CVaR can be equivalently written in the dual formulation (Shapiro et al., 2021) using the risk envelope

$$\mathcal{U}_{\text{CVaR}_\tau}(P) = \left\{\xi : \xi \in \left[0, \frac{1}{\tau}\right], \ \mathbb{E}_P\left[\xi\right] = 1\right\}, \quad (5)$$

and

$$\text{CVaR}_\tau(Z) = \inf_{\xi \in \mathcal{U}_{\text{CVaR}_\tau}(P)} \mathbb{E}_P[\xi Z]. \tag{6}$$

Standard Markov Decision Process. We denote an infinite horizon discounted Markov decision process (MDP) by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, $\gamma \in [0,1)$ is the discount factor, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes the transition kernel that maps a state-action pair to a probability distribution over $\mathcal{S}$, and $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the deterministic reward function. Let S and A denote the sizes of the state and action spaces, respectively. A stationary policy is defined by $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. The value function of a policy π for state s is defined by

$$\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s \right], \forall s \in \mathcal{S}, \tag{7}$$

where s_t and a_t denote the state and action at step t.

Iterated CVaR Objective. The risk-neutral objective defined in (7) fails to consider the risks arising from the stochastic nature of state transitions and the agent's policy decisions. Iterated coherent risk measures (Chu & Zhang, 2014) have been introduced to model and evaluate these types of risks. For notational convenience, let

$$r(s, \pi) := \sum_{a \in \mathcal{A}} \pi(a|s) r(s, a), \tag{8}$$

$$\rho_{s,\pi}(Z(s')) := \sum_{a \in \mathcal{A}} \pi(a|s) \rho_{s,a}(Z(s')), \tag{9}$$

where $\rho_{s,a}$ is a one-step coherent risk measure indexed by $(s, a) \in \mathcal{S} \times \mathcal{A}$ and the distribution of s' follows the transition probability of $P(\cdot|s,a)$. The objective of the risk-sensitive MDP is defined as follows:

$$\max_\pi V^\pi(s_0), \tag{10}$$

where $V^\pi(s_0) = r(s_0, \pi) + \gamma \rho_{s_0,\pi}(r(s_1, \pi) + \gamma \rho_{s_1,\pi}(r(s_2, \pi) + ...))$. The trajectory $\{s_0, s_1, s_2, ...\}$ is generated from the MDP $\mathcal{M}$ and policy π. The objective V^π is defined in a nested pattern rather than through a single static measure of the total discounted reward.

In this paper, we focus on a specific risk measure, the Conditional Value-at-Risk (CVaR), and refer to the objective in (10) as the iterated CVaR objective. For notational convenience, we denote

$$\text{CVaR}_{s,\pi}^\tau(Z(s')) := \sum_{a \in \mathcal{A}} \pi(a|s) \text{CVaR}_{s' \sim P(\cdot|s,a)}^\tau(Z(s')). \tag{11}$$

The value function and Q-function are defined as follows:

$$V^\pi(s_0) = r(s_0, \pi) + \gamma \text{CVaR}_{s_0,\pi}^\tau \bigg(r(s_1, a_1)$$
$$+ \gamma \text{CVaR}_{s_1,\pi}^\tau(r(s_2, \pi) + ...) \bigg), \tag{12}$$

$$Q^\pi(s_0, a_0) = r(s_0, a_0) + \gamma \text{CVaR}_{s_1 \sim P(\cdot|s_0,a_0)}^\tau(V^\pi(s_1)). \tag{13}$$

The choice of the risk-sensitive objective function in (12) guarantees the existence of an optimal policy and the optimal policy is Markovian (Chu & Zhang, 2014). In contrast, the static CVaR objective, which only applies the risk measure to the total discounted reward once, does not have this property.

Optimal Risk-sensitive Policy and Bellman Operator. As shown in Chu & Zhang (2014), there exists a deterministic stationary optimal policy π^* that maximizes the risk-sensitive value function simultaneously for all states:

$$\forall s \in \mathcal{S} : V^*(s) := V^{\pi^*}(s) = \max_\pi V^\pi(s) \tag{14}$$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : Q^*(s, a) := Q^{\pi^*}(s, a) = \max_\pi Q^\pi(s, a). \tag{15}$$

The corresponding Bellman (optimality) equations are as follows:

$$Q^\pi(s, a) = r(s, a) + \gamma \text{CVaR}_\tau(V^\pi(s')), \tag{16a}$$
$$Q^*(s, a) = r(s, a) + \gamma \text{CVaR}_\tau(V^*(s')), \tag{16b}$$

where $s' \sim P(\cdot|s, a)$. The Bellman operator is denoted by $\mathcal{T}^\tau : \mathbb{R}^{SA} \to \mathbb{R}^{SA}$ and defined as follows:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \mathcal{T}^\tau(Q)(s, a) := r(s, a) + \gamma \text{CVaR}_\tau V(s),$$
$$\text{with} \quad V(s) := \max_a Q(s, a). \tag{17}$$

Since $Q^*(s, a)$ is the unique fixed point of $\mathcal{T}^\tau$, we can recover the optimal policy using a value iteration algorithm (see Algorithm 1). This converges rapidly due to the γ-contraction property of the $\mathcal{T}^\tau$ operator w.r.t. the l_∞ norm (Lemma 1).

Connection to Distributional Robust RL. Applying the dual form of CVaR, the Bellman equation can be re-written as

$$Q^\pi(s, a) = r(s, a) + \gamma \inf_{\xi \in \mathcal{U}_{\text{CVaR}_\tau}(P)} \mathbb{E}_P[\xi V^\pi]. \tag{18}$$

This has the same form as the Bellman equation for distributional robust RL (Iyengar, 2005; Nilim & El Ghaoui, 2004), and since $\mathbb{E}_P[\xi] = 1 : P\xi \in \Delta(\mathcal{S})$ is indeed a transition kernel. The uncertainty set of the transition kernel can be defined as follows:

$$\mathcal{U}^\tau(P_{s,a}) = \left\{ \bar{P}_{s,a} \in \Delta(\mathcal{S}), \ 0 \le \frac{\bar{P}_{s,a}(s')}{P_{s,a}(s')} \le \frac{1}{\tau} \right\}, \tag{19}$$

and the uncertainty set satisfied the (s, a)-rectangularity (Iyengar, 2005). We can define the robust Bellman operator for CVaR as:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} :$$
$$\mathcal{T}^\tau(Q)(s, a) := r(s, a) + \gamma \inf_{\bar{P} \in \mathcal{U}^\tau(P_{s,a})} \bar{P}V. \quad (20)$$

Generative Model. Assume we have access to a generative model or a simulator, which can provide samples $s' \sim P(\cdot|s, a)$, for any (s, a). Suppose we call our generative model N times for each state-action pair. Let $\widehat{P}$ be the empirical model defined as follows

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}, \ s_{i,s,a} \overset{\text{i.i.d}}{\sim} P(\cdot|s, a), i = 1, 2, \cdots, N.$$

The total sample size is then NSA.

Goal. Given the collected samples, the goal is to learn the risk-sensitive optimal policy under risk level τ using as few samples as possible. Specifically, given a target accuracy tolerance $\epsilon > 0$, the goal is to find an ϵ-optimal risk-sensitive policy $\widehat{\pi}$ s.t.

$$\forall s \in \mathcal{S} : \quad V^*(s) - V^{\widehat{\pi}}(s) \le \epsilon. \quad (21)$$

4 Algorithm

In this section, we present a model-based approach, which first constructs an empirical nominal transition kernel based on the collected samples and then applies a value iteration-based algorithm ICVaR-VI to compute an optimal risk-sensitive policy in the approximated MDP.

Empirical Nominal Transition Kernel. The empirical nominal transition kernel $\widehat{P}$ can be constructed as follows: $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$

$$\widehat{P}(s'|s, a) := \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{s_{i,s,a} = s'\}. \quad (22)$$

We define $\widehat{\mathcal{M}}$ to be the empirical MDP that is identical to the original M, except that it uses $\widehat{P}$ instead of P for the transition kernel. We use $\widehat{V}^\pi$ and $\widehat{Q}^\pi$ to denote the value and action value functions of a policy π in $\widehat{\mathcal{M}}$. And $\widehat{\pi}^*$, $\widehat{Q}^*$ and $\widehat{V}^*$ are the optimal policy and value functions in $\widehat{\mathcal{M}}$.

Equipped with $\widehat{P_0}$, we can define the empirical Bellman operator $\mathcal{T}^\tau$ as follows: $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\widehat{\mathcal{T}}^\tau(Q)(s, a) := r(s, a) + \gamma \text{CVaR}_{s' \sim P(\cdot|s,a)}^\tau(V(s')), \quad (23)$$

where $V(s) = \max_a Q(s, a)$.

ICVaR-VI: Iterated CVaR Value Iteration. To find the fixed point of $\mathcal{T}^\tau$, we introduce iterated CVaR

value iteration (ICVAR-VI)(Ruszczyński, 2010), which is shown in Algorithm 1. The update rule can be written as:

$$\widehat{Q}_t(s, a) = \widehat{\mathcal{T}}^\tau(\widehat{Q}_{t-1})(s, a)$$
$$= r(s, a) + \gamma \text{CVaR}_\tau(\widehat{V}_{t-1}(s')), \quad (24)$$

where $\widehat{V}_{t-1} = \max_a \widehat{Q}_{t-1}(s, a)$ for all $s \in \mathcal{S}$.

Algorithm 1 ICVaR-VI

Input: Empirical nominal transition kernel $\widehat{P}$; reward function r; risk level τ; number of iterations T.
1: **Initialization:** $\widehat{Q}_0(s, a) = 0$, $\widehat{V}_0(s) = 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3:     **for** $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
4:         Update $\widehat{Q}_t(s, a)$ according to (24);
5:     **end for**
6:     **for** $s \in \mathcal{S}$ **do**
7:         Set $\widehat{V}_t(s) = \max_a \widehat{Q}_t(s, a)$;
8:     **end for**
9: **end for**
10: **Output:** $\widehat{Q}_T$, $\widehat{V}_T$, and policy $\widehat{\pi}(s) = \arg\max_a \widehat{Q}_T(s, a)$.

---

5 Theoretical Results

In this section, we present our main theoretical results. We start with the upper bound on the sample complexity for Iterated CVaR RL.

