Regret Bounds for Markov Decision Processes with Recursive Optimized Certainty Equivalents

Wenhao Xu [1] **Xuefeng Gao** [1] **Xuedong He** [1]

Abstract

The optimized certainty equivalent (OCE) is a family of risk measures that cover important examples such as entropic risk, conditional value-at-risk and mean-variance models. In this paper, we propose a new episodic risk-sensitive reinforcement learning formulation based on tabular Markov decision processes with recursive OCEs. We design an efficient learning algorithm for this problem based on value iteration and upper confidence bound. We derive an upper bound on the regret of the proposed algorithm, and also establish a minimax lower bound. Our bounds show that the regret rate achieved by our proposed algorithm has optimal dependence on the number of episodes and the number of actions.

1. Introduction

Reinforcement learning (RL) studies the problem of sequential decision making in an unknown environment by carefully balancing between exploration and exploitation (Sutton & Barto, 2018). In the classical setting, it describes how an agent takes actions to maximize *expected cumulative rewards* in an environment typically modeled by a Markov decision process (MDP, (Puterman, 2014)). However, optimizing the expected cumulative rewards alone is often not sufficient in many practical applications such as finance, healthcare and robotics. Hence, it may be necessary to take into account of the risk preferences of the agent in the dynamic decision process. Indeed, a rich body of literature has studied *risk-sensitive* (and safe) RL, incorporating risk measures such as the entropic risk measure and conditional value-at-risk (CVaR) in the decision criterion, see, e.g., Shen et al. (2014); Garcıa & Fernández (2015); Tamar et al. (2016); Chow et al. (2017); Prashanth L & Fu (2018); Fei et al. (2020) and the references therein.

In this paper we study risk-sensitive RL for tabular MDPs with unknown transition probabilities in the *finite-horizon, episodic* setting, where an agent interacts with the MDP in episodes of a fixed length with finite state and action spaces. To incorporate risk sensitivity, we consider a broad and important class of risk measures known as Optimized Certainty Equivalent (OCE, (Ben-Tal & Teboulle, 1986; 2007)). The OCE is a (nonlinear) risk function which assigns a random variable X to a real value, and it depends on a *concave* utility function, see Equation (1) for the definition. With an appropriate choice of the *utility function*, OCE covers important examples of risk measures, including the entropic risk, CVaR and mean-variance models, as special cases, so it is popular in financial applications, such as portfolio optimization, and in the machine learning literature. See Section 2.1 for details. Using this unified framework, we aim to develop efficient learning algorithms for risk-sensitive RL with OCEs and provide worst-case regret bounds, where the regret measures the sub-optimality of the learning algorithm compared to an optimal policy should the model parameters be completely known.

We formulate a new risk-sensitive episodic RL problem with recursive OCEs. The conventional objective in risk-sensitive MDPs (when the model is known) is to optimize a static risk measure/functional applied to the (possibly discounted) *cumulative* rewards over the decision horizon (Howard & Matheson, 1972; Marcus et al., 1997). Except for the entropic risk measure, this approach typically suffers from the time-inconsistency issue, which prevents one from directly applying the dynamic programming principle (Artzner et al., 2007). In addition, the optimal policies can be non-Markovian, and are often difficult to compute (Mannor & Tsitsiklis, 2011; Du et al., 2022). In view of the time-inconsistency issue and the computational difficulty, we consider an alternative approach, which is to consider MDPs with recursive risk measures (Ruszczyński, 2010; Shen et al., 2013; Bäuerle & Glauner, 2022). In this approach, instead of optimizing a static risk measure of the cumulative rewards, one optimizes the value function defined by a recursive application of a risk measure *at each period*, which essentially replaces the expectation operator

[1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China. Correspondence to: Xuefeng Gao <xfgao@se.cuhk.edu.hk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

in the standard value iteration by the risk measure (OCEs in our setting). This approach is also partly motivated by recursive utilities in the economics literature (Kreps & Porteus, 1978; Epstein & Zin, 1989). Indeed, our recursive OCE model is a special case of the so-called dynamic mixture-averse preferences, which have an axiomatic foundation (Sarver, 2018) and is a special class of recursive utilities. The recursive structure in our OCE model implies time consistency and dynamic programming, which leads to a Bellman equation in a known environment; see for instance Bäuerle & Glauner (2022). Our formulation of episodic RL with recursive OCEs is built on this Bellman equation. Due to the generality of OCE, our RL formulation unifies and generalizes several existing episodic RL formulations in the literature, including standard risk-neutral RL (see, e.g., Azar et al. (2017)), RL with entropic risk in Fei et al. (2020), and RL with iterated CVaR in Du et al. (2022). See Section 2.2 for details.

A special case of OCE is the entropic risk measure, which is obtained by setting the utility function in OCE to be an exponential function. In this special case, the recursive OCE model is equivalent to applying the entropic risk measure to the cumulative reward over the entire decision horizon. In general, the recursive OCE model and the model of applying OCE to the cumulative reward directly are different and can be applied in different problems to account for different attitudes toward risk. The former tends to lead to more conservative actions than the latter does, because in the former the agent is concerned about risk in every step and in every state; see Du et al. (2022) for a detailed discussion and a concrete example in this regard in the context of recursive CVaR.

In this paper, we develop a model-based algorithm for risk-sensitive RL with recursive OCEs. Our algorithm is a variant of the UCBVI (Upper Confidence Bound Value Iteration) algorithm in Azar et al. (2017) for risk-neural RL. The main novelty in our algorithm design is that the bonus term used to encourage exploration depends on the *utility function* in the specific OCE that one considers. Theoretically, we prove regret bounds for our algorithm in learning MDPs with a wide family of recursive risk measures including the mean-variance criterion, by considering different utility functions in OCEs. Such bounds are new to the literature, to the best of our knowledge.

The regret analysis of algorithms for risk-sensitive RL is difficult mainly due to the nonlinearity of the objective (Fei et al., 2020). Although the structure of our regret analysis of the proposed algorithm follows the optimism principle in provably efficient risk-neutral RL (see, e.g., Azar et al. (2017); Agarwal et al. (2021)), we develop two new ingredients to overcome the difficulty in our risk-sensitive setting: (a) concentration bounds for the OCE of the next-state value function under the estimated transition distributions, and (b) a change-of-measure technique to bound the OCE of the estimated value function under the true transition distribution with an affine functional (see Equation (13)). Our concentration bounds for OCEs of value functions are different from recent results in LA & Bhat (2022) which rely on the Lipschitz continuity of the utility function and a Wasserstein distance approach. Our technique (b) is inspired by the regret analysis in Du et al. (2022) for iterated CVaR, but it is much more general and thus is applicable to OCEs. Conceptually, the main insight is to use the fact that the OCE is a concave risk functional (Ben-Tal & Teboulle, 2007, Theorem 2.1). Its (algebraic) supergradient is a linear functional (Ruszczyński & Shapiro, 2006) which turns out to be in the form of an expectation with respect to a new probability distribution that is related to the true transition distribution via change-of-measure. This linearization method is crucial in carrying out the recursions (in the time parameter) in our regret analysis. Due to change-of-measure, the corresponding Radon-Nikodym derivative naturally appears in our analysis and we need to carefully bound it.

In addition to the regret upper bound, we also establish a minimax lower bound. It shows that the regret rate achieved by our proposed algorithm has optimal dependence on the number of episodes K and the number of actions A, up to logarithmic factors. The proof of our lower bound proof is built on the hard MDP instances constructed in Domingues et al. (2021) for tabular risk-neutral RL. The main novelty in our analysis lies in modifying such hard instances to adapt to the OCEs and bounding value functions which are defined recursively via OCEs that involve an optimization problem.

1.1. Related Work

Despite rich literature in risk-sensitive RL, there are fairly limited number of studies on regret minimization in risk-sensitive MDPs. We provide a concise review below, and leave the detailed comparisons of existing regret bounds (for entropic risk and CVaR only) with our bounds to Section 4.1.

To the best of our knowledge, the first regret bound for risk-sensitive tabular MDP is due to Fei et al. (2020), who study episodic RL with the goal of maximizing the *entropic risk* of the cumulative rewards. By the pleasant properties of exponential functions in entropic risk, their RL formulation is in fact equivalent to our general (iterative) formulation when the OCE is entropic risk.

The results in Fei et al. (2020) have been improved in Fei et al. (2021a) for tabular MDPs with entropic risk, where they design two model-free algorithms with improved regret bounds. In addition, these algorithms have been extended to the function approximation setting in Fei et al. (2021b) and to time-inhomogeneous MDPs with variation budgets in Ding et al. (2022). Liang & Luo (2022) also consider

RL with the entropic risk, and they use tools from distributional RL (Bellemare et al., 2017). They propose algorithms with regret upper bounds matching the results in Fei et al. (2021a).

