PROVABLY EFFICIENT RISK-SENSITIVE REINFORCEMENT LEARNING: ITERATED CVAR AND WORST PATH

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we study a novel episodic risk-sensitive Reinforcement Learning (RL) problem, named Iterated CVaR RL, which aims to maximize the tail of the reward-to-go at each step, and focuses on tightly controlling the risk of getting into catastrophic situations at each stage. This formulation is applicable to real-world tasks that demand strong risk avoidance throughout the decision process, such as autonomous driving, clinical treatment planning and robotics. We investigate two performance metrics under Iterated CVaR RL, i.e., Regret Minimization and Best Policy Identification. For both metrics, we design efficient algorithms `ICVaR-RM` and `ICVaR-BPI`, respectively, and provide nearly matching upper and lower bounds with respect to the number of episodes K. We also investigate an interesting limiting case of Iterated CVaR RL, called Worst Path RL, where the objective becomes to maximize the minimum possible cumulative reward. For Worst Path RL, we propose an efficient algorithm with constant upper and lower bounds. Finally, the techniques we develop for bounding the change of CVaR due to the value function shift and decomposing the regret via a distorted visitation distribution are novel, and can find applications in other risk-sensitive online learning problems.

1 INTRODUCTION

Reinforcement Learning (RL) (Kaelbling et al., 1996; Szepesvári, 2010; Sutton & Barto, 2018) is a classic online decision-making formulation, where an agent interacts with an unknown environment with the goal of maximizing the obtained reward. Despite the empirical success and theoretical progress of recent RL algorithms, e.g., (Szepesvári, 2010; Agrawal & Jia, 2017; Azar et al., 2017; Zanette & Brunskill, 2019), they focus mainly on the risk-neutral criterion, i.e., maximizing the expected cumulative reward, and can fail to avoid rare but disastrous situations. As a result, existing algorithms cannot be applied to tackle real-world risk-sensitive tasks, such as autonomous driving (Wen et al., 2020) and clinical treatment planning (Coronato et al., 2020), where policies that ensure low risk of getting into catastrophic situations at all decision stages are strongly preferred.

Motivated by the above facts, we investigate Iterated CVaR RL, a novel episodic RL formulation equipped with an important risk-sensitive criterion, i.e., Iterated Conditional Value-at-Risk (CVaR) (Hardy & Wirch, 2004). Here, CVaR (Artzner et al., 1999) is a popular static (single-stage) risk measure which stands for the expected tail reward. Iterated CVaR is a dynamic (multi-stage) risk measure defined upon CVaR by backward iteration, and focuses on the worst portion of the reward-to-go at each stage. In the Iterated CVaR RL problem, an agent interacts with an *unknown* episodic Markov Decision Process (MDP) in order to maximize the worst α-portion of the reward-to-go at each step, where $\alpha \in (0, 1]$ is a given risk level. Under this model, we investigate two important performance metrics, i.e., Regret Minimization (RM), where the goal is to minimize the cumulative regret over all episodes, and Best Policy Identification (BPI), where the performance is measured by the number of episodes required for identifying an optimal policy.

In contrast to existing risk-sensitive RL works (Di Castro et al., 2012; La & Ghavamzadeh, 2013; Fei et al., 2020; 2021a;b), which take all successor states into account and mainly quantify the risk by the variance-related or exponential utility criteria, Iterated CVaR RL primarily concerns the worst portion successor states. Also, different from existing CVaR MDP works (Boda & Filar, 2006; Bäuerle & Ott, 2011; Chow et al., 2015), which maximize the CVaR of the total reward with

known transition, Iterated CVaR RL considers optimizing the performance under bad situations and preventing from getting into catastrophic states *at each step* with *unknown* transition. This Iterated CVaR RL formulation enables us to control the risk of falling into disastrous situations throughout the decision process, and is most suitable for applications where such safety-at-all-time is critical, e.g., autonomous driving (Wen et al., 2020) and clinical treatment planning (Coronato et al., 2020). For example, consider an unmanned helicopter control task (Johnson & Kannan, 2002), where we fly an unmanned helicopter to complete some task. There is a small probability that, at each time during execution, the helicopter encounters a sensing or control failure and does not take the prescribed action. In order to guarantee the safety of the surrounding workers and buildings and the helicopter, it is important to make sure that even if accidents occur, e.g., the sensing or control failure occurs, the taken policy ensures that the helicopter does not crash and cause vital damage.

Iterated CVaR RL faces several unique challenges as follows. (i) The importance (contribution to regret) of a state in Iterated CVaR RL is not proportional to its visitation probability. Specifically, there can be states which are critical (risky) but have a small visitation probability. As a result, the regret for Iterated CVaR RL cannot be decomposed into the estimation error at each step with respect to the visitation distribution, as in standard RL analysis (Jaksch et al., 2010; Azar et al., 2017; Zanette & Brunskill, 2019). (ii) In Iterated CVaR RL, the calculation of estimation error involves bounding the change of CVaR when the true value function shifts to optimistic value function, which is very different from typically bounding the change of expected rewards as in existing RL analysis (Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018). Therefore, Iterated CVaR RL demands brand-new algorithm design and analytical techniques. To tackle the above challenges, we design two efficient algorithms `ICVaR-RM` and `ICVaR-BPI` for the RM and BPI metrics, respectively, equipped with delicate CVaR-adapted value iteration and exploration bonuses to allocate more attention on rare but potentially dangerous states. We also develop novel analytical techniques, for bounding the change of CVaR due to the value function shift and decomposing the regret via a distorted visitation distribution. Lower bounds for both metrics are established to demonstrate the optimality of our algorithms with respect to the number of episodes K. Moreover, we present experiments to validate our theoretical results and show the performance superiority of our algorithm (see Appendix A).