Theorem 1 (Sample Complexity Upper Bound). *For any risk level $\tau \in (0, 1]$, the number of samples needed by Algorithm ICVaR-VI to return an ϵ-optimal policy with probability at least $1 - \delta$ is at most $\tilde{\mathcal{O}}\left(\frac{SA}{\tau^2(1-\gamma)^4\epsilon^2}\right)$. In addition, when $\tau \ge \gamma$, the sample complexity can be further improved to $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$.*

Remark 1. *In Theorem 1, the dependencies on S and A match with the result of Risk-neutral RL (Gheshlaghi Azar et al., 2013). Furthermore, our sample complexity matches the dependency on τ and $1 - \gamma$ with the result in Du et al. (2022)[1], and our results improve the dependency on S by a factor of S (Du et al., 2022).*

The sample complexity upper bound holds for any risk level $\tau \in (0, 1]$. In the special case when $\tau = 1$, CVaR

---

[1]In the finite-horizon episodic setting, the horizon H is analogous to $(1 - \gamma)^{-1}$ in the infinite-horizon setting. Du et al. (2022) focuses on the stationary finite-horizon episodic setting, where the transition kernel at each time step is the same, and therefore we incorporate an extra H into the sample complexity for a fair comparison with our results.

reduces to expectation, and Iterated CVaR RL reduces to risk-neutral RL. In this case, our sample complexity bound is $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$, which matches the result of the state-of-the-art sample complexity bound for standard risk-neutral RL (Agarwal et al., 2020; Gheshlaghi Azar et al., 2013). Furthermore, for an arbitrary constant risk level $0 < \tau < \gamma$, the sample complexity is increased by a factor of $(1-\gamma)^{-1}$, which, later in our lower bound analysis, is proved to be necessary.

In standard risk-neutral RL, the sample complexity of $(1-\gamma)^{-3}$ can be obtained using Bernstein inequality in combination with the Bellman property of a policy's variance (Lemma 4 of Agarwal et al. (2020)), which could improve the sample complexity by a factor of $(1-\gamma)^{-1}$ comparing to only using Hoeffding's inequality. However, using Bernstein inequality may not always improve the sample complexity for Iterated CVaR RL by a factor of $(1-\gamma)^{-1}$. In distributional robust RL (which is similar to our case since the dual form of Iterated CVaR can also be written as a distributionally robust optimization form), there is an extra term because the worst-case transition kernel is different from the nominal one, and whether Bernstein is superior to Hoeffding's inequality depends on the specific uncertainty set (Shi et al., 2023; Shi & Chi, 2024; Panaganti & Kalathil, 2022). In our setting, when $\tau < \gamma$, applying Bernstein's inequality to the uncertainty set for CVaR leads to the same sample complexity bound as Hoeffding's inequality. This phenomenon also appears in existing sample complexity analyses for χ^2 distributionally robust RL (Shi & Chi, 2024). Still, for total-variation distance defined robust RL, Bernstein's inequality can improve the upper bound when $\tau \geq \gamma$, the total-variation distance of our uncertainty set is upper bounded by $\frac{1-\tau}{\tau}$, and Bernstein's inequality can reduce the sample complexity by a factor of $(1-\gamma)^{-1}$.

The analysis of Iterated CVaR RL presents the following challenge: The objective is not the expected cumulative reward but rather the Iterated CVaR risk measure, which prevents us from decomposing the error as we do in risk-neutral cases. To address this challenge, we will establish a connection between Iterated CVaR RL and distributionally robust RL. This connection enables us to treat the risk-sensitive objective as the expected cumulative reward under the worst-case transition kernel. We then quantify the deviation between the empirical model and the true underlying model, using Hoeffding's inequality to derive bounds for $\tau < \gamma$. When $\tau \geq \gamma$, we introduce an alternative analytical approach using Bernstein inequality to further tighten the sample complexity bounds.

Below we present a proof sketch for Theorem 1 to highlight our major technical contributions (also see Appendix B for a complete proof).

Proof sketch of Theorem 1. With the connection between Iterated CVaR and distributionally robust RL, we can decompose the error in the following way. Let $\hat{\pi}^*$ denote the optimal risk-sensitive policy in the empirical model $\widehat{\mathcal{M}}$, and let $\hat{\pi}$ represent the policy from Algorithm 1. Additionally, $\widehat{V}$ represents the Iterated CVaR value function in the empirical model.

The sub-optimality gap between π^* and $\hat{\pi}$ can be decomposed as

$$V^* - V^{\hat{\pi}} \overset{(i)}{\leq} \left(V^{\pi^*} - \widehat{V}^{\pi^*}\right) + \frac{2\gamma\epsilon_{opt}}{1-\gamma}\mathbf{1} + \left(\widehat{V}^{\hat{\pi}} - V^{\hat{\pi}}\right), \tag{25}$$

where (i) holds by the optimality of $\hat{\pi}$ in $\widehat{\mathcal{M}}$ and γ-contraction property of risk-sensitive Bellman operator (Lemma 1). $\mathbf{1} \in \mathbb{R}^S$ is the all 1 vector.

To bound $||\widehat{V}^{\pi^*} - V^{\pi^*}||_\infty$ and $||\widehat{V}^{\hat{\pi}} - V^{\hat{\pi}}||_\infty$, we introduce a key inequality:

$$\left|\text{CVaR}^\tau_{s'\sim\widehat{P}(\cdot|s,a)}(V^\pi(s')) - \text{CVaR}^\tau_{s'\sim P(\cdot|s,a)}(V^\pi(s'))\right|$$
$$\leq \frac{1}{\tau}\sup_{t\in\mathbb{R}}\left|\sum_{s'}\left(\widehat{P}_{s,a}(s') - P_{s,a}(s')\right)\right|. \tag{26}$$

Applying the concentration lemma for CVaR (Lemma 3), we get that

$$||\widehat{V}^\pi - V^\pi||_\infty \leq c_0\sqrt{\frac{L}{N\tau^2(1-\gamma)^4}}. \tag{27}$$

L stands for the log term of S, A, N and $\frac{1}{\delta}$. c_0 is a large enough constant. Finally, for a small enough ϵ_{opt}, we obtain the sample complexity upper bound.

For the special case when $\tau \geq \gamma$, we first introduce the total-variation distance bound for the CVaR uncertainty set $\mathcal{U}^\tau$:

$$\forall \bar{P}_{s,a} \in \mathcal{U}^\tau(P_{s,a}) : \|\bar{P}_{s,a} - P_{s,a}\|_1 \leq 2\frac{1-\tau}{\tau}. \tag{28}$$

This bound is useful when $\tau \geq \frac{1}{2}$. When the $\tau \geq \gamma$, the dependence on $(1-\gamma)^{-1}$ of the extra term induced by worst-case transition mismatch ($\mathcal{C}_2$ in Shi et al. (2023)) is reduced when applying Berstein inequality:

$$\left|\text{Var}_{\bar{P}_{s,a}}(V) - \text{Var}_{P_{s,a}}(V)\right| \leq \|\bar{P}_{s,a} - P_{s,a}\|_1\|V\|_\infty^2$$
$$\leq \frac{1-\tau}{\tau}\frac{1}{(1-\gamma)^2} \leq \frac{1}{\gamma(1-\gamma)}, \tag{29}$$

where $\text{Var}_{P_{s,a}}(V)$ is the variance of V respect to distribution $P_{s,a}$. In this case, we can reduce the order

of $(1-\gamma)^{-1}$, which leads to the following bound:

$$\|\widehat{V}^\pi - V^\pi\|_\infty \le c_1 \sqrt{\frac{L}{N(1-\gamma)^3}}, \qquad (30)$$

where c_1 is some universal constant.

The analysis of the above two cases concludes the proof of the sample complexity upper bound to get an ϵ-optimal policy. $\qquad\square$

Additionally, in order to assess the tightness of Theorem 1, we further develop a minimax lower bound as follows, with the proof provided in Appendix C.2.

Theorem 2 (Sample Complexity Lower Bound). *Fix any $\tau \in (0,1]$, $\gamma \in (\frac{1}{2}, 1)$, there exist an MDP s.t. for any algorithm to obtain an ϵ-optimal policy, the sample complexity is at least $\tilde{\mathcal{O}}\left(\frac{(1-\gamma\tau)SA}{\tau(1-\gamma)^4\epsilon^2}\right)$. In addition, when $\tau \ge \gamma$, the sample complexity of any algorithm is at least $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$.*

Remark 2. *Theorem 2 shows that when τ is small, there exists an MDP for which the sample complexity becomes unavoidably large. When the risk tolerance is low—indicating that the agent is highly sensitive to adverse states—more samples are needed for each state-action pair to gather more accurate information about the environment, allowing the agent to develop a safer policy.*

Lower Bound Analysis. In this following, we outline the proof idea for the lower bound in Theorem 2, with the full proof deferred to Appendix C.2. Our proof is inspired by the lower bound construction in distributional robust RL (Shi et al., 2023). We first construct two similar MDPs with close transition kernels that are hard to distinguish. For each MDP, there is an unknown optimal action. If an algorithm is capable of achieving an ϵ-optimal policy in the Iterated CVaR RL problem, it must also be able to identify the optimal action and determine which MDP it is interacting with, with high probability. The challenge then becomes determining how many samples are needed to distinguish between two distributions. One notable difference in our problem is that, in our case, the two transition probabilities to the rewarding state are close to $1-\tau$, rather than $1-\gamma$ as in standard and distributional robust RL. Since CVaR computes the average over the worst τ-quantile of the reward-to-go, the transition probability to lower reward states must be smaller than τ for the worst-case transition kernel to differ. As a result, the final lower bound has a higher order in $1-\gamma$ compared with risk-neutral settings when τ is relatively small.

Recall that the sufficient and necessary sample complexity for learning a standard risk-neutral MDP is

$\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$ (Gheshlaghi Azar et al., 2013; Agarwal et al., 2020). Intuitively, the sample complexity for CVaR should include an additional $\frac{1}{\tau}$ factor compared to the risk-neutral case, since CVaR only considers the worst τ-portion of outcomes and takes an average. Therefore, the number of samples needed should be $\tilde{\mathcal{O}}\left(\frac{SA}{\tau(1-\gamma)^3\epsilon^2}\right)$. This is true for static CVaR RL where Wang et al. (2023) provided a regret lower bound has an extra $\sqrt{\tau^{-1}}$ term comparing to risk-neutral regret lower bound ($\sqrt{\tau^{-1}}$ in regret is equivalent to τ^{-1} in PAC condition). But for Iterated CVaR RL, it is not merely a matter of averaging the worst τ-protion of trajectories. Du et al. (2022) in Section C.2 provides a detailed discussion of the differences between static and iterated CVaR in the episodic setting.

When $\tau \ge \gamma$, our sample complexity lower bound becomes $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$ which matches the minimax optimal sample complexity for the risk-neutral case, in general, Iterated CVaR RL is harder to learn than standard RL. However, when τ is large, then the problem becomes closer to a risk-neutral one.

Nearly Tight Sample Complexity. By combining the upper bound from Theorem 1 with the minimax lower bound from Theorem 2, we confirm that the sample complexity is nearly optimal:

- When $\tau \in (0,1)$ is a constant independent of γ, our sample complexity upper bound $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^4\epsilon^2}\right)$ is tight and matches the minimax lower bound;

- When $\tau \ge \gamma$, our sample complexity upper bound is $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$, and it matches with the minimax lower bound;

- When $\tau \le 1 - \gamma$, our sample complexity $\tilde{\mathcal{O}}\left(\frac{SA}{\tau^2(1-\gamma)^4\epsilon^2}\right)$ has a gap of $\frac{1}{\tau} \ge \frac{1}{1-\gamma}$ compared to the minimax lower bound. This case of small risk level τ will be further discussed in the next section.