Du et al. (2022) propose Iterated CVaR RL, which is an episodic risk-sensitive RL formulation with the objective of maximizing the tail of the reward-to-go at each step. Their RL formulation is a special case of ours. Du et al. (2022) study both regret miminization and best policy identification, and provide matching upper and lower bounds with respect to the number of episodes.

All the aforementioned studies focus on one single risk measure (entropic risk or CVaR only) for regret analysis in risk-sensitive MDPs. Their algorithms and analysis typically rely on the properties of the special risk measure they consider. Bastani et al. (2022) study episodic RL with a class of risk-sensitive objectives known as *spectral risk measures*, which includes CVaR as an example (but not the entropic risk and mean-variance criterion). They develop an upper-confidence-bound style algorithm and obtain a regret upper bound for their algorithm. Although spectral risk measures cover CVaR as an example, their work is different from Du et al. (2022) and ours in that their objective is to optimize the (static) spectral risk of the cumulative rewards, rather than the value function obtained from iterative application of the risk measure at each time step. See Appendix A in Du et al. (2022) and Section 3 in Bastani et al. (2022) for further discussions.

Paper Organization. The rest of the paper is organized as follows: we present the problem formulation in Section 2 and describe the algorithm in Section 3. We state and discuss the main results in Section 4. We provide the proof sketch of our regret upper bound in Section 5, and conclude in Section 6. Due to space constraints, proofs and experiments are given in the Appendix.

2. Problem Formulation

In this section, we introduce the optimized certainty equivalent (OCE), and formulate the risk-sensitive reinforcement learning problem with recursive OCE.

2.1. The Optimized Certainty Equivalent

We introduce OCE, following Ben-Tal & Teboulle (2007). Let $u : \mathbb{R} \to [-\infty, \infty)$ be a nondecreasing, closed, concave utility function with effective domain dom $u = \{x \in \mathbb{R} | u(t) > -\infty\} \neq \emptyset$. Suppose u satisfies $u(0) = 0$ and $1 \in \partial u(0)$, where $\partial u(\cdot)$ denotes the superdifferential of u. We denote this class of normalized utility functions by U_0. The optimized certainty equivalent (OCE) is defined by

$$OCE^u(X) = \sup_{\lambda \in \mathbb{R}} \{\lambda + E[u(X - \lambda)]\}, \qquad (1)$$

Table 1. Popular OCEs and corresponding utility functions. For CVaR, $q(\alpha) = \min\{x | F_X(x) \geq \alpha\}$ where F_X is the cumulative distribution function of X and $[-t]_+ = \max\{-t, 0\}$.

Name	$OCE^u(X)$	Utility function u
Mean	$\mathbf{E}[X]$	$u(t) = t$
Entropic risk	$\frac{1}{\beta} \log \mathbf{E}\left[e^{\beta X}\right]$	$u_\beta(t) = \frac{1}{\beta} e^{\beta t} - \frac{1}{\beta}$
CVaR	$\mathbf{E}\left[X \mid X \leq q(\alpha)\right]$	$u_\alpha(t) = -\frac{1}{\alpha}[-t]_+$
Mean-Variance	$\mathbf{E}[X] - c \cdot \mathrm{Var}(X)$	$u_c(t) = \frac{1}{4c} 1\{t > \frac{1}{2c}\}$ $+(t - ct^2) 1\{t \leq \frac{1}{2c}\}$

where X is a bounded random variable (so that $OCE^u(X)$ is finite). The interpretation for OCE in (1) is as follows: a decision maker can consume part of the future uncertain income of X dollars at present, and this is denoted by λ. The present value of X then becomes $\lambda + E[u(X - \lambda)]$, and the OCE represents the optimal allocation of X between present and future consumption.

OCE captures the risk attitude of a decision maker via the utility function u. With different choices of the utility functions, OCE covers important examples of popular risk measures, including the entropic risk measure, CVaR and mean-variance models, as special cases. See Table 1. Due to its tractability and flexibility, OCE has been applied in many areas including finance and machine learning; see, e.g., Ben-Tal & Teboulle (2007); Lee et al. (2020); LA & Bhat (2022).

2.2. Episodic Risk-Sensitive MDPs with Recursive OCE

Consider a finite-horizon, tabular, time-inhomogeneous Markov decision process (MDP), $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, where $\mathcal{S}$ is the set of states with $|\mathcal{S}| = S$, $\mathcal{A}$ is the set of actions with $|\mathcal{A}| = A$, H is the number of steps in each episode, $\mathcal{P}$ is the transition matrix so that $P_h(\cdot|s, a)$ gives the distribution over states if action a is taken for state s at step $h \in [H]$, where $[H] = \{1, 2, \cdots, H\}$, and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function at step h. We define s_{H+1} as the terminate state, which represents the end of an episode. A policy π is a collection of H functions $\Pi := \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h \in [H]}$.

The reinforcement learning agent repeatedly interacts with the MDP $\mathcal{M} := \mathcal{M}(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ over K episodes. For simplicity (as in many prior studies (Azar et al., 2017; Du et al., 2022)) we assume that the reward function $(r_h(s, a))_{s \in \mathcal{S}, a \in \mathcal{A}}$ is known, but the transition probabilities $(P_h(\cdot|s, a))_{s \in \mathcal{S}, a \in \mathcal{A}}$ are unknown. In each episode $k = 1, 2, \cdots, K$, an arbitrary fixed initial state $s_1^k = s_1 \in \mathcal{S}$ is picked.[1] An algorithm **algo** initializes and implements

[1]The results of the paper can also be extended to the case where the initial states are drawn from a fixed distribution over $\mathcal{S}$.

a policy π^1 for the first episode, and executes policy π^k throughout episode k based on the observed past data (states, actions and rewards) up to the end of episode $k-1$, $k = 2, \cdots, K$.

To capture the (dynamic) risk in the decision making process of the agent, we propose a novel RL formulation with recursive OCEs based on the studies of MDPs with recurisve measures (Ruszczyński, 2010; Bäuerle & Glauner, 2022). Specifically, we use $V_h^\pi : \mathcal{S} \to \mathbb{R}$ to denote the value function at step h under policy π and we use $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to denote the state-action value function at step h. They are recursively defined as follows: for all $h \in [H]$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$Q_h^\pi(s,a) = r_h(s,a) + OCE^u_{s' \sim P_h(\cdot|s,a)}(V_{h+1}^\pi(s')), \quad (2)$$

$$V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)), \qquad V_{H+1}^\pi(s) = 0, \quad (3)$$

where

$$OCE^u_{s' \sim P_h(\cdot|s,a)}(g(s')) = \sup_{\lambda \in \mathbb{R}} \{\lambda + E_{s' \sim P_h(\cdot|s,a)}[u(g(s') - \lambda)]\}, \quad (4)$$

with $g : \mathcal{S} \to \mathbb{R}$ being a real-valued function.

Note that in (2)–(3), the risk measure OCE is applied to the next-state value at each period. Due to the generality of OCEs, the recursions in (2)–(3) cover and unify several existing frameworks: (a) when $u(t) = t$, the OCE becomes the mean, and (2)–(3) become the standard Bellman equation for the policy π in risk-neutral RL; (b) when u is an exponential function given in Table 1, OCE becomes the entropic risk, and (2)–(3) recover the Bellman equation for the policy π in risk-sensitive RL with entropic risk (see Equation (3) in Fei et al. (2021a)); (c) when u is a piecewise linear function and the OCE becomes the CVaR, (2)–(3) reduce to the recursion of value functions in risk-sensitive RL with iterated CVaR (see Equation (1) in Du et al. (2022)). We also remark that when the OCE is a coherent risk measure (e.g. CVaR), it has a dual or robust representation, and the recursion (2)–(3) can be interpreted as the Bellman equation of a distributionally robust MDP, see Section 6 of Bäuerle & Glauner (2022) for detailed discussions.

Because $\mathcal{S}, \mathcal{A}, H$ are finite, by Theorem 4.8 in Bäuerle & Glauner (2022), there exists an optimal Markov policy π^* which gives the optimal value function $V_h^*(s) = \max_{\pi \in \Pi} V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$. The optimal Bellman equation is given by

$$Q_h^*(s,a) = r_h(s,a) + OCE^u_{s' \sim P_h(\cdot|s,a)}(V_{h+1}^*(s')), \quad (5)$$

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a), \qquad V_{H+1}^*(s) = 0. \quad (6)$$

The expected (total) regret for algorithm **algo** over K episodes of interaction with the MDP $\mathcal{M}$ is then defined as

$$Regret(\mathcal{M}, \textbf{algo}, K) = E\left[\sum_{k=1}^K (V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k))\right], \quad (7)$$

where the term $V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)$ measures the performance loss when the agent executes (suboptimal) policy π^k in episode k. Our goal is to propose an efficient learning algorithm with a provable worst-case regret upper bound that scales sublinearly in K, as well as to establish a minimax lower bound.

3. The OCE-VI Algorithm

In this section, we propose a model-based algorithm, denoted by OCE-VI, for risk-sensitive RL with recursive OCE.