We further study an interesting limiting case of Iterated CVaR RL when α approaches 0, called Worst Path RL, where the goal becomes to maximize the minimum possible cumulative reward (optimize the worst path). This setting corresponds to the scenario where the decision maker is extremely risk-adverse and concerns the worst situation (e.g., in clinical treatment planning (Coronato et al., 2020) where the worst case can be disastrous). We emphasize that Worst Path RL cannot be directly solved by taking $\alpha \to 0$ in Iterated CVaR RL's results, as the results there have a dependency on $\frac{1}{\alpha}$ in both upper and lower bounds. To handle this interesting case, we design a simple yet efficient algorithm `MaxWP` for Worst Path RL, and obtain constant upper and lower regret bounds which are independent of K.

The contributions of this paper are summarized as follows.

- We propose a novel Iterated CVaR RL formulation, where an agent interacts with an unknown environment, with the objective of maximizing the worst α-percent tail of the reward-to-go at each step. This formulation allows one to tightly control risk throughout the decision process, and is most suitable for applications where such safety-at-all-time is critical.
- We investigate two important metrics of Iterated CVaR RL, i.e., Regret Minimization (RM) and Best Policy Identification (BPI), and propose efficient algorithms `ICVaR-RM` and `ICVaR-BPI`. We establish nearly matching regret/sample complexity upper and lower bounds with respect to K. Moreover, we also develop novel techniques to bound the change of CVaR due to the value function shift and decompose regret into estimation error via a distorted visitation distribution, which can be applied to other risk-sensitive decision making problems.
- We further investigate a limiting case of Iterated CVaR RL when α approaches 0, called Worst Path RL, where the objective is to maximize the minimum possible cumulative reward. We develop a simple and efficient algorithm `MaxWP`, and provide constant regret upper and lower bounds (independent of K).

Due to space limit, we defer all proofs and experiments to Appendix.

2 Related Work

Below we review the most related works, and defer a full literature review to Appendix B.

CVaR-based MDPs (known transition). Boda & Filar (2006); Bäuerle & Ott (2011); Chow et al. (2015) study the CVaR MDP problem where the objective is to minimize the CVaR of the total cost with known transition, and show that the optimal policy for CVaR MDP is history-dependent and inefficient to exactly compute (see Appendix C.2 for a comparison with CVaR MDP). Hardy & Wirch (2004) firstly define the Iterated CVaR measure, and prove that it is a coherent dynamic risk measure. Osogami (2012); Chu & Zhang (2014); Bäuerle & Glauner (2022) investigate MDPs with iterated coherent risk measures (including Iterated CVaR), and demonstrate the existence of Markovian optimal policies. The above works focus mainly on the planning side, i.e., proposing algorithms and error guarantees for MDPs with *known* transition, while our work develops RL algorithms (interact with the environment) and regret/sample complexity results for *unknown* transition.

Risk-sensitive Reinforcement Learning (unknown transition). In risk-sensitive RL, Di Castro et al. (2012); La & Ghavamzadeh (2013) study variance-related risk criteria, Tamar et al. (2015) aim to maximize the CVaR of the total reward, and Borkar & Jain (2014); Chow & Ghavamzadeh (2014); Chow et al. (2017) investigate CVaR-based constraints. Heger (1994); Coraluppi & Marcus (1997; 1999) consider minimizing the worst-case cost, and design heuristic algorithms. There are other safe RL works (Cheng et al., 2019; Fatemi et al., 2019; 2021) which focus on state-wise safety (constrain the agent within a set of safe states), and their formulations are very different from ours. A comprehensive survey on safe RL can be found in (Garcıa & Fernández, 2015). The above works mainly investigate convergence analysis, and do not provide regret/sample complexity bounds as us.

To our best knowledge, there are only a few risk-sensitive RL works (Fei et al., 2020; 2021a;b) which establish rigorous regret bounds. Fei et al. (2020; 2021a;b) consider risk-sensitive RL with the exponential utility criterion, and develop exponential Bellmen equations and Bellman backup analytical procedures. Their exponential utility criterion takes all successor states into account, while our Iterated CVaR criterion mainly concerns the worst α-portion successor states, and focuses on optimizing the performance under detrimental situations (see Appendix C.2 for a comparison).

3 Problem Formulation

In this section, we present the problem formulations of Iterated CVaR RL and Worst Path RL.

Conditional Value-at-Risk (CVaR). We first introduce two risk measures, i.e., Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR). Let X be a random variable with cumulative distribution function $F(x) = \Pr[X \leq x]$. Given a risk level $\alpha \in (0, 1]$, the VaR at risk level α is the α-quantile of X, i.e., $\text{Var}^\alpha(X) = \min\{x | F(x) \geq \alpha\}$, and the CVaR at risk level α is defined as (Rockafellar et al., 2000):

$$\text{CVaR}^\alpha(X) = \sup_{x \in \mathbb{R}} \left\{ x - \frac{1}{\alpha} \mathbb{E}\left[(x - X)^+\right] \right\},$$

where $(x)^+ := \max\{x, 0\}$. If there is no probability atom at $\text{CVaR}^\alpha(X)$, CVaR can also be written as $\text{CVaR}^\alpha(X) = \mathbb{E}[X | X \leq \text{Var}^\alpha(X)]$ (Shapiro et al., 2021). Intuitively, $\text{CVaR}^\alpha(X)$ is a distorted expectation of X conditioning on its α-portion tail, which depicts the average value when bad situations happen. When $\alpha = 1$, $\text{CVaR}^\alpha(X) = \mathbb{E}[X]$, and when $\alpha \to 0$, $\text{CVaR}^\alpha(X)$ tends to $\min(X)$ (Chow et al., 2015).

Iterated CVaR RL. We consider an episodic Markov Decision Process (MDP) $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, p, r)$. Here $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and H is the length of horizon in each episode. p is the transition distribution, i.e., $p(s'|s, a)$ gives the probability of transitioning to s' when starting from state s and taking action a. $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is a reward function, and $r(s, a)$ gives a deterministic reward for taking action a in state s. A policy π is defined as a collection of H functions, i.e., $\pi = \{\pi_h : \mathcal{S} \mapsto \mathcal{A}\}_{h \in [H]}$, where $[H] := \{1, 2, ..., H\}$.