6 Worst-Path RL

In this section, we investigate the problem with a fixed MDP and consider a limiting case where the risk level is small. This problem is referred to as worst-path RL (Du et al., 2022).

Specifically, consider an MDP, and denote by $p_{\min}$ the minimum non-zero reaching probability from any state-action pair: $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$, and $\forall s' \in \text{supp}(P(\cdot|s,a))$, $P(s'|s,a) \ge p_{\min}$. Consider small risk level τ: $\tau \le p_{\min}$. Here we use $\text{supp}(P)$ to denote the support of a distribution P.

This case is not covered by results in Theorems 1 and 2. Obviously, the sample complexity in Theorems 1 and 2 depends on $\frac{1}{\tau}$ and goes to infinity as $\tau \to 0$. However, as will be shown later, in the case with $\tau \leq p_{\min}$, a sample complexity of $\tilde{\mathcal{O}}(\frac{SA}{p_{\min}})$ is minimax optimal, and it does not depend on ϵ and $1 - \gamma$.

Taking the minimax lower bound result in Theorem 2 as an example to explain the difference. To prove the minimax lower bound in Theorem 2, for any risk level τ, we construct two MDPs with transition probabilities to lower-reward states smaller than the risk level τ. However, such hard examples do not satisfy the problem setting in this section: $\tau \leq p_{\min}$.

As we have mentioned earlier, as $\tau \to 0$, $\mathrm{CVaR}_\tau(Z) \to \mathrm{ess\,inf}\, Z$. In Du et al. (2022), regret bounds independent of the number of episodes were developed, where the bounds also depend on visitation probability to the worst state. Given access to a generative model, the problem is simpler since there is no need to explore. Then the transition to the worst states is simply the frequency of that state in the N samples generated. When τ is smaller than all possible non-zero transition probability, CVaR_τ simply reduces to finding the worst state for the reward-to-go.

Bellman operator and Bellman equations. Since $\tau \leq p_{\min}$, the objective reduces to maximizing the accumulative reward along the worst-case trajectory. The Bellman equations can then be written as follows:

$$
Q^\pi(s,a) = r(s,a) + \gamma \min_{s' \in \mathrm{supp}(P(\cdot|s,a))}(V^\pi(s')),
$$
$$
V^\pi(s) = \max_{a \in \mathcal{A}} Q^\pi(s,a), \tag{31}
$$

where $\min_{s' \in \mathrm{supp}(P(\cdot|s,a))}$ considers the worst-case of all possible next states (with non-zero probability). The algorithm for this problem is similar to Algorithm 1 with a slightly different Bellman operator:

$$
\forall (s,a) \in \mathcal{S} \times \mathcal{A} :
$$
$$
\widehat{\mathcal{T}}^\tau(Q)(s,a) := r(s,a) + \gamma \min_{s':n(s',s,a)>0}(V(s')),
$$
$$
\text{with } V(s) := \max_a Q(s,a) \tag{32}
$$

where $n(s',s,a)$ denotes the number of samples with next state s' in the total N generated samples for state action pair (s,a). With that in mind, we substitute (32) to (23) in Algorithm 1, and we have the algorithm for the worst-path RL problem in this section.

Sample Complexity Analysis. Below we provide the sample complexity upper bound for the algorithm discussed above for the worst-path RL problem.

Theorem 3 (Worst-Path RL Upper Bound). *Consider a risk level $\tau \leq p_{\min}$. With probability at least*

$1 - \delta$, *the number of samples needed to obtain an optimal policy is at most*

$$
\mathcal{O}\left(\frac{SA}{p_{\min}}\left(1 + \log\left(\frac{SA}{\delta}\right)\right)\right). \tag{33}
$$

Remark 3. *For any $\tau < p_{\min}$, CVaR risk measure reduces to the essential infimum, and therefore, the bound does not depend on τ anymore. More importantly, the upper bound now depends on $\frac{1}{p_{\min}}$, which is strictly smaller than $\frac{1}{\tau^2}$.*

The key idea in the proof is to analyze the suboptimality gap using the occurrence of the worst-case state. If the worst-case state occurs with high probability, the Bellman operator behaves as it would under the true underlying model. Otherwise, the suboptimality gap becomes non-vanishing. This also explains why the sample complexity does not depend on ϵ and $\frac{1}{1-\gamma}$ here. This result aligns with our intuition: if we expect a state with probability $p_{\min}$ to occur, we need $\frac{1}{p_{\min}}$ samples for each state-action pair.

To validate the optimality of our sample complexity upper bound, we also provide a minimax lower bound.

Theorem 4 (Worst-Path RL Lower Bound). *For a given risk level τ, there exists an MDP with $P_{\min} \geq \tau$ such that for any algorithm to obtain an optimal policy at a risk level $\tau \leq p_{\min}$, the sample complexity is at least $\tilde{\mathcal{O}}\left(\frac{SA}{p_{\min}}\right)$.*

The minimax lower bound matches with the upper bound in Theorem 3.

7 Acknowledgements

8 Conclusion

In this paper, we investigate Iterated CVaR RL problem in infinite horizon discounted MDP with access to a generative model. We introduce the algorithm ICVaR-VI and provide nearly matching sample complexity upper and lower bounds. Later we study the limit case with an arbitrarily small risk level τ, and provide tight upper and lower bounds. There are several interesting directions for future work, e.g., further closing the gap between the upper and lower sample complexity bounds and extending iterated CVaR to other types of coherent risk measure or general Markov coherent risk.

References

Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Proc. Annual Conference on Learning Theory (CoLT)*, volume 125, pp. 67–83, 2020.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Osbert Bastani, Jason Yecheng Ma, Estelle Shen, and Wanqiao Xu. Regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 35:36259–36269, 2022.

Nicole Bäuerle and Alexander Glauner. Markov decision processes with recursive risk measures. *European Journal of Operational Research*, 296(3):953–966, 2022.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with Average-Value-at-Risk criteria. *Mathematical Methods of Operations Research*, 74:361–379, 2011.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 2024.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 1691–1692. PMLR, 2018.

Gaëlle Chagny. An introduction to nonparametric adaptive estimation. *The Graduate Journal of Mathematics*, 2016(2):105–120, 2016.

Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance. *Computational Economics*, pp. 1–38, 2021.

Yu Chen, Yihan Du, Pihe Hu, Siwei Wang, Desheng Wu, and Longbo Huang. Provably efficient iterated CVaR reinforcement learning with function approximation. *arXiv preprint arXiv:2307.02842*, 2023.

Shanyun Chu and Yi Zhang. Markov decision processes with iterated coherent risk measures. *International Journal of Control*, 87(11):2286–2293, 2014.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Yihan Du, Siwei Wang, and Longbo Huang. Provably efficient risk-sensitive reinforcement learning: Iterated CVaR and worst path. *arXiv preprint arXiv:2206.02678*, 2022.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91:325–349, 2013.

Mary R Hardy and Julia L Wirch. The iterated CTE: a dynamic risk measure. *North American Actuarial Journal*, 8(4):62–75, 2004.

Audrey Huang, Liu Leqi, Zachary C Lipton, and Kamyar Azizzadenesheli. On the convergence and optimality of policy gradient for markov coherent risk. *arXiv preprint arXiv:2103.02827*, 2021.

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

Michael Kearns and Satinder Singh. Finite-Sample Convergence Rates for Q-Learning and Indirect Algorithms. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 11, 1998.

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

Xinyi Ni, Guanlin Liu, and Lifeng Lai. Risk-sensitive reward-free reinforcement learning with CVaR. In *Proc. International Conference on Machine Learning (ICML)*, volume 235, pp. 37999–38017, 2024.

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.

Takayuki Osogami. Iterated risk measures for risk-sensitive markov decision processes with discounted cost. *arXiv preprint arXiv:1202.3755*, 2012.

Long Ouyang et al. Training language models to follow instructions with human feedback. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022.

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

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Andrzej Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125:235–261, 2010.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory.* SIAM, 2021.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

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. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

R. Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 2803–2830, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* The MIT Press, Cambridge, Massachusetts, 2018.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28, 2015.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with CVaR. In *Proc. International Conference on Machine Learning (ICML)*, pp. 35864–35907, 2023.

Wenhao Xu, Xuefeng Gao, and Xuedong He. Regret bounds for Markov decision processes with recursive Optimized Certainty Equivalents. In *Proc. International Conference on Machine Learning (ICML)*, pp. 38400–38427, 2023.

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

Yulai Zhao, Wenhao Zhan, Xiaoyan Hu, Ho-fung Leung, Farzan Farnia, Wen Sun, and Jason D Lee. Provably efficient CVaR RL in low-rank MDPs. *arXiv preprint arXiv:2311.11965*, 2023.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8665–8675, 2019.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A   Notations and useful lemmas

In the appendices, with a slight abuse of notations, we use $P \in \mathbb{R}^{SA \times S}$ to denote the transition matrix of the nominal transition kernel P, and let $P_{s,a}$ denote its (s,a)-th row. Similarly, we could define the transition matrix $\widehat{P}$ for the empirical nominal transition kernel $\widehat{P}$. We further define the following matrix/vector notations for the convenience of presentation. Let the state space $\mathcal{S} = \{0, 1, 2, \ldots, S-1\}$ and action space $\mathcal{A} = \{0, 1, 2, \ldots, A-1\}$. A *deterministic* policy π is a mapping from the state space to the action space, i.e., $\pi(s)$ is an action in $\mathcal{A}$.

- $r \in \mathbb{R}^{SA}$: vector form of the reward function r.

- $\Pi^\pi \in \{0, 1\}^{S \times SA}$: projection matrix associated with a deterministic policy π:

$$
\Pi^\pi = \begin{pmatrix} e_{\pi(0)}^\top & \mathbf{0}^\top & \cdots & \mathbf{0}^\top \\ \mathbf{0}^\top & e_{\pi(1)}^\top & \cdots & \mathbf{0}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^\top & \mathbf{0}^\top & \cdots & e_{\pi(S-1)}^\top \end{pmatrix},
$$

  where $e_{\pi(0)}^\top, e_{\pi(1)}^\top, \ldots, e_{\pi(S-1)}^\top \in \mathbb{R}^A$ are standard basis vectors and $\mathbf{0} \in \mathbb{R}^S$ is all zero vector.

- $r_\pi \in \mathbb{R}^S$: reward vector restricted to the actions chosen by the deterministic policy π, namely, $r_\pi(s) = r(s, \pi(s))$ for all $s \in S$ (or simply, $r_\pi = \Pi^\pi r$).