Before presenting the algorithm, we first introduce some notations. A state-action-state triplet (s, a, s') means that the process is in state s, takes an action a and then moves to state s'. Similarly, a state-action pair (s, a) means that the process is in state s and takes an action a. At the beginning of the k-th episode, we set the observed cumulated visit counts to (s, a, s') at step h up to the end of episode $k-1$ as $N_h^k(s, a, s')$ for $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, and the cumulated visit counts to (s, a) at step h up to the end of episode $k-1$ as $N_h^k(s, a)$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$. When $2 \le k \le K$, for $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$, the formulas for $N_h^k(s, a, s')$ and $N_h^k(s, a)$ are given by

$$N_h^k(s,a,s') = \sum_{i=1}^{k-1} 1\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\},$$
$$N_h^k(s,a) = \sum_{i=1}^{k-1} 1\{(s_h^i, a_h^i) = (s,a)\}. \quad (8)$$

When $k = 1$, we set $N_h^k(s, a, s') = N_h^k(s, a) = 0$ for $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$. Then the empirical transition probabilities are given by

$$\hat{P}_h^k(s'|s,a) = \frac{N_h^k(s,a,s')}{\max\{1, N_h^k(s,a)\}}. \quad (9)$$

In particular, if (s, a) has not been sampled before episode k, $\hat{P}_h^k(s'|s, a) = 0$ for all s'.

Similar to UCBVI in Azar et al. (2017) for risk-neutral RL, the OCE-VI algorithm achieves exploration by awarding some bonus for exploring some state-action pairs during the learning process. We consider the bonus

$$b_h^k(s,a) = |u(-H+h)| \sqrt{\frac{2\log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N_h^k(s,a)\}}}, \quad (10)$$

where $(s, a) \in \mathcal{S} \times \mathcal{A}$, and $\delta \in (0, 1)$ is an input parameter in our algorithm. The details of the OCE-VI algorithm are summarized in Algorithm 1.

Algorithm 1 The OCE-VI Algorithm

Input: Parameters $\delta, \mathcal{S}, \mathcal{A}, H, K, r$ and an utility function $u \in U_0$
for episode $k = 1, \cdots, K$ **do**
  Estimate $\hat{P}_h^k(s'|s,a)$ by (8) and (9) for all $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$
  Initialize $\hat{V}_{H+1}^k(s) \leftarrow 0$ for all $s \in \mathcal{S}$
  for step $h = H, H-1, \cdots, 1$ **do**
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      if $N_h^k(s,a) \geq 1$ **then**
        Compute $b_h^k(s,a)$ by (10) according to the utility function u
        $\hat{Q}_h^k(s,a) \leftarrow \min\{r_h(s,a) + OCE^u_{s' \sim \hat{P}_h^k(\cdot|s,a)}(\hat{V}_{h+1}^k(s')) + b_h^k(s,a), H-h+1\}$
      else
        $\hat{Q}_h^k(s,a) \leftarrow H-h+1$
      end if
      $\hat{V}_h^k(s) \leftarrow \max_{a' \in \mathcal{A}} \hat{Q}_h^k(s,a')$, $\pi_h^k(s) \leftarrow \arg\max_{a' \in \mathcal{A}} \hat{Q}_h^k(s,a')$
    end for
  end for
  Apply policy π^k throughout episode k
end for

Remark 3.1. The dependence of the bonus (10) on the utility function u sheds some light on how the degree of risk aversion affects the degree of exploration. Ben-Tal & Teboulle (2007) show that an agent with OCE preferences is weakly risk averse (i.e., any random payoff is less preferred by the agent to its mean) if and only if the utility function is dominated by the identity function (i.e., $u(x) \leq x, x \in \mathbb{R}$). Now, consider two agents with recursive OCE preferences represented by utility functions u_1 and u_2, respectively. If u_1 is dominated by u_2 (i.e., $u_1(x) \leq u_2(x), x \in \mathbb{R}$), then $|u_1(-H+h)| \geq |u_2(-H+h)|$ because $-H+h \leq 0$ and $u_i(x) \leq 0, x \leq 0, i = 1, 2$. Consequently, the exploration bonus for agent 1 is larger than that for agent 2. Therefore, if we interpret the dominance of u_2 over u_1 as a higher degree of risk aversion of agent 1 than that of agent 2, as suggested by the characterization of weak risk aversion Ben-Tal & Teboulle (2007), then in our algorithm for a more risk averse agent we need to have a larger bonus to encourage her to explore. We also remark that our bonus (10) is based on Chernoff-Hoeffding's concentration inequalities and it scales linearly with $|u(-H+h)|$. It might be possible to design tighter bonuses that may depend on the utility function in a nonlinear manner. This is an open problem for risk-sensitive RL with recursive OCE and we leave it for future work.

Remark 3.2. The OCE-VI algorithm is computationally tractable. In each episode, the computational cost of the algorithm is similar to solving a known MDP with value iteration, except that one needs to to compute the quantity $OCE^u_{s' \sim \hat{P}_h^k(\cdot|s,a)}(\hat{V}_{h+1}^k(s'))$ when updating the Q function. For certain special utility functions such as those in Table 1, this quantity can be explicitly computed because the state space is finite. In general, computing this OCE is equivalent to solving the optimization problem $\sup_{\lambda \in \mathbb{R}}\{\lambda + E_{s' \sim \hat{P}_h^k(\cdot|s,a)}[u(\hat{V}_{h+1}^k(s') - \lambda)]\}$. This is a one-dimensional concave optimization problem because the utility function u is concave and $\hat{P}_h^k(\cdot|s,a)$ is a probability distribution. Because the state space is finite, we can exchange the expectation and the derivative/supergradient with respect to λ in the first order optimality condition of the above optimization problem. Thus, when the utility function is differentiable, this concave optimization problem can be solved efficiently using the gradient descent or Newton's method. When the utility function is nondifferentiable, it can be solved with efficient proximal gradient methods; see, e.g., Parikh et al. (2014).

4. Main Results

In this section, we present our main results. Our first main result is an upper bound on the expected regret of the proposed OCE-VI algorithm.

Theorem 4.1. *The expected regret of the OCE-VI algorithm satisfies*

$$Regret(\mathcal{M}, \textbf{\textit{OCE-VI}}, K) \leq \tilde{\mathcal{O}}\left(\sum_{h=1}^{H} |u(-H+h)| S \sqrt{\prod_{i=1}^{h-1} u'_{-}(-H+i) AK}\right),$$

where $\tilde{\mathcal{O}}(\cdot)$ ignores the logarithmic factors in S, A, H and K and $u'_{-}(\cdot)$ is the left derivative of u.

The regret upper bound depends on the utility function u in the OCE (1) via the term $|u(-H+h)|$, which comes from the bonus (10), and the term $\prod_{i=1}^{h-1} u'_{-}(-H+i)$, which comes from bounding the Radon-Nikodym derivative arising from the linearization of the OCE as a concave functional in our regret analysis (see Equation (16)). We provide a sketch of the proof of Theorem 4.1 in Section 5, and give the full details in Appendix B.

We next present our second main result, which provides a minimax regret lower bound for RL with recursive OCE. We first state the following assumption.

Assumption 4.2. The number of states and actions satisfy $S \geq 6, A \geq 2$, and there exists an integer d such that $S = 3 + \frac{A^d - 1}{A-1}$. In addition, the horizon H satisfies $H \geq c_2 d$, where $c_2 > 2$ is a constant.

Assumption 4.2 is adapted from Assumption 1 in Domingues et al. (2021), who provide a minimax lower bound in the risk-neutral episodic RL setting. This assumption is imposed to simplify the analysis, more precisely the construction of hard MDP instances, and it can be relaxed following the discussion in Appendix D of Domingues et al. (2021).

Theorem 4.3. *Under Assumption 4.2, for any algorithm **algo**, there exists an MDP $\mathcal{M}$ whose transition probabilities depend on h such that*

$$Regret(\mathcal{M}, \textbf{\textit{algo}}, K) \geq \frac{\sqrt{SAHK}}{18\sqrt{c_1 c_2}} \cdot \left[u\left(\left(1 - \frac{2}{c_2}\right) H - \lambda^* \right) - u(-\lambda^*) \right]$$

for all $K \geq \frac{c_1 HSA}{2c_2}$, where the constants $c_1 \geq 4, c_2 > 2$ and λ^* satisfies

$$1 \in \left(1 - \frac{2}{c_1}\right) \partial u \left(\left(1 - \frac{2}{c_2}\right) H - \lambda^* \right) + \frac{2}{c_1} \partial u(-\lambda^*).$$

Note that when $u(t) = t$, OCE becomes expectation, and our regret lower bound in Theorem 4.3 is $\Omega(H\sqrt{SAHK})$, by choosing for instance $c_1 = 4$ and $c_2 = 3$. This recovers the (tight) regret lower bound in Domingues et al. (2021) in learning risk-neutral tabular MDP. For a general utility function u in OCE, the choices of constants $c_1 \geq 4, c_2 > 2$ should be based on the specific utility function to generate tighter lower bounds. For illustrations, we provide some examples in Section 4.1.