The *episodic* RL game is as follows. In each episode k, an agent chooses a policy π^k, and starts from a fixed initial state s_1, i.e., $s_1^k := s_1$, as assumed in many prior RL works (Fiechter, 1994; Kaufmann et al., 2021; Ménard et al., 2021). At each step $h \in [H]$, the agent observes the state s_h^k and takes an action $a_h^k = \pi_h^k(s_h^k)$. After that, it receives a reward $r(s_h^k, a_h^k)$ and transitions to a next

state s_{h+1}^k according to the transition distribution $p(\cdot|s_h^k, a_h^k)$. The episode ends after H steps and the agent enters the next episode.

In Iterated CVaR RL, for any risk level $\alpha \in (0, 1]$ and a policy π, we use value function $V_h^{\alpha,\pi} : \mathcal{S} \mapsto \mathbb{R}$ and Q-value function $Q_h^{\alpha,\pi} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ to denote the cumulative reward that can be obtained when the agent transitions to the worst α-portion states at each step, starting from s and (s, a) at step h, respectively. For simplicity of notation, when the value of α is clear, we omit the superscript α and use the notations V_h^π and Q_h^π. Formally, Q_h^π and V_h^π are recurrently defined in Eq. (i) below. Since $\mathcal{S}$, $\mathcal{A}$ and H are finite and the maximization of $V_h^\pi(s)$ in Iterated CVaR RL satisfies the optimal substructure property, there exists an optimal policy π^* which gives the optimal value $V_h^*(s) = \max_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$ (Chu & Zhang, 2014). Therefore, the Bellman equation and the Bellman optimality equation are given in Eqs. (i),(ii) below, respectively (Chu & Zhang, 2014).

$$\begin{cases} Q_h^\pi(s,a) = r(s,a) + \text{CVaR}^\alpha_{s'\sim p(\cdot|s,a)}(V_{h+1}^\pi(s')) \\ V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)) \\ V_{H+1}^\pi(s) = 0, \ \forall s \in \mathcal{S}, \end{cases} \quad \text{(i)} \qquad \begin{cases} Q_h^*(s,a) = r(s,a) + \text{CVaR}^\alpha_{s'\sim p(\cdot|s,a)}(V_{h+1}^*(s')) \\ V_h^*(s) = \max_{a\in\mathcal{A}} Q_h^*(s,a) \\ V_{H+1}^*(s) = 0, \ \forall s \in \mathcal{S}, \end{cases} \quad \text{(ii)}$$

where $\text{CVaR}^\alpha_{s'\sim p(\cdot|s,a)}(V_{h+1}^\pi(s'))$ denotes the CVaR value of random variable $V_{h+1}^\pi(s')$ with $s' \sim p(\cdot|s, a)$ at risk level α.

We consider two performance metrics for Iterated CVaR RL, i.e., Regret Minimization (RM) and Best Policy Identification (BPI). In Iterated CVaR RL-RM, the agent aims to minimize the cumulative regret in K episodes, defined as

$$\mathcal{R}(K) = \sum_{k=1}^{K} \left(V_1^*(s_1) - V_1^{\pi_k}(s_1)\right). \tag{1}$$

In Iterated CVaR RL-BPI, given a confidence parameter $\delta \in (0, 1]$ and an accuracy parameter $\varepsilon > 0$, the agent needs to use as few trajectories (episodes) as possible to identify an ε-optimal policy $\hat{\pi}$, which satisfies $V_1^{\hat{\pi}}(s_1) \geq V_1^*(s_1) - \varepsilon$, with probability as least $1 - \delta$. That is, the performance of BPI is measured by the number of trajectories used, i.e., sample complexity.

Worst Path RL. Furthermore, we investigate an interesting limiting case of Iterated CVaR RL when α approaches 0, called Worst Path RL. In this case, the objective becomes maximizing the minimum possible reward (Heger, 1994). The Bellman (optimality) equations become

$$\begin{cases} Q_h^\pi(s,a) = r(s,a) + \min_{s'\sim p(\cdot|s,a)}(V_{h+1}^\pi(s')) \\ V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)) \\ V_{H+1}^\pi(s) = 0, \ \forall s \in \mathcal{S}, \end{cases} \quad \begin{cases} Q_h^*(s,a) = r(s,a) + \min_{s'\sim p(\cdot|s,a)}(V_{h+1}^*(s')) \\ V_h^*(s) = \max_{a\in\mathcal{A}} Q_h^*(s,a) \\ V_{H+1}^*(s) = 0, \ \forall s \in \mathcal{S}, \end{cases} \tag{2}$$

where $\min_{s'\sim p(\cdot|s,a)}(V_{h+1}^\pi(s'))$ denotes the minimum value of random variable $V_{h+1}^\pi(s')$ with $s' \sim p(\cdot|s, a)$. From Eq. (2), one sees that

$$Q_h^\pi(s,a) = \min_{(s_t,a_t)\sim\pi}\left[\sum_{t=h}^{H} r(s_t,a_t)\Big| s_h = s, a_h = a, \pi\right], \ V_h^\pi(s) = \min_{(s_t,a_t)\sim\pi}\left[\sum_{t=h}^{H} r(s_t,a_t)\Big| s_h = s, \pi\right].$$

Thus, $Q_h^\pi(s, a)$ and $V_h^\pi(s)$ denote the minimum possible cumulative reward under policy π, starting from (s, a) and s at step h, respectively. The optimal policy π^* maximizes the minimum possible cumulative reward (i.e., optimizes the worst path) for all starting states and steps. Formally, π^* gives the optimal value $V_h^*(s) = \max_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$.

For Worst Path RL, in this paper we mainly consider the regret minimization setting, where the regret is defined the same as Eq. (1). Note that this case cannot be directly solved by taking $\alpha \to 0$ in Iterated CVaR RL, as the results there have a dependency on $\frac{1}{\alpha}$. Thus, changing from $\text{CVaR}(\cdot)$ to $\min(\cdot)$ in Worst Path RL requires a different algorithm design and analysis.