- $P^V \in \mathbb{R}^{SA \times S}$, $\widehat{P}^V \in \mathbb{R}^{SA \times S}$: worst-case transition matrices for a vector $V \in \mathbb{R}^S$. We denote $P_{s,a}^V$ (resp. $\widehat{P}_{s,a}^V$) as the (s,a)-th row. Specifically,

$$
P_{s,a}^V = \underset{\bar{P}_{s,a} \in \mathcal{U}^\tau(P_{s,a})}{\arg\min} \bar{P}_{s,a}V, \quad \widehat{P}_{s,a}^V = \underset{\bar{P}_{s,a} \in \mathcal{U}^\tau(\widehat{P}_{s,a})}{\arg\min} \bar{P}_{s,a}V.
$$

  Furthermore, we make use of the following short-hand notation:

$$
P_{s,a}^{\pi,V} := \underset{\bar{P}_{s,a} \in \mathcal{U}^\tau(P_{s,a})}{\arg\min} \bar{P}_{s,a}V^\pi, \quad P_{s,a}^{\pi,\widehat{V}} := \underset{\bar{P}_{s,a} \in \mathcal{U}^\tau(P_{s,a})}{\arg\min} \bar{P}_{s,a}\widehat{V}^\pi,
$$

$$
\widehat{P}_{s,a}^{\pi,V} := \underset{\bar{P}_{s,a} \in \mathcal{U}^\tau(\widehat{P}_{s,a})}{\arg\min} \bar{P}_{s,a}V^\pi, \quad \widehat{P}_{s,a}^{\pi,\widehat{V}} := \underset{\bar{P}_{s,a} \in \mathcal{U}^\tau(\widehat{P}_{s,a})}{\arg\min} \bar{P}_{s,a}\widehat{V}^\pi.
$$

  The corresponding probability transition matrices are denoted by $P^{\pi,V} \in \mathbb{R}^{SA \times S}$, $P^{\pi,\widehat{V}} \in \mathbb{R}^{SA \times S}$, $\widehat{P}^{\pi,V} \in \mathbb{R}^{SA \times S}$ and $\widehat{P}^{\pi,\widehat{V}} \in \mathbb{R}^{SA \times S}$, respectively.

- $P^\pi \in \mathbb{R}^{S \times S}$, $\widehat{P}^\pi \in \mathbb{R}^{S \times S}$, $\underline{P}^{\pi,V} \in \mathbb{R}^{S \times S}$, $\underline{P}^{\pi,\widehat{V}} \in \mathbb{R}^{S \times S}$, $\underline{\widehat{P}}^{\pi,V} \in \mathbb{R}^{S \times S}$ and $\underline{\widehat{P}}^{\pi,\widehat{V}} \in \mathbb{R}^{S \times S}$: probability transition matrices w.r.t. policy π over the states:

$$
P^\pi := \Pi^\pi P, \quad \widehat{P}^\pi := \Pi^\pi \widehat{P}, \quad \underline{P}^{\pi,V} := \Pi^\pi P^{\pi,V}, \quad \underline{P}^{\pi,\widehat{V}} := \Pi^\pi P^{\pi,\widehat{V}},
$$

$$
\underline{\widehat{P}}^{\pi,V} := \Pi^\pi \widehat{P}^{\pi,V}, \quad \underline{\widehat{P}}^{\pi,\widehat{V}} := \Pi^\pi \widehat{P}^{\pi,\widehat{V}}. \tag{34}
$$

- $\mathrm{Var}_{\bar{P}}(V) \in \mathbb{R}^{SA}$: for any transition kernel $\bar{P} \in \mathbb{R}^{SA \times S}$ and any vector $V \in \mathbb{R}^S$, the (s,a)-th row of $\mathrm{Var}_P(V)$ is

$$
\mathrm{Var}_P(s,a) = \mathrm{Var}_{P_{s,a}}(V), \tag{35}
$$

  where $\mathrm{Var}_{P_{s,a}}(V) := P_{s,a}(V^2) - (P_{s,a}V)^2$.

Lemma 1. *(Ruszczyński, 2010, Lemma 2). For any $\gamma \in [0, 1)$, the robust Bellman operator $\mathcal{T}^\tau(\cdot)$ is a γ-contraction w.r.t. $\|\cdot\|_\infty$. Namely, for any $Q_1, Q_2 \in \mathbb{R}^{SA}$ s.t. $Q_1(s,a), Q_2(s,a) \in [0, \frac{1}{1-\gamma}]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, one has*

$$
\|\mathcal{T}^\tau(Q_1) - \mathcal{T}^\tau(Q_2)\|_\infty \leq \gamma\|Q_1 - Q_2\|_\infty. \tag{36}
$$

Lemma 2 (Shi et al., 2023, Lemma 5). *Let $\widehat{Q}_0 = \mathbf{0}$. The iterates $\{\widehat{Q}_t\}$, $\{\widehat{V}_t\}$ of ICVaR-VI (Algorithm 1) obey*

$$\forall t \geq 0: \quad \|\widehat{Q}_t - \widehat{Q}^*\|_\infty \leq \frac{\gamma^t}{1-\gamma} \quad \|\widehat{Q}_t - \widehat{Q}^*\|_\infty \leq \frac{\gamma^t}{1-\gamma}. \tag{37}$$

Furthermore, the output policy $\widehat{\pi}$ obeys

$$\|\widehat{V}^* - \widehat{V}^{\widehat{\pi}}\|_\infty \leq \frac{2\gamma\epsilon_{opt}}{1-\gamma}, \quad where \quad \epsilon_{opt} := \|\widehat{V}^* - \widehat{V}_{T-1}\|_\infty. \tag{38}$$

B   Proof of Theorem 1: Sample Complexity Upper Bound

Step 1: Decomposing the error. The optimality gap can be decomposed as

$$V^* - V^{\widehat{\pi}} = \left(V^{\pi^*} - \widehat{V}^{\pi^*}\right) + \left(\widehat{V}^{\pi^*} - \widehat{V}^{\widehat{\pi}^*}\right) + \left(\widehat{V}^{\widehat{\pi}^*} - \widehat{V}^{\widehat{\pi}}\right) + \left(\widehat{V}^{\widehat{\pi}} - V^{\widehat{\pi}}\right)$$

$$\overset{(i)}{\leq} \left(V^{\pi^*} - \widehat{V}^{\pi^*}\right) + \left(\widehat{V}^{\widehat{\pi}^*} - \widehat{V}^{\widehat{\pi}}\right) + \left(\widehat{V}^{\widehat{\pi}} - V^{\widehat{\pi}}\right)$$

$$\overset{(ii)}{\leq} \left(V^{\pi^*} - \widehat{V}^{\pi^*}\right) + \frac{2\gamma\epsilon_{opt}}{1-\gamma}\mathbf{1} + \left(\widehat{V}^{\widehat{\pi}} - V^{\widehat{\pi}}\right), \tag{39}$$

where (i) holds by the fact that $\widehat{\pi}^*$ is the optimal policy under transition kernel $\widehat{P}$; and (ii) follows from Lemma 2.

The first and third terms in the sub-optimality gap in (39) can be bounded in the same way as follows

where (i) holds by the fact that corresponds to the worst-case transition kernel in $\mathcal{U}^\tau(P^\pi)$ for V^π. By decomposing the error in a symmetric way, we can similarly obtain that

$$\widehat{V}^\pi - V^\pi \leq \gamma(I - \gamma\underline{\widehat{P}}^{\pi,V})^{-1}(\underline{\widehat{P}}^{\pi,V}V^\pi - \underline{P}^{\pi,V}V^\pi). \tag{40}$$

Combining (**??**) and (40), we arrive at

Step 2: Controlling $\|\widehat{V}^\pi - V^\pi\|_\infty$ and $\|\widehat{V}^{\widehat{\pi}} - V^{\widehat{\pi}}\|_\infty$ in (39).

Lemma 3. *For any $\delta \in (0,1)$, with probability at least $1 - \delta$, one has that*

$$\left\|\underline{\widehat{P}}^{\pi,V}V^\pi - \underline{P}^{\pi,V}V^\pi\right\|_\infty \leq \frac{2}{\tau}\sqrt{\frac{2\log(6SAN/\delta)}{N(1-\gamma)^2}}. \tag{41}$$

Proof. Note that $V^\pi(s) \in \left[0, \frac{1}{1-\gamma}\right]$ for any $s \in \mathcal{S}$. Therefore, $\sup_{x \in \mathbb{R}}$ is equivalent to $\sup_{x \in [0, \frac{1}{1-\gamma}]}$. We first show that

$$|(\widehat{P}_{s,a}^{\pi,V} - P_{s,a}^{\pi,V})V^\pi| = \left|\sup_{x \in [0, \frac{1}{1-\gamma}]}\left\{x - \frac{1}{\tau}\mathbb{E}_{s' \sim \widehat{P}_{s,a}}[(x - V^\pi(s'))_+]\right\} - \sup_{x \in [0, \frac{1}{1-\gamma}]}\left\{x - \frac{1}{\tau}\mathbb{E}_{s' \sim P_{s,a}}[(x - V^\pi(s'))_+]\right\}\right|$$

$$\leq \frac{1}{\tau}\sup_{x \in [0, \frac{1}{1-\gamma}]}\left|\mathbb{E}_{s' \sim \widehat{P}_{s,a}}[(x - V^\pi(s'))_+] - \mathbb{E}_{s' \sim P_{s,a}}[(x - V^\pi(s'))_+]\right|$$

$$= \frac{1}{\tau}\sup_{x \in [0, \frac{1}{1-\gamma}]}\left|\mathbb{E}_{s' \sim \widehat{P}_{s,a}}[x-] - \mathbb{E}_{s' \sim P_{s,a}}[x-]\right|$$

$$= \frac{1}{\tau}\sup_{x \in [0, \frac{1}{1-\gamma}]}\left|\sum_{s'}\left(\widehat{P}_{s,a}(s') - P_{s,a}(s')\right)\right|. \tag{42}$$

We define the function g as the difference in the expectation of V between the two transition kernel for a fix x and (s, a),

$$g_{s,a}(x, V) := \left|\sum_{s'}\left(\widehat{P}_{s,a}(s') - P_{s,a}(s')\right)\right|. \tag{43}$$