The proof of Theorem 4.3 is based on extending the proof of Theorem 9 in Domingues et al. (2021) to our risk-sensitive setting. There are essential difficulties in this extension. These include how to construct hard MDP instances that adapt to the OCE, and how to bound the value functions defined recursively via OCE that involves an optimization problem. Due to space limitations, we provide the proof details in Appendix C.

Remark 4.4. For the simplicity of presentation, we focus on OCE in (1), which exhibits the risk aversion property with $OCE^u(X) \leq E[X]$, due to the concavity of the utility function; see Proposition 2.2 in Ben-Tal & Teboulle (2007). Our main results in the paper hold in the risk-seeking setting as well, where $OCE^u(X)$ is defined by $\inf_{\lambda \in \mathbb{R}} \{\lambda + E[u(X - \lambda)]\}$ with a convex utility function u. In this case, we need to use a bonus $b_h^k(s,a) = |u(-H+h)| \sqrt{\frac{2S \log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N_h^k(s,a)\}}}$ in the OCE-VI algorithm. Compared with (10), this bonus has an extra term $\sqrt{S}$, which arises from a technical step in the proof for the risk-seeking case (see inequality (2) of Lemma B.2). The regret bounds still hold in this setting.

4.1. Examples and Comparisons to Related Work

We consider several specific utility functions and the resulting OCEs to illustrate our regret bounds in Theorems 4.1 and 4.3.

4.1.1. Mean-variance Model

When the utility function is $u_c(t) = (t - ct^2) 1\{t \leq \frac{1}{2c}\} + \frac{1}{4c} 1\{t > \frac{1}{2c}\}$, the corresponding OCE is the celebrated mean-variance model (Markowitz, 1952), where $c > 0$ is a given risk parameter representing the degree of risk aversion. To the best of our knowledge, the following results are the first regret bounds for risk-sensitive MDPs with the recursive mean-variance model.

- Upper bound. Our regret upper bound in Theorem 4.1 is $\tilde{\mathcal{O}}\left((1+2cH)^{\frac{H-1}{2}} (H^2 + cH^3) S\sqrt{AK}\right)$.
- Lower bound. We can choose $c_1 = 8, c_2 = 4$, and then $\lambda^* = \left(1 - \frac{2}{c_1}\right)\left(1 - \frac{2}{c_2}\right) H = 3H/8$. The regret lower bound in Theorem 4.3 becomes $\Omega\left((H + \frac{1}{4}cH^2)\sqrt{SAHK}\right)$.

4.1.2. (Iterated) CVaR

When the utility function is $u_\alpha(t) = -\frac{1}{\alpha}[-t]_+, \alpha > 0$, the corresponding OCE is CVaR, where $\alpha > 0$ is the risk level of CVaR. Our RL formulation in Section 2.2 reduces to the one in Du et al. (2022), and our OCE-VI algorithm becomes their ICVaR algorithm with a smaller exploration bonus.

- Upper bound. Our regret upper bound in Theorem 4.1 becomes $\tilde{\mathcal{O}}\left(\frac{(\frac{1}{\sqrt{\alpha}})^H - 1 - H(\frac{1}{\sqrt{\alpha}} - 1)}{(1-\sqrt{\alpha})^2} S\sqrt{AK}\right)$. When $0 < \alpha \leq \frac{3-\sqrt{5}}{2}$, this upper bound can be further bounded by $\tilde{\mathcal{O}}\left(\left(\frac{1}{\sqrt{\alpha^{H+1}}} - \frac{H}{\sqrt{\alpha}}\right) S\sqrt{AK}\right)$. When $\frac{3-\sqrt{5}}{2} < \alpha < 1$, the regret bound can be further bounded by $\tilde{\mathcal{O}}\left(\frac{H^2 S\sqrt{AK}}{\sqrt{\alpha^{H+1}}}\right)$. Du et al. (2022) design the ICVaR algorithm and can obtain a worst-case regret upper bound of $\tilde{\mathcal{O}}\left(\frac{H^2 S\sqrt{AK}}{\sqrt{\alpha^{H+1}}}\right)$.[2] Our result improves the result of Du et al. (2022) by a factor of H^2 when $0 < \alpha \leq \frac{3-\sqrt{5}}{2}$. This is due to a smaller exploration bonus used in our algorithm compared with theirs.

- Lower bound. We can choose $c_1 = \frac{2}{\alpha}$ and $c_2 = 4$ in Theorem 4.3, and let $\lambda^* = \left(1 - \frac{3}{c_2}\right) H$. Then, our regret lower bound becomes $\Omega\left(H\sqrt{\frac{SAHK}{\alpha}}\right)$ and it is problem-independent. This is in contrast with Du et al. (2022), who derive a regret lower bound that depends on some problem-dependent quantity, specifically, the minimum probability of visiting an available state under any feasible policy.

4.1.3. ENTROPIC RISK

When the utility function is $u_\beta(t) = \frac{1}{\beta}e^{\beta t} - \frac{1}{\beta}, \beta < 0$, the corresponding OCE is entropic risk, where $\beta < 0$ is a given risk parameter representing the degree of risk aversion. In this case, our RL formulation in Section 2.2 is equivalent to the one in Fei et al. (2020). Note, however, that our OCE-VI algorithm is model-based and is different from the model-free algorithms proposed in Fei et al. (2020).

- Upper bound. Our regret upper bound in Theorem 4.1 for the OCE-VI algorithm is $\tilde{\mathcal{O}}\left(\exp(-\frac{\beta H^2}{4})\frac{\exp(-\beta H)-1}{-\beta}S\sqrt{AK}\right)$. This bound has a factor that is exponential in $|\beta|H^2$, which is similar as the bounds in Fei et al. (2020). Recently, Fei et al. (2021a) propose the RSVI2 and RSQ2 algorithms, and they manage to remove this factor. Their algorithms are based on the nice properties of the exponential utility, in particular, the so-called exponential Bellman equation which takes the exponential on both sides of the Bellman equation in Fei et al. (2020). However, such techniques can not be applied to our general setting, because general utility functions do not possess the same nice properties as the exponential function. Even though our upper bound is worse than the one in Fei et al. (2021a), we show numerically that our algorithm can outperform their algorithms on randomly generated MDP instances. See Figure 1 for an example and Appendix D for experimental details.
- Lower Bound. We can choose $c_1 = \frac{2}{e^{-\beta}-1} \cdot \exp\left(-\beta\left(1-\frac{2}{c_2}\right)H\right)$ and $c_2 = 6$ in Theorem 4.3, and the corresponding $\lambda^* = \frac{1}{\beta}\log\left(\left(1-\frac{2}{c_1}\right)\exp\left(\beta\left(1-\frac{2}{c_2}\right)H\right) + \frac{2}{c_1}\right)$. Then our regret lower bound becomes $\Omega\left(\frac{\exp(-\frac{1}{3}\beta H)-1}{-\beta}\sqrt{SAHK}\right)$. By contrast, Fei et al. (2020) derive a regret lower bound of $\Omega\left(\frac{\exp(\frac{1}{2}|\beta|H)-1}{|\beta|}\sqrt{K}\right)$, which does not depend on S or A (due to the simple structure of the hard instances they construct). Liang & Luo (2022) derive a regret lower bound $\Omega\left(\frac{\exp(\frac{1}{6}|\beta|H)-1}{|\beta|}\sqrt{SAHK}\right)$ in the risk-seeking setting when $\beta > 0$, but they mention that it is unclear whether a similar bound holds in the risk-averse setting when $\beta < 0$; see page 30 of their paper.

[2] Du et al. (2022) consider time-homogeneous MDPs, and we modify their regret bounds to adapt to our time-inhomogeneous setting.

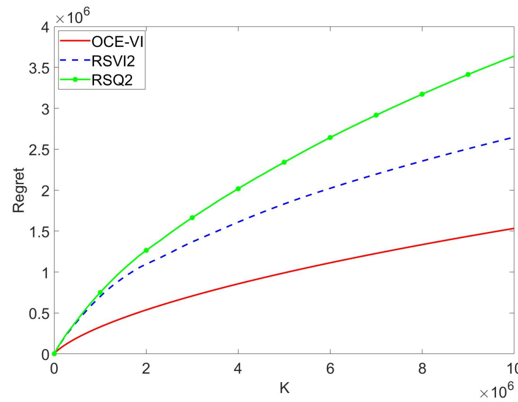


Figure 1. Performance comparison of the OCE-VI algorithm with RSVI2 and RSQ2 algorithms in Fei et al. (2021a) on a randomly generated MDP with $(H, S, A) = (6, 20, 3)$ and the entropic risk objective.