The best policy identification setting of Worst Path RL, on the other hand, is very challenging. This is because we cannot establish confidence intervals under the $\min(\cdot)$ operation, and it is difficult to determine when the estimated optimal policy is accurate enough and when the algorithm should stop. We will further investigate this setting in future work.

Algorithm 1: `ICVaR-RM`

Input: δ, α, $\delta' := \frac{\delta}{5}$, $L := \log(\frac{KHSA}{\delta'})$, $\bar{V}^k_{H+1}(s) = 0$ for any $k > 0$ and $s \in \mathcal{S}$

1 **for** $k = 1, 2, \ldots, K$ **do**
2 **for** $h = H, H-1, \ldots, 1$ **do**
3 **for** $s \in \mathcal{S}$ **do**
4 **for** $a \in \mathcal{A}$ **do**
5 $\bar{Q}^k_h(s,a) \leftarrow \min\left\{r(s,a) + \text{CVaR}^{\alpha}_{s' \sim \hat{p}^k(\cdot|s,a)}(\bar{V}^k_{h+1}(s')) + \frac{H}{\alpha}\sqrt{\frac{L}{n_k(s,a)}},\ H\right\}$;
6 $\bar{V}^k_h(s) \leftarrow \max_{a \in \mathcal{A}} \bar{Q}^k_h(s,a)$. $\pi^k_h(s) \leftarrow \text{argmax}_{a \in \mathcal{A}} \bar{Q}^k_h(s,a)$;
7 Play the episode k with policy π^k, and update $n_{k+1}(s,a)$ and $\hat{p}^{k+1}(s'|s,a)$;

4 Iterated CVaR RL with Regret Minimization

In this section, we consider regret minimization (Iterated CVaR RL-RM). We propose an algorithm `ICVaR-RM` with CVaR-adapted exploration bonuses, and demonstrate its sample efficiency.

4.1 Algorithm `ICVaR-RM` and Regret Upper Bound

We propose a value iteration-based algorithm `ICVaR-RM` (Algorithm 1), which adopts CVaR-adapted (Brown-type (Brown, 2007)) exploration bonuses and pays more attention to rare but risky states. Specifically, in each episode k, `ICVaR-RM` computes the empirical CVaR for the values of next states $\text{CVaR}^{\alpha}_{s' \sim \hat{p}^k(\cdot|s,a)}(\bar{V}^k_{h+1}(s'))$ and Brown-type exploration bonuses $\frac{H}{\alpha}\sqrt{\frac{L}{n_k(s,a)}}$. Here $n^k(s,a)$ is the number of times (s,a) was visited up to episode k, and $\hat{p}^k(s'|s,a)$ is the empirical estimate of transition probability $p(s'|s,a)$. Then, `ICVaR-RM` constructs optimistic Q-value function $\bar{Q}^k_h(s,a)$, optimistic value function $\bar{V}^k_h(s)$, and a greedy policy π^k with respect to $\bar{Q}^k_h(s,a)$. After calculating the value functions and policy, `ICVaR-RM` plays episode k with policy π^k, observes a trajectory, and updates $n_k(s,a)$ and $\hat{p}^{k+1}(s'|s,a)$. The calculation of CVaR (Line 5) can be implemented efficiently, and costs $O(S \log S)$ computation complexity (Shapiro et al., 2021).

We summarize the regret performance of `ICVaR-RM` as follows.

Theorem 1 (Regret Upper Bound)**.** *With probability at least $1-\delta$, the regret of algorithm* `ICVaR-RM` *is bounded by*

$$O\left(\mathbf{min}\left\{\frac{1}{\sqrt{\min_{\pi,h,s:\ w_{\pi,h}(s)>0} w_{\pi,h}(s)}},\ \frac{1}{\sqrt{\alpha^{H-1}}}\right\} \cdot \frac{HS\sqrt{KHA}}{\alpha} \log\left(\frac{KHSA}{\delta}\right)\right),$$

where $w_{\pi,h}(s)$ denotes the probability of visiting state s at step h under policy π.

Remark 1. The regret depends on the *minimum* of an MDP-intrinsic visitation factor $[\min_{\pi,h,s:\ w_{\pi,h}(s)>0} w_{\pi,h}(s)]^{-\frac{1}{2}}$ and $\frac{1}{\sqrt{\alpha^{H-1}}}$. When α is small, the first term dominates the bound, which stands for the minimum probability of visiting an available state under any feasible policy. Note that $\min_{\pi,h,s:\ w_{\pi,h}(s)>0} w_{\pi,h}(s)$ takes the minimum over only the policies under which s is reachable, and thus, this factor will never be zero. Indeed, this factor also exists in the lower bound (see Section 4.2). Thus, it characterizes the essential problem hardness, i.e., when the agent is highly risk-adverse, her regret will be heavily influenced by exploring critical but hard-to-reach states.

When α is large, $\frac{1}{\sqrt{\alpha^{H-1}}}$ instead dominates the bound. The intuition behind the factor $\frac{1}{\sqrt{\alpha^{H-1}}}$ is that for any state-action pair, the ratio of the visitation probability conditioning on transitioning to bad successor states over the original visitation probability can be upper bounded by $\frac{1}{\alpha^{H-1}}$. This ratio is critical and will appear in the regret bound (see Lemma 9 for a formal statement).

In the special case when $\alpha = 1$, our Iterated CVaR RL problem reduces to the classic RL formulation, and our regret bound becomes $\tilde{O}(HS\sqrt{KHA})$, which matches the result in existing classic RL work (Jaksch et al., 2010). This bound has a gap of $\sqrt{HS}$ to the state-of-the-art regret bound for classic RL (Azar et al., 2017; Zanette & Brunskill, 2019). This is because our algorithm is mainly designed for general risk-sensitive cases (which require CVaR-adapted exploration bonuses), and

does not use the Bernstein-type exploration bonuses (which only work for classic expectation maximization criterion). Such phenomenon also appears in existing risk-sensitive RL works (Fei et al., 2020; 2021a). Designing an algorithm which achieves an optimal regret simultaneously for both risk-sensitive cases and classic expectation maximization case is still an open problem, which we leave for future work. To validate our theoretical analysis, we also conduct experiments to exhibit the influences of parameters α, δ, H, S, A and K on the regret of `ICVaR-RM` in practice, and the empirical results well match our theoretical bound (see Appendix A).