Using Hoeffding's inequality, one has that with probability at least $1 - \delta$,

$$g_{s,a}(x, V) \leq \sqrt{\frac{2 \log(2/\delta)}{N(1-\gamma)^2}}. \tag{44}$$

It can be easily shown that $g_{s,a}(s, V)$ is 1-Lipschitz in x for any V such that $\|V\|_\infty \leq \frac{1}{1-\gamma}$. To obtain the union bound, we construct an ϵ_1-net $\mathcal{N}_{\epsilon_1}$ over $[0, \frac{1}{1-\gamma}]$, with size $|\mathcal{N}_{\epsilon_1}| \leq \frac{3}{(1-\gamma)\epsilon_1}$ (Vershynin, 2018). By the union bound, we have that with probability at least $1 - \delta$,

$$
\begin{aligned}
\sup_{x \in \mathbb{R}} \left| \sum_{s'} \left(\widehat{P}_{s,a}(s') - P_{s,a}(s') \right) \right| &\overset{(i)}{\leq} \epsilon_1 + \sup_{x \in \mathcal{N}_{\epsilon_1}} \left| \sum_{s'} \left(\widehat{P}_{s,a}(s') - P_{s,a}(s') \right) \right| \\
&\overset{(ii)}{\leq} \sqrt{\frac{2 \log(2SA|\mathcal{N}_{\epsilon_1}|/\delta)}{N(1-\gamma)^2}} + \epsilon_1 \\
&\overset{(iii)}{\leq} 2\sqrt{\frac{2 \log(6SAN/\delta)}{N(1-\gamma)^2}},
\end{aligned} \tag{45}
$$

where (i) follows from that the optimal x falls into an ϵ_1-ball centered around some point in $\mathcal{N}_{\epsilon_1}$ and $g_{s,a}(x, V)$ is 1-Lipschitz in x; (ii) stems from applying the results in (44) and the union bound over $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{N}_{\epsilon_1}$; and (iii) follows if we let $\epsilon_1 = \sqrt{\frac{2 \log(6SAN/\delta)}{N(1-\gamma)^2}}$ and then $|\mathcal{N}_{\epsilon_1}| \leq \frac{3}{\epsilon_1(1-\gamma)} \leq 3N$. Substituting (45) back into (42), we have that with probability at least $1 - \delta$,

$$\left| (\widehat{P}^{\pi,V}_{s,a} - P^{\pi,V}_{s,a}) V^\pi \right| \leq \frac{2}{\tau} \sqrt{\frac{2 \log(6SAN/\delta)}{N(1-\gamma)^2}}. \tag{46}$$

From (46), it can be shown that with probability at least $1 - \delta$,

$$\left\| \underline{\widehat{P}}^{\pi,V} V^\pi - \underline{P}^{\pi,V} V^\pi \right\|_\infty \leq \frac{2}{\tau} \sqrt{\frac{2 \log(6SAN/\delta)}{N(1-\gamma)^2}}. \tag{47}$$

This concludes the proof of Lemma 3. $\qquad\square$

Substituting (47) back into (**??**) we get that with probability at least $1 - \delta$,

$$
\begin{aligned}
\|\widehat{V}^\pi - V^\pi\|_\infty &\overset{(i)}{\leq} \gamma \max\Bigg\{ \frac{2}{\tau} (I - \gamma \underline{\widehat{P}}^{\pi,\widehat{V}})^{-1} \left\| \underline{\widehat{P}}^{\pi,V} V^\pi - \underline{P}^{\pi,V} V^\pi \right\|_\infty \mathbf{1}, \\
&\qquad\qquad \frac{2}{\tau} (I - \gamma \underline{\widehat{P}}^{\pi,V})^{-1} \left\| \underline{\widehat{P}}^{\pi,V} V^\pi - \underline{P}^{\pi,V} V^\pi \right\|_\infty \mathbf{1} \Bigg\} \\
&\overset{(ii)}{\leq} \frac{2\gamma}{\tau(1-\gamma)} \sqrt{\frac{2 \log(6SAN/\delta)}{N(1-\gamma)^2}},
\end{aligned} \tag{48}
$$

where (i) holds by $\left(I - \gamma \underline{\widehat{P}}^{\pi,V} \right)^{-1} = \sum_{t=0}^\infty \gamma^t (\underline{\widehat{P}}^{\pi,V})^t \geq 0$, (ii) follows from

$$\left(I - \gamma \underline{\widehat{P}}^{\pi,V} \right)^{-1} \mathbf{1} = \sum_{t=0}^\infty \gamma^t (\underline{\widehat{P}}^{\pi,V})^t \mathbf{1} = \frac{1}{1-\gamma} \mathbf{1}. \tag{49}$$

Finally take $\epsilon_{opt} \leq \frac{\sqrt{2 \log(6SAN/\delta)}}{\tau(1-\gamma)\sqrt{N}}$, and plug (48) back to (39). We then have that with probability at least $1 - \delta$,

$$
\begin{aligned}
\|V^* - V^{\widehat{\pi}}\|_\infty &\leq \|V^{\pi^*} - \widehat{V}^{\pi^*}\|_\infty + \frac{2\gamma\epsilon_{\text{opt}}}{1-\gamma} + \|\widehat{V}^{\widehat{\pi}} - V^{\widehat{\pi}}\|_\infty \\
&\leq \frac{6\gamma}{\tau(1-\gamma)} \sqrt{\frac{2 \log(6SAN/\delta)}{N(1-\gamma)^2}} \\
&\leq 6 \sqrt{\frac{2 \log(6SAN/\delta)}{N(1-\gamma)^4 \tau^2}}.
\end{aligned} \tag{50}
$$

B.1 Tighter bound using Berstein inequality when $\tau \geq \gamma$

Note that the bound in (50) applied to any τ and γ. In this subsection, we consider the scenario where $\tau \geq \gamma$. We show that a tighter bound can be achieved.

Then, for a fixed x that is independent with $P_{s,a}$, using Bernstein inequality, one has that with probability at least $1 - \delta$,

$$
\begin{aligned}
g_{s,a}(x, V) &\leq \sqrt{\frac{2\log(\frac{2}{\delta})}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(\min(V, x))} + \frac{2\log\left(\frac{2}{\delta}\right)}{3N(1-\gamma)} \\
&\leq \sqrt{\frac{2\log(\frac{2}{\delta})}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(V)} + \frac{2\log\left(\frac{2}{\delta}\right)}{3N(1-\gamma)}.
\end{aligned}
\tag{51}
$$

To derive the union bound, we can similarly construct a ϵ_1-net over $[0, \frac{1}{1-\gamma}]$ with size $|\mathcal{N}_{\epsilon_1}| \leq \frac{3}{\epsilon_1(1-\gamma)}$. By the union bound and (51), it holds with probability at least $1 - \frac{\delta}{SA}$ for all $x \in \mathcal{N}_{\epsilon_1}$,

$$
g_{s,a}(x, V) \leq \sqrt{\frac{2\log(\frac{2SA|\mathcal{N}_{\epsilon_1}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(V)} + \frac{2\log\left(\frac{2SA|\mathcal{N}_{\epsilon_1}|}{\delta}\right)}{3N(1-\gamma)}.
\tag{52}
$$

From (42), we can show that with probability at least $1 - \frac{\delta}{SA}$,

$$
\begin{aligned}
\left|(\widehat{P}_{s,a}^{\pi,V} - P_{s,a}^{\pi,V})V^\pi\right| &\leq \frac{1}{\tau} \sup_{x \in [0, \frac{1}{1-\gamma}]} \left| \sum_{s'} \left(\widehat{P}_{s,a}(s') - P_{s,a}(s')\right) \right| \\
&\overset{(i)}{\leq} \frac{1}{\tau} \left(\epsilon_1 + \sqrt{\frac{2\log(\frac{2SA|\mathcal{N}_{\epsilon_1}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(V^\pi)} + \frac{2\log\left(\frac{2SA|\mathcal{N}_{\epsilon_1}|}{\delta}\right)}{3N(1-\gamma)} \right) \\
&\overset{(ii)}{=} \frac{1}{\tau} \left(\sqrt{\frac{2\log(\frac{2SA|\mathcal{N}_{\epsilon_1}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(V^\pi)} + \frac{\log\left(\frac{2SA|\mathcal{N}_{\epsilon_1}|}{\delta}\right)}{N(1-\gamma)} \right) \\
&\overset{(iii)}{\leq} \frac{1}{\tau} \left(2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(V^\pi)} + \frac{\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)} \right),
\end{aligned}
\tag{53}
$$

where (i) follows from that the optimal x falls into an ϵ_1-ball centered around some point in $\mathcal{N}_{\epsilon_1}$ and $g_{s,a}$ is 1-Lipschitz in x. (ii) stems from taking $\epsilon_1 = \frac{\log(2SA|\mathcal{N}_{\epsilon_1}|/\delta)}{3N(1-\gamma)}$; and (iii) is shown by $|\mathcal{N}_{\epsilon_1}| \leq \frac{3}{\epsilon_1(1-\gamma)} \leq 9N$.

To bound the term on the right-hand side of (**??**), for any policy π, we show that

$$
\begin{aligned}
&\left(I - \gamma\underline{\widehat{P}}^{\pi,V}\right)^{-1} \left(\underline{\widehat{P}}^{\pi,V}V^\pi - \underline{P}^{\pi,V}V^\pi\right) \\
&\overset{(i)}{\leq} \left(I - \gamma\underline{\widehat{P}}^{\pi,V}\right)^{-1} \\
&\overset{(ii)}{\leq} \frac{\log\left(\frac{18SAN}{\delta}\right)}{N\tau(1-\gamma)} \left(I - \gamma\underline{\widehat{P}}^{\pi,V}\right)^{-1}\mathbf{1} + \underbrace{\frac{2}{\tau}\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N}} \left(I - \gamma\underline{\widehat{P}}^{\pi,V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi,V}}(V^\pi)}}_{=:\mathcal{C}_1} \\
&\quad + \underbrace{\frac{2}{\tau}\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N}} \left(I - \gamma\underline{\widehat{P}}^{\pi,V}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\widehat{P}^\pi}(V^\pi) - \mathrm{Var}_{\underline{\widehat{P}}^{\pi,V}}(V^\pi)\right|}}_{=:\mathcal{C}_2} \\
&\quad + \underbrace{\frac{2}{\tau}\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N}} \left(I - \gamma\underline{\widehat{P}}^{\pi,V}\right)^{-1}\left(\sqrt{\mathrm{Var}_{P^\pi}(V^\pi)} - \sqrt{\mathrm{Var}_{\widehat{P}^\pi}(V^\pi)}\right)}_{=:\mathcal{C}_3},
\end{aligned}
\tag{54}
$$

where (i) holds by $\left(I - \gamma \widehat{\underline{P}}^{\pi,V}\right)^{-1} = \sum_{t=0}^{\infty} \gamma^t (\widehat{\underline{P}}^{\pi,V})^t \geq 0$; and (ii) holds with high probability by the bound in (53) and

$$\sqrt{\text{Var}_{P^\pi}(V^\pi)} \leq \left(\sqrt{\text{Var}_{P^\pi}(V^\pi)} - \sqrt{\text{Var}_{\widehat{P}^\pi}(V^\pi)}\right) + \sqrt{\left|\text{Var}_{\widehat{P}^\pi}(V^\pi) - \text{Var}_{\widehat{\underline{P}}^{\pi,V}}(V^\pi)\right|} + \sqrt{\text{Var}_{\widehat{\underline{P}}^{\pi,V}}(V^\pi)}. \quad (55)$$

Lemma 4 (Agarwal et al. (2020), Lemma 4). *For any policy π and transition kernel P,*

$$\left\| (I - \gamma P^\pi)^{-1} \sqrt{\text{Var}_{P^\pi}(V^\pi)} \right\|_\infty \leq \sqrt{\frac{2}{(1-\gamma)^3}}. \quad (56)$$