4.2. Discussions on tightness of our regret bounds

Theorems 4.1 and 4.3 imply that the OCE-VI algorithm achieves a regret rate with the optimal dependence on the number of episodes K and the number of actions A, up to logarithmic factors. While the bounds on K are the most important as they imply the convergence rates of learning algorithms, it remains an important open question whether one can improve the dependence of these bounds on H and S to narrow down the gap between the upper and lower bounds in the risk-sensitive RL setting. We elaborate further on this issue below.

From Theorems 4.1 and 4.3, we can see that the gap between our upper and lower bounds in terms of S is $\sqrt{S}$, where S is the number of states. The extra $\sqrt{S}$ in our regret upper bound arises from a step in our proof where we apply an L^1 concentration bound for the S-dimensional empirical transition probability vector, see Equation (12) in Section 5. This extra $\sqrt{S}$ factor can be removed in RL for *risk-neutral* MDPs by directly maintaining confidence intervals on the optimal value function; see, e.g., Azar et al. (2017); Zanette

& Brunskill (2019). However, it is not clear how to adapt this technique to our risk-sensitive setting, i.e., remove $\sqrt{S}$ in (12). This is primarily because the estimated value functions $\hat{V}_h^k$ in our algorithm are not only random, but they also involve OCE which is nonlinear and defined by an optimization problem (so the optimizer is also random).

There is an exponential gap in terms of H between our upper and lower bounds. This gap is due to the linearization of OCE in the recursive procedure of the regret analysis of our algorithm. Indeed, if $u(t) = t$, the corresponding regret upper bound is $\tilde{\mathcal{O}}(H^2 S\sqrt{AK})$, which does not have the exponential term of H. In the risk-neutral setting, one can improve the dependence of the upper bound on H by considering Bernstein-style exploration bonus which is built from the empirical variance of the estimated value function $\hat{V}_{h+1}^k$ at the next state, see, e.g., Azar et al. (2017). However, it is still an open problem how to use Bernstein-type bonus to improve the regret bound in risk-sensitive RL (Fei et al., 2021a; Du et al., 2022). In our RL setting with recursive OCEs, it is possible to design a Bernstein-type bonus, but it may not lead to improved regret bounds, at least within our current analysis framework. We provide some informal discussions below including the challenges in improving bounds.

First, to ensure optimism with Bernstein-type (or variance-related) bonuses, we need analogous results to Lemmas B.1 and B.2 in the appendix, which provide concentration bounds for OCEs of next-state value functions. Using Bernstein inequality instead of Hoeffding inequality, the confidence bound in Lemma B.1 becomes

$$\sqrt{\frac{2\,\mathrm{Var}_{s'\sim P_h(\cdot|s,a)}\left(u(V_{h+1}^*(s')-\lambda_{h+1}^*)\right)\log\left(\frac{SAHK}{\delta}\right)}{N_h^k(s,a)}}$$

$+$ lower order term.

This bound allows us to design a Bernstein-type bonus $b_h^k(s,a)$ in the form of

$$\underbrace{2\sqrt{\frac{\mathrm{Var}_{s'\sim \hat{P}_h^k(\cdot|s,a)}\left(u(\hat{V}_{h+1}^k(s')-\hat{\lambda}_{h+1}^k)\right)\log\left(\frac{SAHK}{\delta}\right)}{N_h^k(s,a)}}}_{\text{main term}}$$

$+$ lower order term.

Compared with the Bernstein bonus in the risk-neutral RL setting (see e.g. Azar et al. (2017)), $\hat{V}_{h+1}^k(s')$ inside the variance operator is replaced by $u(\hat{V}_{h+1}^k(s')-\hat{\lambda}_{h+1}^k)$ in our risk-sensitive RL setting. We use this approach because the OCE involves an optimization problem and we need to 'linearize' it (i.e., remove the sup in the definition of OCE) and work with the utility u applied to the value function first in order to derive concentration bounds for OCEs. With this new bonus, we might be able to get the same regret bound as the one presented in the current paper.

However, it is difficult to get improved bounds as we explain below. In the risk-neutral setting, Azar et al. (2017) use an iterative-Bellman-type-Law of Total Variance so that the sum of the variances of V_{h+1}^* over H steps is bounded by the variance of the sum of H-step rewards; see Equation (26) in Azar et al. (2017) and Lemma C.5 in Jin et al. (2018) for a proof of this result. This is a key technical result in obtaining improved bounds in H. However, this result does not hold in our setting for two reasons: first, our value function is not the expected sum of H-step rewards; second, while the value V_{h+1}^* satisfies a Bellman recursion, the quantity $u(V_{h+1}^*(s')-\lambda_{h+1}^*)$ (that appears in the variance operator) does not. Therefore, we may still have to use the crude bound for the variance term in the Bernstein-type bonus by using a maximum bound for $u(V_{h+1}^*(s')-\lambda_{h+1}^*)$. This leads to the same bound as in our current paper and we do not obtain improvements in the regret with respect to H.

5. Proof Sketch of Theorem 4.1

The structure of the proof of Theorem 4.1 follows the optimism principle in provably efficient risk-neutral RL (see, e.g., Agarwal et al. (2021, Chapter 7)), however, we provide two new ingredients in our analysis: (a) concentration bounds for the OCE of the next-state value function under estimated transitions (see (11) and (12)), and (b) a change-of-measure technique to bound the OCE of the estimated value function (under the true transition) with an affine function (see (13)), and bound the the Radon-Nikodym derivative (see (16)). For notational simplicity, we use P_h to denote $P_h(s_{h+1}^k|s_h^k,a_h^k)$ and use $\hat{P}_h^k$ to denote $\hat{P}_h^k(s_{h+1}^k|s_h^k,a_h^k)$ when there is no ambiguity.

Step 1: Optimism. We can first show optimism, i.e., the event $\hat{V}_h^k \geq V_h^*$ for all h,k holds with a high probability, where $\hat{V}_h^k$ is the estimated value function in our algorithm in episode k. This step relies on a concentration bound of the OCE of the optimal value function under the estimated transitions $\hat{P}_h^k$: with probability $1-\delta$ (where $\delta\in(0,1)$),

$$OCE_{P_h}^u(V_{h+1}^*) - OCE_{\hat{P}_h^k}^u(V_{h+1}^*) \leq b_h^k. \tag{11}$$

This bound can be proved by using the representation of the OCE in (1), together with similar martingale arguments used in the risk-neutral RL setting (Agarwal et al., 2021, Lemma 7.3). By optimism, the regret in (7) is upper bounded by $E[\sum_{k=1}^K(\hat{V}_1^k - V_1^{\pi^k})]$.

Step 2: Bounding $\hat{V}_h^k - V_h^{\pi^k}, \forall k, h$. By definition,

$$\begin{aligned}\hat{V}_h^k - V_h^{\pi^k} &\leq b_h^k + OCE_{\hat{P}_h^k}^u(\hat{V}_{h+1}^k) - OCE_{P_h}^u(V_{h+1}^{\pi^k})\\ &= b_h^k + \left[OCE_{\hat{P}_h^k}^u(\hat{V}_{h+1}^k) - OCE_{P_h}^u(\hat{V}_{h+1}^k)\right]\\ &\quad + \left[OCE_{P_h}^u(\hat{V}_{h+1}^k) - OCE_{P_h}^u(V_{h+1}^{\pi^k})\right].\end{aligned}$$

Step 2.1: The second term in the above equation can be bounded by using a concentration result for the OCE of the estimated value function $\hat{V}_{h+1}^k$: with probability $1-\delta$,

$$OCE^u_{\hat{P}_h^k}(\hat{V}_{h+1}^k) - OCE^u_{P_h}(\hat{V}_{h+1}^k) \le \sqrt{S} \cdot b_h^k. \tag{12}$$

The extra $\sqrt{S}$ factor, compared with (11), is because both $\hat{V}_{h+1}^k$ and $\hat{P}_h^k$ are random and we use L^1 concentration bounds for $||\hat{P}_h^k - P_h||_1$ as in Jaksch et al. (2010).