Challenges and Novelty in Regret Analysis. There are several unique challenges in the regret analysis for Iterated CVaR RL. (i) First of all, in Iterated CVaR RL, the contribution of a state to the regret is not proportional to its visitation probability as in standard RL analysis (Jaksch et al., 2010; Azar et al., 2017; Zanette & Brunskill, 2019). Instead, the regret is influenced more by risky but hard-to-reach states. Thus, the regret cannot be decomposed into estimation error with respect to visitation distribution. (ii) Second, unlike existing RL analysis (Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018) which typically calculates the change of expected rewards between optimistic and true value functions, in Iterated CVaR RL, we need to instead analyze the change of CVaR due to the value function shift. To tackle these challenges, we develop novel analytical techniques, to bound the deviation of CVaR between optimistic and true value functions and decompose regret into estimation error via a distorted visitation distribution. Below we present a proof sketch for Theorem 1 (see Appendix D.1 for a complete proof).

Proof sketch of Theorem 1. First, we introduce a key inequality (Eq. (3)) to bound the change of CVaR when the true value function shifts to an optimistic one. To this end, let $\beta^{\alpha,V}(\cdot|s,a) \in \mathbb{R}^S$ denote the conditional transition probability conditioning on transitioning to the worst α-portion successor states s', i.e., with the lowest values $V(s')$. It satisfies that $\sum_{s'\in\mathcal{S}} \beta^{\alpha,V}(s'|s,a)\cdot V(s') = \mathrm{CVaR}^{\alpha}_{s'\sim p(\cdot|s,a)}(V(s'))$. Then, for any (s,a) and value functions $\bar{V}, V$ such that $\bar{V}(s') \geq V(s')$ for any $s' \in \mathcal{S}$, we have

$$\mathrm{CVaR}^{\alpha}_{s'\sim p(\cdot|s,a)}(\bar{V}(s')) - \mathrm{CVaR}^{\alpha}_{s'\sim p(\cdot|s,a)}(V(s')) \leq \beta^{\alpha,V}(\cdot|s,a)^{\top}\left(\bar{V} - V\right). \tag{3}$$

Eq. (3) implies that the deviation of CVaR between optimistic and true value functions can be bounded by their value deviation under a distorted transition distribution, which resolves the aforementioned challenge (ii), and serves as the basis of our recurrent regret decomposition.

Now, since $\bar{V}^k_h$ is an optimistic estimate of V^*_h, we decompose the regret in episode k as

$$\begin{aligned}
\bar{V}^k_1(s^k_1) - V^{\pi^k}_1(s^k_1) &\overset{\text{(a)}}{=} \frac{H}{\alpha}\sqrt{\frac{L}{n_k(s^k_1,a^k_1)}} + \mathrm{CVaR}^{\alpha}_{s'\sim\hat{p}^k(\cdot|s^k_1,a^k_1)}(\bar{V}^k_2(s')) - \mathrm{CVaR}^{\alpha}_{s'\sim p(\cdot|s^k_1,a^k_1)}(\bar{V}^k_2(s')) \\
&\quad + \mathrm{CVaR}^{\alpha}_{s'\sim p(\cdot|s^k_1,a^k_1)}(\bar{V}^k_2(s')) - \mathrm{CVaR}^{\alpha}_{s'\sim p(\cdot|s^k_1,a^k_1)}(V^{\pi^k}_2(s')) \\
&\overset{\text{(b)}}{\leq} \frac{H}{\alpha}\sqrt{\frac{L}{n_k(s^k_1,a^k_1)}} + \frac{4H}{\alpha}\sqrt{\frac{SL}{n_k(s^k_1,a^k_1)}} + \beta^{\alpha,V^{\pi^k}_2}(\cdot|s^k_1,a^k_1)^{\top}(\bar{V}^k_2 - V^{\pi^k}_2) \\
&\overset{\text{(c)}}{\leq} \sum_{h=1}^{H}\sum_{(s,a)} w^{\mathrm{CVaR},\alpha,V^{\pi^k}}_{kh}(s,a)\cdot\frac{H\sqrt{L}+4H\sqrt{SL}}{\alpha\sqrt{n_k(s,a)}}
\end{aligned} \tag{4}$$

Here $w^{\mathrm{CVaR},\alpha,V^{\pi^k}}_{kh}(s,a)$ denotes the conditional probability of visiting (s,a) at step h of episode k, conditioning on transitioning to the worst α-portion successor states s' (i.e., with the lowest α-portion values $V^{\pi^k}_{h'+1}(s')$) at each step $h' = 1,\ldots,h-1$. Intuitively, $w^{\mathrm{CVaR},\alpha,V^{\pi^k}}_{kh}(s,a)$ is the distorted visitation probability under the conditional transition probability $\beta^{\alpha,V^{\pi^k}}(\cdot|\cdot,\cdot)$. Inequality (b) uses the concentration of CVaR and Eq. (3). Inequality (c) follows from recurrently applying steps (a)-(b) to unfold $\bar{V}^k_h(\cdot) - V^{\pi^k}_h(\cdot)$ for $h = 2,\ldots,H$, and the fact that $w^{\mathrm{CVaR},\alpha,V^{\pi^k}}_{kh}(s,a)$ is a visitation probability under conditional transition probability $\beta^{\alpha,V^{\pi^k}}(\cdot|\cdot,\cdot)$. Eq. (4) decomposes the regret into estimation error at all state-action pairs via the distorted visitation distribution $w^{\mathrm{CVaR},\alpha,V^{\pi^k}}_{kh}(s,a)$, which overcomes the aforementioned challenge (i).