Applying Lemma 4, then $\mathcal{C}_1$ in (54) can be bounded as follows

$$\mathcal{C}_1 = \frac{2}{\tau} \sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N}} \left(I - \gamma \widehat{\underline{P}}^{\pi,V}\right)^{-1} \sqrt{\text{Var}_{\widehat{\underline{P}}^{\pi,V}}(V^\pi)} \leq 2\sqrt{\frac{2\log\left(\frac{18SAN}{\delta}\right)}{N\tau^2(1-\gamma)^3}}\mathbf{1}. \quad (57)$$

Lemma 5. *Consider a CVaR uncertainty set $\mathcal{U}^\tau$ defined in (19), then it holds that*

$$\forall \bar{P}_{s,a} \in \mathcal{U}^\tau(P_{s,a}) : \|P_{s,a} - \bar{P}_{s,a}\|_1 \leq \frac{2(1-\tau)}{\tau}. \quad (58)$$

Then we have for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $\bar{P}_{s,a} \in \mathcal{U}^\tau(P_{s,a})$:

$$\left|\text{Var}_{\bar{P}_{s,a}}(V^\pi) - \text{Var}_{P_{s,a}}(V^\pi)\right| \leq 3\left\|\bar{P}_{s,a} - P_{s,a}\right\|_1 \|V^\pi\|_\infty^2 \leq 6\frac{1-\tau}{\tau}\frac{1}{(1-\gamma)^2} \overset{(i)}{\leq} \frac{6}{\tau(1-\gamma)}, \quad (59)$$

and (i) holds when $\tau \geq \gamma$. We then have that

$$\begin{aligned}
\mathcal{C}_2 &= \frac{2}{\tau}\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N}} \left(I - \gamma\widehat{\underline{P}}^{\pi,V}\right)^{-1} \sqrt{\left|\text{Var}_{\widehat{P}^\pi}(V^\pi) - \text{Var}_{\widehat{\underline{P}}^{\pi,V}}(V^\pi)\right|} \\
&= \frac{2}{\tau}\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N}} \left(I - \gamma\widehat{\underline{P}}^{\pi,V}\right)^{-1} \sqrt{\left|\Pi^\pi\left(\text{Var}_{\widehat{P}}(V^\pi) - \text{Var}_{\widehat{P}^\pi,V}(V^\pi)\right)\right|} \\
&\leq \frac{2}{\tau}\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N}} \left(I - \gamma\widehat{\underline{P}}^{\pi,V}\right)^{-1} \sqrt{\left\|\text{Var}_{\widehat{P}}(V^\pi) - \text{Var}_{\widehat{P}^\pi,V}(V^\pi)\right\|_\infty}\mathbf{1} \\
&\overset{(i)}{\leq} \frac{2}{\tau}\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N}} \left(I - \gamma\widehat{\underline{P}}^{\pi,V}\right)^{-1} \sqrt{\frac{6(1-\tau)}{\tau(1-\gamma)^2}}\mathbf{1} \\
&\overset{(ii)}{=} 2\sqrt{\frac{6(1-\tau)\log\left(\frac{18SAN}{\delta}\right)}{N\tau^3(1-\gamma)^4}}\mathbf{1},
\end{aligned} \quad (60)$$

where (i) follows from (59); and (ii) follows from (49).

In the following, we then bound $\mathcal{C}_3$. The following lemma plays an important role.

Lemma 6 (Panaganti & Kalathil, 2022, Lemma 6). *Consider any $\delta \in (0,1)$. For any fixed policy π and fixed value function vector $V \in \mathbb{R}^S$, one has that with probability at least $1 - \delta$,*

$$\left|\sqrt{\text{Var}_{\widehat{P}^\pi}(V)} - \sqrt{\text{Var}_{P^\pi}(V)}\right| \leq \sqrt{\frac{2\|V\|_\infty^2 \log\left(\frac{2SA}{\delta}\right)}{N}}\mathbf{1}. \quad (61)$$

Plugging (61) into (54), we can show that with high probability

$$\begin{aligned}
\mathcal{C}_3 &= \frac{2}{\tau}\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N}} \left(I - \gamma\widehat{\underline{P}}^{\pi,V}\right)^{-1} \left(\sqrt{\text{Var}_{P^\pi}(V^\pi)} - \sqrt{\text{Var}_{\widehat{P}^\pi}(V^\pi)}\right) \\
&\leq \frac{4}{1-\gamma}\frac{\log\left(\frac{18SAN}{\delta}\right)\|V^\pi\|_\infty}{N\tau}\mathbf{1} \leq \frac{4\log\left(\frac{18SAN}{\delta}\right)}{N\tau(1-\gamma)^2}\mathbf{1}.
\end{aligned} \quad (62)$$

Finally, plugging the results of $\mathcal{C}_1$ in (57), $\mathcal{C}_2$ in (60) and $\mathcal{C}_3$ in (62) back into (54), we have

$$
\begin{aligned}
\left(I - \gamma \widehat{\underline{P}}^{\pi,V}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,V} V^\pi - \underline{P}^{\pi,V} V^\pi\right) &\leq 2\sqrt{\frac{2\log\left(\frac{18SAN}{\delta}\right)}{N\tau^2(1-\gamma)^3}}\mathbf{1} + 2\sqrt{\frac{6(1-\tau)\log\left(\frac{18SAN}{\delta}\right)}{N\tau^3(1-\gamma)^4}}\mathbf{1} \\
&\quad + \frac{4\log\left(\frac{18SAN}{\delta}\right)}{N\tau(1-\gamma)^2}\mathbf{1} + \frac{\log\left(\frac{18SAN}{\delta}\right)}{N\tau(1-\gamma)^2}\mathbf{1} \\
&= \frac{5\log\left(\frac{18SAN}{\delta}\right)}{N\tau(1-\gamma)^2}\mathbf{1} + \sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N\tau^2(1-\gamma)^3}}\left(2\sqrt{2} + 2\sqrt{6\frac{1-\tau}{\tau(1-\gamma)}}\right)\mathbf{1} \\
&\overset{(i)}{\leq} \frac{10\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)^2}\mathbf{1} + 32\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)^3}}\mathbf{1},
\end{aligned}
\tag{63}
$$

where (i) holds when $\tau \geq \gamma \geq \frac{1}{2}$.

Plugging (63) back into (**??**), we have that

$$
\left\|V^\pi - \widehat{V}^\pi\right\|_\infty \leq \frac{10\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)^2} + 32\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)^3}}.
\tag{64}
$$

Summing up the results for $\widehat{\pi}$ and π^* and plugging back to (39) complete the proof as follows: taking $\epsilon_{\text{opt}} \leq \frac{\log\left(\frac{18SAN}{\delta}\right)}{\gamma(1-\gamma)N}$ and $N \geq \frac{\log\left(\frac{18SAN}{\delta}\right)}{(1-\gamma)^2}$, with probability at least $1 - \delta$,

$$
\begin{aligned}
\left\|V^* - V^{\widehat{\pi}}\right\|_\infty &\leq \frac{2\gamma\epsilon_{\text{opt}}}{1-\gamma} + \left\|\widehat{V}^{\widehat{\pi}} - V^{\widehat{\pi}}\right\|_\infty + \left\|V^{\pi^*} - \widehat{V}^{\pi^*}\right\|_\infty \\
&\leq \frac{2\gamma\epsilon_{\text{opt}}}{1-\gamma} + \frac{20\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)^2} + 64\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)^3}} \\
&\leq \frac{22\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)^2} + 64\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)^3}} \\
&\leq 86\sqrt{\frac{\log\left(\frac{18SAN}{\delta}\right)}{N(1-\gamma)^3}},
\end{aligned}
\tag{65}
$$

where the last inequality holds if $N \geq \frac{\log\left(\frac{18SAN}{\delta}\right)}{(1-\gamma)^2}$, with probability at least $1 - \delta$.

C Proof of Theorem 2: Sample Complexity Lower Bound

C.1 Construction of hard problem instances

Construction of two hard MDPs. Suppose there are two standard MDPs defined as below:

$$
\mathcal{M}_\phi = \{(\mathcal{S}, \mathcal{A}, P^\phi, r, \gamma) | \phi = 0, 1\}.
$$

Here, γ is the discount factor, $\mathcal{S} = \{0, 1, ..., S-1\}$ is the state space. Given any state $s \in \{2, 3, ..., S-1\}$, the coresponding action spaces are $\mathcal{A} = \{0, 1, 2, ..., A-1\}$. While for state $s = 0$ and $s = 1$, the action space is only $\mathcal{A}' = \{0, 1\}$. For any $\phi \in \{0, 1\}$, the transition kernel P^ϕ of the constructed MDP $\mathcal{M}_\phi$ is defined as

$$
P^\phi(s'|s, a) = \begin{cases} p\mathbb{1}(s' = 1) + (1-p)\mathbb{1}(s' = 0), & \text{if} \quad (s, a) = (0, \phi) \\ q\mathbb{1}(s' = 1) + (1-q)\mathbb{1}(s' = 0), & \text{if} \quad (s, a) = (0, 1-\phi), \\ \mathbb{1}(s' = 1), & \text{if} \quad s \geq 1 \end{cases}
\tag{66}
$$

where p and q are set to satisfy

$$
0 \leq p \leq 1 \qquad \text{and} \qquad 0 \leq q = p - \Delta
\tag{67}
$$

for some p and $\Delta > 0$. The above transition kernel P^ϕ implies that State 1 is an absorbing state.

Then we define the reward function as:

$$r(s,a) = \begin{cases} 1, & \text{if } s = 1 \\ 0, & \text{otherwise.} \end{cases}$$

Additionally, we choose the following initial state distribution:

$$\varphi(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{otherwise.} \end{cases}$$

Uncertainty set of the transition kernel. Recall that the uncertainty set of CVaR is:

$$\mathcal{U}^\tau(P_{s,a}) = \left\{ \bar{P}_{s,a} \in \Delta(\mathcal{S}), \ 0 \leq \frac{\bar{P}_{s,a}(s')}{P_{s,a}(s')} \leq \frac{1}{\tau} \right\}. \tag{68}$$

We define p and Δ:

$$p = (1-\tau) + c\tau(1-\gamma) \quad and \quad \Delta \leq c\tau(1-\gamma), \tag{69}$$

where $c \in (0,1)$. Consequently, applying (67) directly leads to

$$p \geq q \geq 1 - \tau. \tag{70}$$

For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we denote the smallest transition probability of moving to the next state $s' \in \{0,1\}$ in the uncertainty set as

$$\underline{P}^\phi(1|s,a) := \inf_{P_{s,a} \in \mathcal{U}^\tau(P_{s,a}^\phi)} P(1|s,a) = 1 - \min\left\{\frac{1}{\tau}P_{s,a}^\phi(0), 1\right\} = \frac{1}{\tau}\max\left\{P_{s,a}^\phi(1) - (1-\tau), 0\right\}, \tag{71a}$$

$$\underline{P}^\phi(0|s,a) := \inf_{P_{s,a} \in \mathcal{U}^\tau(P_{s,a}^\phi)} P(0|s,a) = \min\left\{\frac{1}{\tau}P_{s,a}^\phi(0), 1\right\}, \tag{71b}$$

where the last equation can be verified by the definition of $\mathcal{U}^\tau(P)$. We further define the following notation

$$\underline{p} := \underline{P}^\phi(1|0,\phi) = \frac{1}{\tau}\max\{p - (1-\tau), 0\}, \quad \underline{q} := \underline{P}^\phi(1|0, 1-\phi) = \frac{1}{\tau}\max\{q - (1-\tau), 0\}, \tag{72}$$

which follows from the fact that $p \geq q \geq 1 - \tau$ in (70).