Step 2.2: The third term $OCE^u_{P_h}(\hat{V}_{h+1}^k) - OCE^u_{P_h}(V_{h+1}^{\pi^k})$ is more difficult to bound. Because the OCE is a concave nonlinear functional, we expect that $OCE^u_{P_h}(\hat{V}_{h+1}^k) - OCE^u_{P_h}(V_{h+1}^{\pi^k}) \le \ell(\hat{V}_{h+1}^k - V_{h+1}^{\pi^k})$, where $\ell(\cdot)$ is a linear function of random variables and it is a supergradient of the OCE. We actually show that ℓ can be represented in the form of an expectation:

$$OCE^u_{P_h}(\hat{V}_{h+1}^k) - OCE^u_{P_h}(V_{h+1}^{\pi^k}) \le E_{B_h}(\hat{V}_{h+1}^k - V_{h+1}^{\pi^k}), \tag{13}$$

where the expectation $E_{B_h}[\cdot]$ is taken with respect to a new probability distribution B_h that is linked to the true transition distribution P_h by change-of-measure. Specifically, using the first order optimality condition for $OCE^u_{P_h}(V_{h+1}^{\pi^k})$ as a concave optimization problem, we have $1 \in E_{s' \sim P_h}[\partial u(V_{h+1}^{\pi^k}(s') - \lambda_{h+1}^k)]$ where λ_{h+1}^k is an optimal solution. We can find a supergradient $\Lambda_{h+1}^k(s') \in \partial u(V_{h+1}^{\pi^k}(s') - \lambda_{h+1}^k)$ that satisfies $E_{P_h}[\Lambda_{h+1}^k] = 1$, and define the new distribution B_h by

$$B_h(s'|s,a) = P_h(s'|s,a)\Lambda_{h+1}^k(s'), \quad \forall s' \in \mathcal{S}.$$

Here, Λ_{h+1}^k is the Radon-Nikodym derivative.

By combining Steps 2.1 and 2.2, we obtain

$$\hat{V}_h^k - V_h^{\pi^k} \le 2\sqrt{S} \cdot b_h^k + E_{B_h}[\hat{V}_{h+1}^k - V_{h+1}^{\pi^k}]. \tag{14}$$

Step 3: Bounding the regret. Applying (14) recursively over h and using (10), we have that with probability $1-2\delta$,

$$\sum_{k=1}^K \left(\hat{V}_1^k - V_1^{\pi^k}\right) \tag{15}$$
$$\le \sum_{h=1}^H \sum_{k=1}^K E_{w_{hk}^B}\left[2\sqrt{2}|u(-H+h)|\sqrt{\frac{S\log\left(\frac{SAHK}{\delta}\right)}{N_h^k}}\right],$$

where w_{hk}^B is the probability of π^k visiting (s_h^k, a_h^k) at step h starting from s_1^k under probability measures $B_i(\cdot|s_i^k, a_i^k), i = 1, \cdots, h-1$. Specifically,

$$E_{w_{hk}^B}[\cdot] := \begin{cases} 1 & h = 1, \\ E_{B_1}\left[E_{B_2}\left[\cdots E_{B_{h-1}}[\cdot]\right]\right] & h \ge 2. \end{cases}$$

The main difficulty in bounding $E[\sum_{k=1}^K(\hat{V}_1^k - V_1^{\pi^k})]$ is that w_{hk}^B is built upon the probability measure B_h for any $k \in [K], h \in [H]$ while we have to take expectation under probability measure P_h outside the summation over $k \in [K]$. To address this issue, we first note that $E_{w_{hk}^B}\left[\frac{1}{\sqrt{N_h^k}}\right] = E\left[\Lambda_2^k \cdots \Lambda_h^k \frac{1}{\sqrt{N_h^k}} \middle| s_1^k, a_1^k\right]$, which implies

$$E\left[\sum_{k=1}^K E_{w_{hk}^B}\left[\frac{1}{\sqrt{N_h^k}}\right]\right] = \sum_{k=1}^K E\left[\Lambda_2^k \cdots \Lambda_h^k \frac{1}{\sqrt{N_h^k}}\right].$$

Using Cauchy–Schwarz inequality, this term can be upper bounded by $\sqrt{\sum_{k=1}^K E[\Lambda_2^k \cdots \Lambda_h^k]^2 \cdot \sum_{k=1}^K E\left[\frac{1}{N_h^k}\right]}$. It is well-known that $\sum_{k=1}^K \frac{1}{N_h^k} \le SA\log(3K)$ (Azar et al., 2017). One can show that $E[\Lambda_2^k \cdots \Lambda_h^k] = 1$ and $0 \le \Lambda_{i+1}^k \le u'_-(-H+i)$. Then we have

$$\sum_{k=1}^K E[\Lambda_2^k \cdots \Lambda_h^k]^2 \le \prod_{i=1}^{h-1} u'_-(-H+i)K. \tag{16}$$

Summing over h, choosing $\delta = \frac{1}{2KH}$ and applying a standard argument (see, e.g., Chapter 7.3 of Agarwal et al. (2021)), we obtain the bound in Theorem 4.1.

6. Conclusion and Future Work

In this paper we have proposed a risk-sensitive RL formulation based on episodic finite MDPs with recursive OCEs. We develop a learning algorithm, OCE-VI, and establish a worst-case regret upper bound. We also prove a regret lower bound, showing that the regret rate achieved by our proposed algorithm actually has the optimal dependence on the numbers of episodes and actions. Because OCEs encompass a wide family of risk measures, our paper generates new regret bounds for episodic risk-sensitive RL problems with those risk measures.

Regret minimization for risk-sensitive MDPs is still largely unexplored. For future work, one important direction is to improve regret bounds in the number of states and the horizon length. Other interesting directions include, to name a few, studying large or continuous state/action spaces, considering risk measures other than OCEs, and obtaining problem-dependent regret bounds.

Acknowledgments

We thank the area chair and anonymous referees for many constructive comments and suggestions. Xuefeng Gao and Xuedong He acknowledges support from the Hong Kong Research Grants Council [Grants 14201520, 14201421, 14212522, 14205718 and 14218022].

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. https://rltheorybook.github.io/rltheorybook_AJKS.pdf, 2021.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., and Ku, H. Coherent multiperiod risk adjusted values and bellman's principle. *Annals of Operations Research*, 152(1):5–22, 2007.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Bastani, O., Ma, Y. J., Shen, E., and Xu, W. Regret bounds for risk-sensitive reinforcement learning. *arXiv preprint arXiv:2210.05650*, 2022.

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

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.

Ben-Tal, A. and Teboulle, M. Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science*, 32(11):1445–1466, 1986.

Ben-Tal, A. and Teboulle, M. An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Dann, C. *Strategic Exploration in Reinforcement Learning-New Algorithms and Learning Guarantees*. PhD thesis, Carnegie Mellon University, 2019.

Ding, Y., Jin, M., and Lavaei, J. Non-stationary risk-sensitive reinforcement learning: Near-optimal dynamic regret, adaptive detection, and separation design. *arXiv preprint arXiv:2211.10815*, 2022.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Du, Y., Wang, S., and Huang, L. Risk-sensitive reinforcement learning: Iterated cvar and the worst path. *arXiv preprint arXiv:2206.02678*, 2022.

Epstein, L. G. and Zin, S. E. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework. *Econometrica*, 57:937–969, 1989.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Fei, Y., Yang, Z., Chen, Y., and Wang, Z. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20436–20446, 2021a.

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

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

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Howard, R. A. and Matheson, J. E. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. URL http://jmlr.org/papers/v11/jaksch10a.html.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Kreps, D. M. and Porteus, E. L. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica: journal of the Econometric Society*, pp. 185–200, 1978.

LA, P. and Bhat, S. P. A wasserstein distance approach for concentration of empirical risk estimates. *Journal of Machine Learning Research*, 23:1–61, 2022.

Lee, J., Park, S., and Shin, J. Learning bounds for risk-sensitive learning. *Advances in Neural Information Processing Systems*, 33:13867–13879, 2020.

Liang, H. and Luo, Z.-Q. Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds. *arXiv preprint arXiv:2210.14051*, 2022.

Mannor, S. and Tsitsiklis, J. N. Mean-variance optimization in markov decision processes. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 177–184, 2011.

Marcus, S. I., Fernández-Gaucherand, E., Hernández-Hernandez, D., Coraluppi, S., and Fard, P. Risk sensitive markov decision processes. In *Systems and control in the twenty-first century*, pp. 263–279. Springer, 1997.

Markowitz, H. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Parikh, N., Boyd, S., et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

Prashanth L, A. and Fu, M. Risk-sensitive reinforcement learning. *arXiv e-prints*, pp. arXiv:1810.09126, 2018.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Ruszczyński, A. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

Ruszczyński, A. and Shapiro, A. Optimization of convex risk functions. *Mathematics of operations research*, 31 (3):433–452, 2006.

Sarver, T. Dynamic mixture-averse preferences. *Econometrica*, 86(4):1347–1382, 2018.

Shen, Y., Stannat, W., and Obermayer, K. Risk-sensitive markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

Shen, Y., Tobia, M. J., Sommer, T., and Obermayer, K. Risk-sensitive reinforcement learning. *Neural computation*, 26 (7):1298–1328, 2014.

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

Tamar, A., Di Castro, D., and Mannor, S. Learning the variance of the reward-to-go. *The Journal of Machine Learning Research*, 17(1):361–396, 2016.

Zanette, A. and Brunskill, E. 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.

A. Preliminary Lemmas

In this section, we present some preliminary lemmas that will be used in the proofs of Theorems 4.1 and 4.3.

Lemma A.1 is Ben-Tal & Teboulle (2007, Theorem 2.1) and it summarizes some fundamental properties of OCE.