Summing Eq. (4) over all episodes $k \in [K]$ and using the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(K)] \leq & \frac{5H\sqrt{SL}}{\alpha} \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)} \frac{w_{kh}^{\mathrm{CVaR},\alpha,V^{\pi^k}}(s,a)}{n_k(s,a)}} \cdot \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)} w_{kh}^{\mathrm{CVaR},\alpha,V^{\pi^k}}(s,a)} \\
\overset{\text{(d)}}{=} & \frac{5H\sqrt{SL}\cdot\sqrt{KH}}{\alpha} \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)} \frac{w_{kh}^{\mathrm{CVaR},\alpha,V^{\pi^k}}(s,a)}{w_{kh}(s,a)} \cdot \frac{w_{kh}(s,a)}{n_k(s,a)} \cdot \mathbb{1}\{w_{kh}(s,a) \neq 0\}} \\
\overset{\text{(e)}}{\leq} & \frac{5H\sqrt{KHSL}}{\alpha} \sqrt{\min\left\{\frac{1}{\min\limits_{\pi,h,(s,a):\, w_{\pi,h}(s,a)>0} w_{\pi,h}(s,a)}, \frac{1}{\alpha^{H-1}}\right\} \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)} \frac{w_{kh}(s,a)}{n_k(s,a)}},
\end{aligned}$$

Here $w_{kh}(s,a)$ denotes the probability of visiting (s,a) at step h of episode k. (d) uses the facts that $\sum_{(s,a)} w_{kh}^{\mathrm{CVaR},\alpha,V^{\pi^k}}(s,a) = 1$, and if the visitation probability $w_{kh}(s,a) = 0$, the conditional visitation probability $w_{kh}^{\mathrm{CVaR},\alpha,V^{\pi^k}}(s,a)$ must be 0 as well. (e) is due to that $w_{kh}^{\mathrm{CVaR},\alpha,V^{\pi^k}}(s,a)/w_{kh}(s,a)$ can be bounded by both $1/\min_{\pi,h,(s,a):\, w_{\pi,h}(s,a)>0} w_{\pi,h}(s,a)$ and $1/\alpha^{H-1}$. Specifically, the bound $1/\min_{\pi,h,(s,a):\, w_{\pi,h}(s,a)>0} w_{\pi,h}(s,a)$ follows from $\min_{\pi,h,(s,a):\, w_{\pi,h}(s,a)>0} w_{\pi,h}(s,a) \leq w_{kh}(s,a)$, and the bound $1/\alpha^{H-1}$ comes from the fact that the conditional visitation probability $w_{kh}^{\mathrm{CVaR},\alpha,V^{\pi^k}}(s,a)$ is at most $1/\alpha^{H-1}$ times the visitation probability $w_{kh}(s,a)$. Having established the above, we can use a similar analysis as that in classic RL (Azar et al., 2017; Zanette & Brunskill, 2019) to bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)} \frac{w_{kh}(s,a)}{n_k(s,a)}$, and then, we can obtain Theorem 1. □

4.2 Regret Lower Bound

We now present a regret lower bound to demonstrate the optimality of algorithm `ICVaR-RM`.

Theorem 2 (Regret Lower Bound)**.** *There exists an instance of Iterated CVaR RL-RM, where* $\min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s) > \alpha^{H-1}$ *and the regret of any algorithm is at least*

$$\Omega\left(H\sqrt{\frac{AK}{\alpha \min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s)}}\right). \tag{5}$$

In addition, there exists an instance of Iterated CVaR RL-RM, where $\min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s) < \alpha^{H-1}$ *and the regret of any algorithm is at least* $\Omega(\sqrt{\frac{AK}{\alpha^{H-1}}})$.

Remark 2. Theorem 2 demonstrates that when α is small, the factor $\min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s)$ is inevitable in general. This reveals the intrinsic hardness of Iterated CVaR RL, i.e., when the agent is highly sensitive to bad situations, she must suffer a regret due to exploring risky but hard-to-reach states. This lower bound also validates that `ICVaR-RM` is near-optimal with respect to K.

Lower Bound Analysis. Here we give the proof idea of the first lower bound (Eq. (5)) in Theorem 2, and defer a complete proof to Appendix D.2 due to space limit. We construct an instance with a hard-to-reach bandit state (which has an optimal action and multiple sub-optimal actions), and show that this state is critical for minimizing the regret, but difficult for any algorithm to learn. As shown in Figure 1, we consider an MDP with A actions, n regular states $s_1, \ldots, s_n$ and three absorbing states x_1, x_2, x_3, and the length of each episode $H \gg n$. The reward function $r(s,a)$ depends only on the states, i.e., $s_1, \ldots, s_n$ generate zero reward and x_1, x_2, x_3 generate rewards 1, 0.8 and 0.2, respectively. Under all actions, state s_1 transitions to s_2, x_1, x_2, x_3 with probabilities α, $1-3\alpha$, α and α, respectively, where α is the risk level, and state s_i $(2 \leq i \leq n-1)$ transitions to s_{i+1}, x_1 with probabilities α and $1-\alpha$, respectively. For the bandit state s_n, under the optimal action, s_n transitions to x_2, x_3 with probabilities $1-\alpha+\eta$ and $\alpha-\eta$, respectively. Under sub-optimal actions, s_n transitions to x_2, x_3 with probabilities $1-\alpha$ and α, respectively.

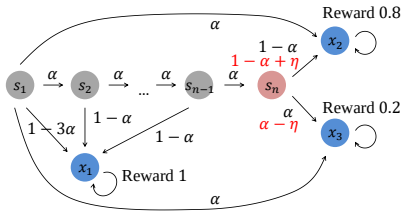

Figure 1: Instance for the lower bound.