Robust value function and optimal robust policies. For any MDP $\mathcal{M}^\phi$ with the above uncertainty set, we denote π_ϕ^* as the optimal policy, and the robust value function of any policy π as V_ϕ^π. Then we introduce the following lemma, which describes some important properties of the robust value function:

Lemma 7. *For any $\phi = \{0,1\}$ and any policy π, the value function satisfies*

$$V_\phi^\pi(0) = \frac{\gamma z_\phi^\pi}{(1-\gamma)(1 - \gamma(1 - z_\phi^\pi))}, \tag{73}$$

where z_ϕ^π is defined as

$$z_\phi^\pi = 1 - \frac{1}{\tau}(1 - p\pi(\phi|0) - q\pi(1 - \phi|0))$$
$$= \underline{p}\pi(\phi|0) + \underline{q}\pi(1 - \phi|0). \tag{74}$$

The optimal value function and the optimal policy satisfy

$$V_\phi^*(0) = \frac{\gamma\underline{p}}{(1-\gamma)(1 - \gamma(1 - \underline{p}))}, \tag{75a}$$

$$\pi_\phi^*(\phi|s) = 1, \quad for \ s \in \mathcal{S}. \tag{75b}$$

C.2 Establishing the minimax lower bound

Note that our goal is to quantify the sub-optimality gap between the policy estimator $\widehat{\pi}$ and the optimal policy π^* on the initial state distribution φ, which is

$$\langle \varphi, V_\phi^* - V_\phi^{\widehat{\pi}} \rangle = V_\phi^*(0) - V_\phi^{\widehat{\pi}}(0).$$

Step 1: Equivalence to estimating ϕ. With $\epsilon \leq \frac{c}{16(1-\gamma)}$, let $\Delta = 16(1-\gamma)^2 \tau \epsilon$. Applying (73) and (75a) we have that

$$
\begin{aligned}
V_\phi^*(0) - V_\phi^{\widehat{\pi}}(0) &= \frac{\gamma \underline{p}}{(1-\gamma)(1-\gamma(1-\underline{p}))} - \frac{\gamma z_\phi^{\widehat{\pi}}}{(1-\gamma)(1-\gamma(1-z_\phi^{\widehat{\pi}}))} \\
&= \frac{\gamma(\underline{p} - z_\phi^{\widehat{\pi}})}{(1-\gamma(1-\underline{p}))(1-\gamma(1-z_\phi^{\widehat{\pi}}))} \\
&= \frac{\gamma(\underline{p}-q)}{(1-\gamma)^2 \left(1 + \frac{\gamma \underline{p}}{1-\gamma}\right)\left(1 + \frac{\gamma z_\phi^{\widehat{\pi}}}{1-\gamma}\right)}(1 - \widehat{\pi}(\phi|0)) \\
&\overset{(i)}{=} \frac{\gamma \Delta}{\tau(1-\gamma)^2 \left(1 + \frac{\gamma \underline{p}}{1-\gamma}\right)\left(1 + \frac{\gamma z_\phi^{\widehat{\pi}}}{1-\gamma}\right)}(1 - \widehat{\pi}(\phi|0)) \\
&\overset{(ii)}{\geq} \frac{\gamma}{(1-\gamma)^2} \frac{1}{\left(1 + \frac{\gamma \underline{p}}{1-\gamma}\right)^2}\left(\frac{\Delta}{\tau}\right)(1 - \widehat{\pi}(\phi|0)) \\
&\overset{(iii)}{=} \frac{\gamma}{\tau(1-\gamma^2)} \frac{1}{(1+c\gamma)^2}\Delta(1 - \widehat{\pi}(\phi|0)) \\
&\overset{(iv)}{\geq} 2\epsilon(1 - \widehat{\pi}(\phi|0)),
\end{aligned}
\tag{76}
$$

where (i) holds by the definition of $\underline{p}$ and q; (ii) follows from $z_\phi^{\widehat{\pi}} \leq \underline{p}$; (iii) follows from the definition of $\underline{p}$; (iv) follows from $\gamma \geq \frac{1}{2}$.

With this connection between the sub-optimality gap and the policy $\widehat{\pi}(\phi|0)$, if the policy $\widehat{\pi}$ is ϵ-optimal with a high probability, i.e.,

$$P\left(\langle \varphi, V_\phi^* - V_\phi^{\widehat{\pi}} \rangle \leq \epsilon\right) \geq 1 - \delta, \tag{77}$$

then, we need $\widehat{\pi}(\phi|0) \geq \frac{1}{2}$ with probably at least $1 - \delta$. With this in mind, we can construct an estimator $\widehat{\phi}$ for the better action ϕ

$$\widehat{\phi} = \operatorname*{arg\,max}_{a \in \{0,1\}} \widehat{\pi}(a|0), \tag{78}$$

which satisfies

$$P\left(\widehat{\phi} = \phi\right) = P\left(\widehat{\pi}(\phi|0) > \frac{1}{2}\right) \geq 1 - \delta. \tag{79}$$

The problem now becomes to produce a correct estimator $\widehat{\phi}$ with high probability. Subsequently, the goal is to demonstrate that (79) cannot hold without a sufficient number of samples.

Step 2: Probability of error in testing two hypotheses. Equipped with the aforementioned ground-work, we can now delve into differentiating between the two hypotheses $\phi \in \{0,1\}$. To achieve this, we consider the concept of minimax probability of error, defined as follows:

$$p_e := \inf_\psi \max\{P_0(\Psi \neq 0), P_1(\Psi \neq 1)\}. \tag{80}$$

Here, the infimum is taken over all possible tests Ψ constructed from the samples generated from the nominal transition kernel P^ϕ.

Moving forward, let us denote μ_ϕ (resp. $\mu_\phi(\text{s})$) as the distribution of a sample tuple (s_i, a_i, s_i') under the nominal transition kernel P^ϕ associated with $\mathcal{M}_\phi$ and the samples are generated independently. Applying standard results from Chagny (2016) and the additivity of the KL divergence, we obtain

$$
\begin{aligned}
p_e &\geq \frac{1}{4} \exp\left(-NSA \cdot \text{KL}\left(\mu_0 || \mu_1\right)\right) \\
&= \exp\left\{-N\left(\text{KL}(P^0(\cdot|0,0)||P^1(\cdot|0,0)) + \text{KL}(P^0(\cdot|0,1)||P^1(\cdot|0,1))\right)\right\},
\end{aligned}
\tag{81}
$$

where the last inequality holds by observing that

$$
\begin{aligned}
\text{KL}(\mu_0||\mu_1) &= \frac{1}{SA} \sum_{s',a',s'} \text{KL}\left(P^0(s'||s,a)||P^1(s'||s,a)\right) \\
&= \frac{1}{SA} \sum_{a \in \{0,1\}} \text{KL}\left(P^0(s'||0,a)||P^1(s'||0,a)\right).
\end{aligned}
\tag{82}
$$

Now our focus shifts to bound the KL divergence in (81). Applying Lemma 2.7 in Iyengar (2005) gives

$$
\begin{aligned}
\text{KL}\left(P^0(s'||0,1)||P^1(s'||0,1)\right) = \text{Kl}(p||q) &\leq \frac{(p-q)^2}{p(1-p)} = \frac{\Delta^2}{p(1-p)} \\
&= \frac{256(1-\gamma)^4 \tau^2 \epsilon^2}{p(1-p)} \\
&\overset{(i)}{\leq} \frac{256(1-\gamma)^4 \tau^2 \epsilon^2}{\gamma\tau(1-\gamma\tau)} \\
&\overset{(ii)}{\leq} \frac{512(1-\gamma)^4 \tau \epsilon^2}{1-\gamma\tau},
\end{aligned}
\tag{83}
$$

where (i) stems from $1 - p = \tau - c\tau(1-\gamma) = \tau(1 - c(1-\gamma)) \geq \gamma\tau$, and $p(1-p) \geq \min\{\tau(1-\tau), \gamma\tau(1-\gamma\tau)\} = \gamma\tau(1-\gamma\tau)$; and (ii) follows from $\gamma \geq \frac{1}{2}$. Note that $\text{KL}\left(P^0(s'||0,0)||P^1(s'||0,0)\right)$ can be bounded in the same procedure. Substitute (83) back in to (81), if the sample size is selected as

$$
N \leq \frac{\log\left(\frac{1}{4\delta}\right)(1-\gamma\tau)}{1024(1-\gamma)^4 \tau \epsilon^2},
\tag{84}
$$

then one necessarily has

$$
p_e \geq \frac{1}{4} \exp\left\{-N \cdot \frac{1024(1-\gamma)^4 \tau \epsilon^2}{1-\gamma\tau}\right\} \geq \delta.
\tag{85}
$$

Step 3: putting the results together. Combined all the results above, suppose there is a policy $\widehat{\pi}$ such that

$$
P_0\{\langle \varphi, V_0^* - V_0^{\widehat{\pi}} \rangle > \epsilon\} < \delta \qquad and \qquad P_1\{\langle \varphi, V_1^* - V_1^{\widehat{\pi}} \rangle > \epsilon\} < \delta.
\tag{86}
$$

According to the discussion in Step 1, the estimated $\widehat{\phi}$ must satisfy

$$
P_0(\widehat{\phi} \neq 0) < \delta \qquad and \qquad P_1(\widehat{\phi} \neq 1) < \delta.
\tag{87}
$$

However, this cannot be satisfied with the sample size being too small (84). Thus we have completed the proof of the lower bound.