Lemma A.1. *(Main Properties of OCE) For any utility function $u \in U_0$, and any bounded random variable X the following properties hold:*
(a)Shift Additivity. $OCE^u(X + c) = OCE^u(X) + c, \forall c \in \mathbb{R}$.
(b)Consistency. $OCE^u(c) = c$, for any constant $c \in \mathbb{R}$.
(c)Monotonicity. Let Y be any random variable such that $X(w) \leq Y(w), \forall w \in \Omega$. Then,

$$OCE^u(X) \leq OCE^u(Y).$$

(d)Concavity. For any random variables X_1, X_2 and any $\mu \in (0, 1)$, we have

$$OCE^u(\mu X_1 + (1-\mu)X_2) \geq \mu OCE^u(X_1) + (1-\mu)OCE^u(X_2).$$

The following lemma provides preliminary bounds for the value functions (see (2) and (3)) and the regret of learning algorithms.

Lemma A.2. *For any $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], \pi \in \Pi$ and $u \in U_0$, we have $Q_h^\pi(s,a) \in [0, H-h+1]$ and $V_h^\pi(s) \in [0, H-h+1]$. Consequently, for each $K \geq 1$, we have $0 \leq Regret(\mathcal{M}, \textbf{algo}, K) \leq KH$ for any **algo**.*

Proof. Recall that $Q_h^\pi(s,a) = r_h(s,a) + OCE^u_{s'\sim P_h(\cdot|s,a)}(V^\pi_{h+1}(s'))$ and $V^\pi_{H+1}(s) = 0, \forall s \in \mathcal{S}$. Then, we can calculate that

$$Q_H^\pi(s,a) = r_H(s,a) + OCE^u_{s'\sim P_H(\cdot|s,a)}(V^\pi_{H+1}(s')) \overset{(1)}{=} r_H(s,a) \in [0,1],$$
$$V_H^\pi(s) = \max_{a\in\mathcal{A}} r_H(s,a) \in [0,1],$$

where equality (1) is due to property (b) in Lemma A.1. Hence, we have

$$Q^\pi_{H-1}(s,a) = r_{H-1}(s,a) + OCE^u_{s'\sim P_{H-1}(\cdot|s,a)}(V^\pi_H(s')) \overset{(1)}{\leq} r_{H-1}(s,a) + OCE^u_{s'\sim P_{H-1}(\cdot|s,a)}(1) \in [1,2],$$
$$Q^\pi_{H-1}(s,a) = r_{H-1}(s,a) + OCE^u_{s'\sim P_{H-1}(\cdot|s,a)}(V^\pi_H(s')) \overset{(2)}{\geq} r_{H-1}(s,a) + OCE^u_{s'\sim P_{H-1}(\cdot|s,a)}(0) \in [0,1],$$
$$V^\pi_{H-1}(s) = \max_{a\in\mathcal{A}} Q^\pi_{H-1}(s,a) \in [0,2],$$

where inequalities (1) and (2) hold due to properties (b) and (c) in Lemma A.1. Carrying out this procedure repeatedly until step h, we can get

$$Q_h^\pi(s,a) \in [0, H-h+1], \quad \text{and} \quad V_h^\pi(s) \in [0, H-h+1].$$

Using the definition (7), we then immediately obtain that $0 \leq Regret(\mathcal{M}, \textbf{algo}, K) \leq KH$ for any **algo**. ☐

With Lemma A.2, we can obtain the following result, which shows that the optimization problem in $OCE^u_{s'\sim P_h(\cdot|s,a)}(V^\pi_{h+1}(s'))$ has an optimal solution in the support of the random variable $V^\pi_{h+1}(s')$.

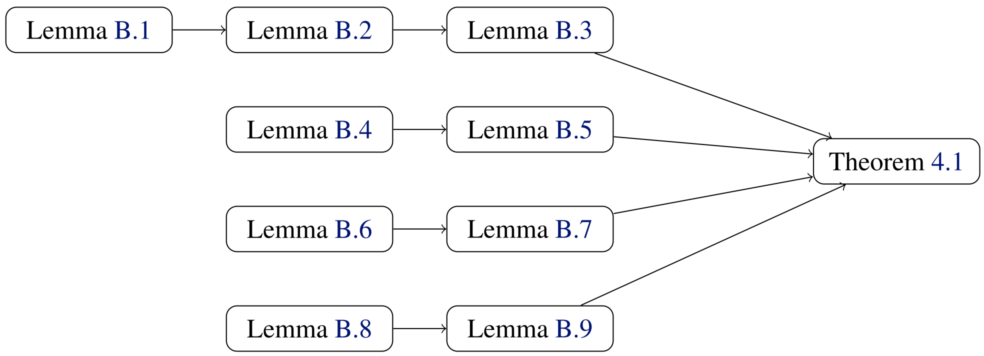
Lemma A.3. *For any probability measure $P_h(\cdot|s,a)$, any $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$, suppose $V^\pi_{h+1}(s') \in [0, H-h]$ for $s' \sim P_h(\cdot|s,a)$. Then, we have*

$$OCE^u_{s'\sim P_h(\cdot|s,a)}(V^\pi_{h+1}(s')) = \max_{\lambda\in[0,H-h]} \{\lambda + E_{s'\sim P_h(\cdot|s,a)}[u(V^\pi_{h+1}(s') - \lambda)]\}. \tag{17}$$

Proof. Note that $V^\pi_{h+1}(s') \in [0, H-h]$ for $s' \sim P_h(\cdot|s,a)$ by Lemma A.2. By the concavity and continuity of u, we deduce that the function $G(\lambda) := \lambda + E_{s'\sim P_h(\cdot|s,a)}[u(V^\pi_{h+1}(s') - \lambda)]$ is concave and continuous, and moreover, $G(\lambda) \leq E_{s'\sim P_h(\cdot|s,a)}[V^\pi_{h+1}(s')] < \infty$ for all $\lambda \in \mathbb{R}$ due to the fact that $u(x) \leq x$ for all x. In addition, $\partial G(\lambda) = 1 - E_{s'\sim P_h(\cdot|s,a)}[\partial u(V^\pi_{h+1}(s') - \lambda)]$ due to the finite state space $\mathcal{S}$, and thus, $G(\lambda)$ will be nonincreasing when $\lambda \geq H-h$ due to the fact that $\eta \geq 1$ for all $\eta \in \partial u(x), x \leq 0$. It follows that the set of optimal solutions to the problem $\sup_{\lambda\in\mathbb{R}} G(\lambda)$ is nonempty. Hence, we can apply Proposition 2.1 in Ben-Tal & Teboulle (2007) and obtain the desired result. ☐

B. Proof of Theorem 4.1

We present a series of lemmas in Section B.1, and prove Theorem 4.1 in Section B.2. The relation of different lemmas is given below.



B.1. Preparations for the Proof of Theorem 4.1

In this subsection, we state and prove a few lemmas needed for the proof of theorem 4.1. Recall that the bonus in the OCE-VI algorithm is $b_h^k(s,a) = |u(-H+h)|\sqrt{\frac{2\log(\frac{SAHK}{\delta})}{\max\{1,N_h^k(s,a)\}}}$ for any $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$.

Lemma B.1 provides a bound for the difference between $E_{s'\sim P_h(\cdot|s,a)}\left[u(V_{h+1}^*(s') - \lambda_{h+1}^*)\right]$ and its estimation for all $h\in[H]$, where

$$\lambda_{h+1}^* \in \operatorname*{arg\,max}_{\lambda\in[0,H-h]} \{\lambda + E_{s'\sim P_h(\cdot|s,a)}[u(V_{h+1}^*(s') - \lambda)]\}.$$

Note that both V_{h+1}^* and λ_{h+1}^* are deterministic quantities. To facilitate the presentation, we let

$$\mathbb{H}_h^k = ((\mathcal{S}\times\mathcal{A})^{H-1}\times\mathcal{S})^{k-1}\times(\mathcal{S}\times\mathcal{A})^{h-1}\times\mathcal{S} \tag{18}$$

be the set of possible histories up to step h in episode k. Then, one sample of the history up to step h in episode k is

$$\mathcal{H}_h^k = (s_1^1, a_1^1, s_2^1, a_2^1, \cdots, s_H^1, \cdots, s_1^k, a_1^k, \cdots, s_{h-1}^k, a_{h-1}^k, s_h^k) \in \mathbb{H}_h^k.$$

Lemma B.1. *For any $\delta\in(0,1)$, we have*

$$\begin{aligned} P\Bigg(& E_{s'\sim P_h(\cdot|s,a)}\left[u(V_{h+1}^*(s') - \lambda_{h+1}^*)\right] - E_{s'\sim \hat{P}_h^k(\cdot|s,a)}\left[u(V_{h+1}^*(s') - \lambda_{h+1}^*)\right] \\ & \le \left|u(H-h-\lambda_{h+1}^*) - u(-\lambda_{h+1}^*)\right| \sqrt{\frac{2\log(\frac{SAHK}{\delta})}{\max\{1,N_h^k(s,a)\}}}, \\ & V_{h+1}^*: \mathcal{S}\to[0,H-h], \lambda_{h+1}^*\in[0,H-h], \forall (s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]\Bigg) \ge 1-\delta. \end{aligned} \tag{19}$$

Proof. We adapt the proof of Lemma 7.3 in Agarwal et al. (2021) who consider the risk neural episodic RL setting. For each fixed $(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]$, we have to consider two cases.