Algorithm 2: `MaxWP`

Input: δ, $\delta' := \frac{\delta}{2}$, $L := \log(\frac{SA}{\delta'})$, $\hat{V}^k_{H+1}(s) = 0$ for any $k > 0$ and $s \in \mathcal{S}$

for $k = 1, 2, \ldots, K$ **do**
  for $h = H, H-1, \ldots, 1$ **do**
    for $s \in \mathcal{S}$ **do**
      for $a \in \mathcal{A}$ **do**
        $\hat{Q}^k_h(s,a) \leftarrow r(s,a) + \min_{s' \sim \hat{p}^k(\cdot|s,a)}(\hat{V}^k_{h+1}(s'))$;
      $\hat{V}^k_h(s) \leftarrow \max_{a \in \mathcal{A}} \hat{Q}^k_h(s,a)$;
      $\pi^k_h(s) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}^k_h(s,a)$;
  Play the episode k with policy π^k, and update $n_{k+1}(s,a)$ and $\hat{p}^{k+1}(s'|s,a)$;

In this MDP, under the Iterated CVaR criterion, V^π_1 mainly depends on the path $s_1 \to s_2 \to \cdots \to s_n \to x_2/x_3$, and especially on the action choice in the bandit state s_n. However, when α is small, it is difficult for any policy to reach (and collect information from) s_n. Thus, to learn the optimal action in s_n, any algorithm must suffer a regret dependent on the probability of visiting s_n, which is exactly the minimum visitation probability of any state under a feasible policy $\min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s)$.

5 Iterated CVaR RL with Best Policy Identification

In this section, we design an efficient algorithm `ICVaR-BPI`, and establish sample complexity upper and lower bounds for Iterated CVaR RL with best policy identification (BPI). To our best knowledge, they are the first BPI results for risk-sensitive RL. Due to space limit, we defer the algorithm pseudocode, a detailed algorithm description and a formal statement of lower bound to Appendix E.

5.1 Algorithm `ICVaR-BPI` and Sample Complexity Upper Bound

Algorithm `ICVaR-BPI` is a value iteration-based algorithm. In each episode, `ICVaR-BPI` builds optimistic and pessimistic value functions with CVaR-adapted exploration bonuses, and calculates the estimation error of value functions. As the number of played episodes increases, the estimation error decreases. Once the estimation error shrinks with a given accuracy parameter ε, `ICVaR-BPI` returns the hypothesized optimal policy. The sample complexity of `ICVaR-BPI` is presented as follows.

Theorem 3 (Sample Complexity Upper Bound)**.** *The number of trajectories used by algorithm* `ICVaR-BPI` *to return an ε-optimal policy with probability at least $1-\delta$ is bounded by*

$$O\left(\mathbf{min}\left\{\frac{1}{\min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s)}, \frac{1}{\alpha^{H-1}}\right\}\frac{H^3S^2A}{\varepsilon^2\alpha^2}\cdot C\right),$$

where $C := \log^2(\min\{\frac{1}{\min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s)}, \frac{1}{\alpha^{H-1}}\}\frac{HSA}{\varepsilon\alpha\delta})$.

Similar to Theorem 1, $\min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s)$ and α^{H-1} dominate the bound for a large α and a small α, respectively. When $\alpha = 1$, the problem reduces to the classic RL formulation with best policy identification, and our sample complexity becomes $\tilde{O}(\frac{H^3S^2A}{\varepsilon^2})$, which recovers the result in prior classic RL work (Dann et al., 2017). Similar to Theorem 1, this bound has a gap of HS to the state-of-the-art sample complexity for classic RL (Ménard et al., 2021). This gap is due to the fact that the result in (Ménard et al., 2021) is obtained using the Bernstein-type exploration bonuses, which are more fine-grained for the classic RL problem but do not work for general risk-sensitive cases, because it cannot be used to quantify the estimation error of CVaR.

To validate the tightness of Theorem 3, we further provide sample complexity lower bounds $\Omega(\frac{H^2A}{\varepsilon^2\alpha\min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s)}\log\left(\frac{1}{\delta}\right))$ and $\Omega(\frac{A}{\alpha^{H-1}\varepsilon^2}\log\left(\frac{1}{\delta}\right))$ for different instances, which demonstrate that the factor $\min\{1/\min_{\pi,h,s:\, w_{\pi,h}(s)>0} w_{\pi,h}(s),\ 1/\alpha^{H-1}\}$ is indispensable in general (see Appendix E.3 for a formal statement of lower bound).

6 WORST PATH RL

In this section, we investigate an interesting limiting case of Iterated CVaR RL when $\alpha \to 0$, called Worst Path RL, in which case the agent aims to maximize the minimum possible cumulative reward.

Worst Path RL has a *unique feature* that, the value function (Eq. (2)) concerns only the minimum value of successor states, which are independent of specific transition probabilities. Therefore, once we learn the connectivity among states, we can perform a planning to compute the optimal policy. Yet, this feature does not make the Worst Path RL problem trivial, because it is still challenging to distinguish whether a successor state is hard to reach or does not exist. As a result, a careful scheme is needed to both explore undetected successor states and exploit observations to minimize regret.

6.1 ALGORITHM MaxWP AND REGRET UPPER BOUND

We design a simple yet efficient algorithm MaxWP (Algorithm 2), which combines the exploration of undetected successor states and the exploitation of current best actions. Specifically, in episode k, MaxWP constructs empirical Q-value/value functions $\hat{Q}_h^k(s,a), \hat{V}_h^k(s)$ using the estimated lowest value of next states, and then, takes a greedy policy $\pi_h^k(s)$ with respect to $\hat{Q}_h^k(s,a)$ in this episode.

The intuition behind MaxWP is as follows. Since the Q-value function for Worst Path RL uses the $\min$ operator, if the Q-value function is not accurately estimated, it can only be over-estimated (not under-estimated). If over-estimation happens, MaxWP will be exploring an over-estimated action and urging its empirical Q-value to get back to its true Q-value. Otherwise, if the Q-value function is already accurate, MaxWP just selects the optimal action. In other words, MaxWP combines the exploration of over-estimated actions (which lead to undetected successor states) and exploitation of current best actions. Below we provide the regret guarantee for algorithm MaxWP.

Theorem 4. *With probability at least $1-\delta$, the regret of algorithm* MaxWP *is bounded by*

$$O\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{H}{\min_{\pi:\, \upsilon_\pi(s,a)>0} \upsilon_\pi(s,a) \cdot \min_{s'\in \text{supp}(p(\cdot|s,a))} p(s'|s,a)} \log\left(\frac{SA}{\delta}\right)\right),$$

where $\upsilon_\pi(s,a)$ denotes the probability (s,a) is visited at least once in an episode under policy π.