D Proof for Worst Path RL

D.1 Proof of Theorem 3

For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $n(s', s, a)$ denotes the number of times when the next state is s'. Using Lemma F.4 in Dann et al. (2017), we have that

$$
P\left(n(s', s, a) \geq \frac{1}{2}NP(s'|s,a) - 2\log\left(\frac{SA}{\delta}\right), \ \forall(s,a) \in \mathcal{S} \times \mathcal{A}\right) \geq 1 - \delta.
\tag{88}
$$

When $N \geq \frac{2}{p_{\min}} \left(1 + 2\log\left(\frac{SA}{\delta}\right)\right)$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathrm{supp}\left(P(\cdot|s,a)\right)$, we can show that

$$P\left(n(s',s,a) \geq \frac{1}{2}NP(s'|s,a) - 2\log\left(\frac{SA}{\delta}\right)\right)$$
$$\leq P\left(n(s',s,a) \geq \frac{P(s'|s,a)}{p_{\min}}\left(1 + 2\log\left(\frac{SA}{\delta}\right)\right) - 2\log\left(\frac{SA}{\delta}\right)\right) \overset{(i)}{\leq} P\left(n(s',s,a) \geq 1\right), \quad (89)$$

where (i) is derived from $P(s'|s,a) \geq p_{\min}$ for $s' \in \mathrm{supp}\left(P(\cdot|s,a)\right)$. Combining with (88), we have that

$$\forall s' \in \mathrm{supp}\left(P(\cdot|s,a)\right), \ P\left(n(s',s,a) \geq 1\right) \geq 1 - \delta. \quad (90)$$

We decompose the sub-optimality gap in the same manner as in (39). We notice that with probability at least $1 - \delta$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$(\widehat{P}_{s,a}^{\pi,V} - P_{s,a}^{\pi,V})V^\pi = \min_{s' \in \mathrm{supp}(\widehat{P}(\cdot|s,a))} V^\pi(s') - \min_{s' \in \mathrm{supp}(P(\cdot|s,a))} V^\pi(s') = 0. \quad (91)$$

Plugging (91) back into (**??**), we have

$$\left\|\widehat{V}^\pi - V^\pi\right\|_\infty = 0. \quad (92)$$

Then the first and the third term in (39) then disappears, and we get that

$$\|V^* - V^{\widehat{\pi}}\|_\infty \leq \frac{2\gamma\epsilon_{\mathrm{opt}}}{1 - \gamma}. \quad (93)$$

When $N \geq \frac{2}{p_{\min}}\left(1 + 2\log\left(\frac{SA}{\delta}\right)\right)$, for any $\epsilon > 0$, if we take $\epsilon_{opt} \leq \frac{\epsilon(1-\gamma)}{2\gamma}$, then with probability at least $1 - \delta$, we have that

$$\|V^* - V^{\widehat{\pi}}\|_\infty \leq \epsilon, \quad (94)$$

which concludes the proof.

D.2 Proof of Theorem 4

In the following, we establish a sample complexity lower bound for Worst Path RL. We first construct two hard MDP instances similar to those in Section C.2. Suppose there are two standard MDPs defined below

$$\mathcal{M}_\phi = \{(\mathcal{S}, \mathcal{A}, P^\phi, r, \gamma) | \phi = 0, 1\}.$$

The transition kernel is defined as

$$P^\phi(s'|s,a) = \begin{cases} \mathbf{1}(s'=1), & \text{if} \quad (s,a) = (0,\phi) \\ (1 - p_{\min})\mathbf{1}(s'=0) + p_{\min}\mathbf{1}(s'=1), & \text{if} \quad (s,a) = (0, 1-\phi) \\ \mathbf{1}(s'=1), & \text{if} \quad s \geq 1, \end{cases} \quad (95)$$

The reward function is defined as

$$r(s,a) = \begin{cases} 1, & \text{if} \quad s = 1 \\ 0, & \text{otherwise}. \end{cases} \quad (96)$$

The initial state distribution is also the same

$$\varphi(s) = \begin{cases} 1, & \text{if} \quad s = 0 \\ 0, & \text{otherwise}. \end{cases} \quad (97)$$

Since state $s = 1$ is an absorbing state and has reward 1, the value function at state 1 for any policy π is $V^\pi(1) = \frac{1}{1-\gamma}$. At state $s \in \{2, 3, ..., S-1\}$, applying the Bellman operator we have $V^\pi(s) = \frac{\gamma}{1-\gamma}$. At state $s = 0$, we have that

$$V_\phi^\pi(0) = \frac{\gamma}{1-\gamma}\pi(\phi|0), \quad (98)$$

$$V_\phi^*(0) - V_\phi^\pi(0) = \frac{\gamma}{1-\gamma}\left(1 - \pi(\phi|0)\right). \quad (99)$$

If we want a policy π to be ϵ-optimal with high probability:

$$P\left\{\langle\varphi, V_\phi^* - V_\phi^\pi\rangle \le \epsilon\right\} \ge 1 - \delta, \tag{100}$$

then if $\epsilon \le \frac{\gamma}{2(1-\gamma)}$, we necessarilly need $\pi(\phi|0) \ge \frac{1}{2}$ with probability at least $1 - \delta$. Following the same procedure in the proof of the lower bound for Iterated CVaR RL, we constructed the estimator $\widehat{\phi}$ for the better action (78) that satisfies (79). We also notice that if $n(0, 0, a) \ge 1$ in the N samples, we can definitely tell the other action is superior. Otherwise, we cannot tell the difference between the two actions, and in this case, the probability of a correct guess is $\frac{1}{2}$. With this in mind, we have

$$P\left\{\widehat{\phi} = \phi\right\} = 1 - (1 - p_{\min})^N + (1 - p_{\min})^N \frac{1}{2}. \tag{101}$$

Inserting (101) into 79 we have that

$$1 - \delta \le P\left\{\widehat{\phi} = \phi\right\} = 1 - \frac{1}{2}(1 - p_{\min})^N, \tag{102}$$

which further implies that

$$N \ge \frac{\log(1/2\delta)}{\log(1/(1-p_{\min}))} \sim \tilde{\mathcal{O}}\left(\frac{1}{p_{\min}}\right). \tag{103}$$

The overall sample complexity lower bound is then $\tilde{\mathcal{O}}\left(\frac{SA}{p_{\min}}\right)$.

E Proof of Lemmas

E.1 Proof of Lemma 5

For any $\bar{P}_{s,a} \in \mathcal{U}^\tau(P_{s,a}) = \left\{\bar{P}_{s,a} \in \Delta(\mathcal{S}), \ 0 \le \frac{\bar{P}_{s,a}(s')}{P_{s,a}(s')} \le \frac{1}{\tau}\right\}$, the total-variation between $\bar{P}_{s,a}$ and $P_{s,a}$ is defined as:

$$\left\|\bar{P}_{s,a} - P_{s,a}\right\|_{TV} = \sup_{\mathcal{S}' \subset \mathcal{S}} \left|\bar{P}_{s,a}(\mathcal{S}') - P_{s,a}(\mathcal{S}')\right| = \frac{1}{2}\left\|\bar{P}_{s,a} - P_{s,a}\right\|_1. \tag{104}$$

Here $\mathcal{S}'$ is a subset of state space $\mathcal{S}$ and $P_{s,a}(\mathcal{S}') = \sum_{s' \in \mathcal{S}'} P_{s,a}(s')$, $\bar{P}_{s,a}(\mathcal{S}') = \sum_{s' \in \mathcal{S}'} \bar{P}_{s,a}(s')$.

By the definition of the uncertainty set (19), for any $\mathcal{S}' \in \mathcal{S}$, we have that

$$0 \le \bar{P}_{s,a}(\mathcal{S}') \le \frac{1}{\tau} P_{s,a}(\mathcal{S}'), \tag{105}$$

$$0 \le 1 - \bar{P}_{s,a}(\mathcal{S}') \le \frac{1}{\tau}(1 - P_{s,a}(\mathcal{S}')). \tag{106}$$

This further implies that

$$1 - \frac{1}{\tau}(1 - P_{s,a}(\mathcal{S}')) \le \bar{P}_{s,a}(\mathcal{S}') \le \frac{1}{\tau} P_{s,a}(\mathcal{S}'). \tag{107}$$

For any $\mathcal{S}' \in \mathcal{S}$,

$$\begin{aligned}
\left|\bar{P}_{s,a}(\mathcal{S}') - P_{s,a}(\mathcal{S}')\right| &\le \max\left\{\left(\frac{1}{\tau} - 1\right) P_{s,a}(\mathcal{S}'), \left(\frac{1}{\tau} - 1\right)(1 - P_{s,a}(\mathcal{S}'))\right\} \\
&= \left(\frac{1}{\tau} - 1\right)\max\left\{P_{s,a}(\mathcal{S}'), 1 - P_{s,a}(\mathcal{S}')\right\} \\
&\le \frac{1}{\tau} - 1.
\end{aligned} \tag{108}$$

Using the second definition of total variation, we obtain that

$$\frac{1}{2}\left\|\bar{P}_{s,a} - P_{s,a}\right\|_1 \le \frac{1}{\tau} - 1. \tag{109}$$

E.2 Proof of Lemma 7

For any $\mathcal{M}_\phi \in \{0,1\}$, we can easily find the value function at state $s = 1$ or any states $s \in \{2,3,\cdots,S-1\}$

$$V_\phi^\pi(1) \overset{(i)}{=} 1 + \gamma V_\phi^\pi(1) = \frac{1}{1-\gamma}, \tag{110a}$$

$$\forall s \in \{2,3,\cdots,S\}: \qquad V_\phi^\pi(s) \overset{(ii)}{=} 0 + \gamma V_\phi^\pi(1) = \frac{\gamma}{1-\gamma}, \tag{110b}$$

where (i) and (ii) are according to the fact that the transitions defined over states $s \geq 1$ in (66) give only one possible next state 1, and by the definition of the uncertainty set in (19), there exists only one transition kernel within the uncertainty set, which is the kernel itself.

The value function at state $s = 0$ with policy π satisfies

$$
\begin{aligned}
V_\phi^\pi(0) &= \mathbb{E}_{a\sim\pi(\cdot|0)}[r(0,a)] + \gamma \sum_{a\in\mathcal{A}} \pi(a|0) \inf_{\bar{P}\in\mathcal{U}^\tau(P_{0,a}^\phi)} \bar{P}V_\phi^\pi \\
&= \gamma \sum_{a\in\mathcal{A}} \pi(a|0) \inf_{\bar{P}\in\mathcal{U}^\tau(P_{0,a}^\phi)} \bar{P}V_\phi^\pi \\
&= \gamma \left(\pi(\phi|0) \inf_{\bar{P}\in\mathcal{U}^\tau(P_{0,\phi}^\phi)} \bar{P}V_\phi^\pi + (1-\pi(\phi|0)) \inf_{\bar{P}\in\mathcal{U}^\tau(P_{0,1-\phi}^\phi)} \bar{P}V_\phi^\pi \right) \\
&\overset{(i)}{=} \gamma \left(\pi(\phi|0) \left(V_\phi^\pi(1)\underline{p} + V_\phi^\pi(0)(1-\underline{p}) \right) + (1-\pi(\phi|0)) \left(V_\phi^\pi(1)\underline{q} + V_\phi^\pi(0)(1-\underline{q}) \right) \right) \\
&= \frac{\gamma z_\phi^\pi}{1-\gamma(1-z_\phi^\pi)} V_\phi^\pi(1) \\
&= \frac{\gamma z_\phi^\pi}{(1-\gamma)(1-\gamma(1-z_\phi^\pi))},
\end{aligned}
\tag{111}
$$

where (i) follows from (71a), (71b) and (72). Note that $V_\phi^\pi(0)$ is increasing in z_ϕ^π and z_ϕ^π is upper bounded by $\underline{p}$ and reaches the upper bound when $\pi(\phi|0) = 1$. Taking $z_\phi^\pi = \underline{p}$ in (111), we get the optimal value function

$$V_\phi^*(0) = \frac{\gamma \underline{p}}{(1-\gamma)(1-\gamma(1-\underline{p}))}. \tag{112}$$