Remark 3. The factor $\min_{\pi:\, \upsilon_\pi(s,a)>0} \upsilon_\pi(s,a)$ stands for the minimum probability of visiting (s,a) at least once in an episode over all feasible policies, and $\min_{s'\in \text{supp}(p(\cdot|s,a))} p(s'|s,a)$ denotes the minimum transition probability over all successor states of (s,a). Note that this result cannot be implied by Theorem 1, because the result for Iterated CVaR RL there depends on $\frac{1}{\alpha}$, and simply taking $\alpha \to 0$ leads to a vacuous bound.

Theorem 4 demonstrates that algorithm MaxWP enjoys a constant regret with respect to K. This constant regret is made possible by the unique feature of Worst Path RL that, under the worst path metric, once the agent determines the connectivity among states, she can accurately estimate the value function and find the optimal policy. Furthermore, determining the connectivity among states (with a given confidence) only requires a number of samples independent of K. MaxWP effectively adapts to this problem feature, and efficiently explores the connectivity among states.

To validate the optimality of our regret upper bound, we also provide a lower bound $\Omega(\max_{(s,a):\exists h,\, a\neq\pi_h^*(s)} \frac{H}{\min_{\pi:\upsilon_\pi(s,a)>0} \upsilon_\pi(s,a)\cdot\min_{s'\in\text{supp}(p(\cdot|s,a))} p(s'|s,a)})$ for Worst Path RL, which demonstrates the tightness of the factors $\min_{\pi:\, \upsilon_\pi(s,a)>0} \upsilon_\pi(s,a)$ and $\min_{s'\in \text{supp}(p(\cdot|s,a))} p(s'|s,a)$.

7 CONCLUSION

In this paper, we investigate a novel Iterated CVaR RL problem with the regret minimization and best policy identification metrics. We design two efficient algorithms ICVaR-RM and ICVaR-BPI, and provide nearly matching regret/sample complexity upper and lower bounds with respect to K. We also study an interesting limiting case called Worst Path RL, and propose a simple and efficient algorithm MaxWP with rigorous regret guarantees. There are several interesting directions for future work, e.g., further closing the gap between upper and lower bounds, and extending our model and results from the tabular setting to the function approximation framework.

8 REPRODUCIBILITY STATEMENT

To make our results reproducible, we provide complete proofs for all our theoretical results in Appendix. Moreover, we also include the code and implementation instructions for our experiments in the supplementary material. One can easily reproduce our theoretical and experimental results by following our proofs and running our code.

REFERENCES

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

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(3):361–379, 2011.

Kang Boda and Jerzy A Filar. Time consistent dynamic risk measures. *Mathematical Methods of Operations Research*, 63(1):169–186, 2006.

Vivek Borkar and Rahul Jain. Risk-constrained markov decision processes. *IEEE Transactions on Automatic Control*, 59(9):2574–2579, 2014.

Vivek S Borkar. A sensitivity formula for risk-sensitive cost and the actor–critic algorithm. *Systems & Control Letters*, 44(5):339–346, 2001.

Vivek S Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2): 294–311, 2002.

David B Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3387–3395, 2019.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a CVaR optimization approach. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(1): 6070–6120, 2017.

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

Stefano P Coraluppi and Steven I Marcus. Mixed risk-neutral/minimax control of markov decision processes. In *Proceedings 31st Conference on Information Sciences and Systems*. Citeseer, 1997.

Stefano P Coraluppi and Steven I Marcus. Risk-sensitive and minimax control of discrete-time, finite-state markov decision processes. *Automatica*, 35(2):301–309, 1999.

Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109: 101964, 2020.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2818–2826, 2015.

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

Dotan Di Castro, Aviv Tamar, and Shie Mannor. Policy gradients with variance related risk criteria. *arXiv preprint arXiv:1206.6404*, 2012.

Mehdi Fatemi, Shikhar Sharma, Harm Van Seijen, and Samira Ebrahimi Kahou. Dead-ends and secure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 1873–1881. PMLR, 2019.

Mehdi Fatemi, Taylor W Killian, Jayakumar Subramanian, and Marzyeh Ghassemi. Medical dead-ends and learning to identify high-risk states and treatments. *Advances in Neural Information Processing Systems*, 34:4856–4870, 2021.

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22384–22395, 2020.

Yingjie Fei, Zhuoran Yang, Yudong Chen, and Zhaoran Wang. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021a.

Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, pp. 3198–3207. PMLR, 2021b.

Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Conference on Computational Learning Theory*, pp. 88–97, 1994.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

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

William B Haskell and Rahul Jain. A convex analytic approach to risk-aware markov decision processes. *SIAM Journal on Control and Optimization*, 53(3):1569–1598, 2015.

Matthias Heger. Consideration of risk in reinforcement learning. In *International Conference on Machine Learning*, pp. 105–111. Elsevier, 1994.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Eric Johnson and Suresh Kannan. Adaptive flight control for an autonomous unmanned helicopter. In *AIAA Guidance, Navigation, and Control Conference and Exhibit*, pp. 4439, 2002.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pp. 865–891. PMLR, 2021.

Prashanth La and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. *Advances in Neural Information Processing Systems*, 26, 2013.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021.

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

Jonathan Theodor Ott. A markov decision model for a surveillance application and risk-sensitive markov decision processes. 2010.

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

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

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

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the CVaR via sampling. In *AAAI Conference on Artificial Intelligence*, 2015.

Philip Thomas and Erik Learned-Miller. Concentration inequalities for conditional value-at-risk. In *International Conference on Machine Learning*, pp. 6225–6233. PMLR, 2019.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the $\ell 1$ deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Lu Wen, Jingliang Duan, Shengbo Eben Li, Shaobing Xu, and Huei Peng. Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization. In *IEEE International Conference on Intelligent Transportation Systems*, pp. 1–7. IEEE, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.