Firstly, according to section 3, when $N_h^k(s,a) = 0$, we have $\hat{P}_h^k(s'|s,a) = 0$ for all $s'\in\mathcal{S}$. According to Lemma A.2 and Lemma A.3, $V_{h+1}^*(s_{h+1}^i)\in[0,H-h]$ and $\lambda_{h+1}^*\in[0,H-h]$. Thus, we have $u(V_{h+1}^*(s_{h+1}^i) - \lambda_{h+1}^*) \in$

$[u(-\lambda^*_{h+1}), u(H-h-\lambda^*_{h+1})]$, where $u(-\lambda^*_{h+1}) \leq 0$ and $u(H-h-\lambda^*_{h+1}) \geq 0$. Then, we have

$$\begin{aligned}
&E_{s'\sim P_h(\cdot|s,a)}\left[u(V^*_{h+1}(s')-\lambda^*_{h+1})\right] - E_{s'\sim \hat{P}^k_h(\cdot|s,a)}\left[u(V^*_{h+1}(s')-\lambda^*_{h+1})\right] \\
&\overset{(1)}{=} E_{s'\sim P_h(\cdot|s,a)}\left[u(V^*_{h+1}(s')-\lambda^*_{h+1})\right] \\
&\overset{(2)}{\leq} |u(H-h-\lambda^*_{h+1}) - u(-\lambda^*_{h+1})| \\
&\overset{(3)}{\leq} \left|u(H-h-\lambda^*_{h+1}) - u(-\lambda^*_{h+1})\right| \sqrt{\frac{2\log(\frac{SAHK}{\delta})}{\max\{1, N^k_h(s,a)\}}},
\end{aligned}$$

where equality (1) holds because $\hat{P}^k_h(s'|s,a) = 0$ for all $s' \in \mathcal{S}$, inequality (2) holds because

$$u(V^*_{h+1}(s^i_{h+1}) - \lambda^*_{h+1}) \leq u(H-h-\lambda^*_{H+1}) \leq |u(H-h-\lambda^*_{h+1}) - u(-\lambda^*_{h+1})|,$$

and inequality (3) holds because $N^k_h(s,a) = 0$ and $\log(\frac{SAHK}{\delta}) > 1$. Therefore,

$$\begin{aligned}
P\Bigg(&E_{s'\sim P_h(\cdot|s,a)}\left[u(V^*_{h+1}(s')-\lambda^*_{h+1})\right] - E_{s'\sim \hat{P}^k_h(\cdot|s,a)}\left[u(V^*_{h+1}(s')-\lambda^*_{h+1})\right] \\
&\leq \left|u(H-h-\lambda^*_{h+1}) - u(-\lambda^*_{h+1})\right| \sqrt{\frac{2\log(\frac{SAHK}{\delta})}{\max\{1, N^k_h(s,a)\}}}, \\
&V^*_{h+1}: \mathcal{S} \to [0, H-h], \lambda^*_{h+1} \in [0, H-h]\Bigg) = 1 \geq 1 - \frac{\delta}{SAHK}.
\end{aligned}$$

Secondly, when $N^k_h(s,a) \geq 1$, by the definition of $\hat{P}^k_h$, we have

$$E_{s'\sim \hat{P}^k_h(\cdot|s,a)}\left[u(V^*_{h+1}(s')-\lambda^*_{h+1})\right] = \frac{1}{N^k_h(s,a)} \sum_{i=1}^{k-1} 1_{\{(s^i_h, a^i_h)=(s,a)\}} u(V^*_{h+1}(s^i_{h+1}) - \lambda^*_{h+1}).$$

Remark that when $N^k_h(s,a) \geq 1$, we have $k \geq 2$. We define for $i = 1, \dots, k-1$,

$$X_i = E\left[1_{\{(s^i_h, a^i_h)=(s,a)\}} u(V^*_{h+1}(s^i_{h+1}) - \lambda^*_{h+1}) \middle| \mathcal{H}^i_h\right] - 1_{\{(s^i_h, a^i_h)=(s,a)\}} u(V^*_{h+1}(s^i_{h+1}) - \lambda^*_{h+1}).$$

By the same argument as in the previous case, we conclude that $u(V^*_{h+1}(s^i_{h+1}) - \lambda^*_{h+1}) \in [u(-\lambda^*_{h+1}), u(H-h-\lambda^*_{h+1})]$. Thus, we have

$$u(-\lambda^*_{h+1}) - u(H-h-\lambda^*_{h+1}) \leq X_i \leq u(H-h-\lambda^*_{h+1}) - u(-\lambda^*_{h+1}).$$

In addition, it is evident that $E[X_i|\mathcal{H}^i_h] = 0$, which implies that (X_i) is a martingale difference sequence. Then, by Azuma-Hoeffding's inequality for martingales, with a probability of at least $1 - \frac{\delta}{SAHK}$, we have

$$\begin{aligned}
\sum_{i=1}^{k-1} X_i &= N^k_h(s,a) E_{s'\sim P_h(\cdot|s,a)}[u(V^*_{h+1}(s') - \lambda^*_{h+1})] - \sum_{i=1}^{k-1} 1_{\{(s^i_h, a^i_h)=(s,a)\}} u(V^*_{h+1}(s^i_{h+1}) - \lambda^*_{h+1}) \\
&\leq |u(H-h-\lambda^*_{h+1}) - u(-\lambda^*_{h+1})| \sqrt{2N^k_h(s,a)\log(\frac{SAHK}{\delta})}
\end{aligned}$$

Divided by $N^k_h(s,a)$ on both sides of the above inequality, combining the above two cases and using a union bound over all $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, we obtain (19). □

By Lemma B.1, we can derive the following concentration bound for the OCE applied to the optimal value function at the next state (under the estimated transition distribution).

Lemma B.2. *For any $\delta \in (0,1)$, we have that*

$$P\Bigg(OCE^u_{s'\sim P_h(\cdot|s,a)}(V^*_{h+1}(s')) - OCE^u_{s'\sim \hat{P}^k_h(\cdot|s,a)}(V^*_{h+1}(s')) \le |u(-H+h)|\sqrt{\frac{2\log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N^k_h(s,a)\}}},$$
$$V^*_{h+1}: \mathcal{S} \to [0, H-h], \forall (s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]\Bigg) \ge 1-\delta. \tag{20}$$

Proof. According to Lemma B.1, with a probability of at least $1-\delta$, for any $k \in [K], s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$, we have

$$\begin{aligned}
&OCE^u_{s'\sim P_h(\cdot|s,a)}(V^*_{h+1}(s')) - OCE^u_{s'\sim \hat{P}^k_h(\cdot|s,a)}(V^*_{h+1}(s')) \\
&\overset{(1)}{=} \max_{\lambda\in[0,H-h]}\{\lambda + E_{s'\sim P_h(\cdot|s,a)}[u(V^*_{h+1}(s')-\lambda)]\} - \max_{\lambda\in[0,H-h]}\{\lambda + E_{s'\sim \hat{P}^k_h(\cdot|s,a)}[u(V^*_{h+1}(s')-\lambda)]\} \\
&\overset{(2)}{\le} \lambda^*_{h+1} + E_{s'\sim P_h(\cdot|s,a)}[u(V^*_{h+1}(s')-\lambda^*_{h+1})] - \lambda^*_{h+1} - E_{s'\sim \hat{P}^k_h(\cdot|s,a)}[u(V^*_{h+1}(s')-\lambda^*_{h+1})] \\
&\overset{(3)}{\le} \left|u(H-h-\lambda^*_{h+1}) - u(-\lambda^*_{h+1})\right|\sqrt{\frac{2\log(\frac{SAHK}{\delta})}{\max\{1, N^k_h(s,a)\}}} \\
&\le \sup_{\lambda\in[0,H-h]}|u(H-h-\lambda) - u(-\lambda)|\sqrt{\frac{2\log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N^k_h(s,a)\}}},
\end{aligned}$$

where equality (1) follows from Lemma A.3, inequality (2) holds because λ^*_{h+1} is the optimal solution to $\max_{\lambda\in[0,H-h]}\{\lambda + E_{s'\sim P_h(\cdot|s,a)}[u(V^*_{h+1}(s')-\lambda)]\}$ and inequality (3) follows from Lemma B.1. One can check that $u(H-h-\lambda) - u(-\lambda)$ is a nondecreasing function of $\lambda \in [0, H-h]$. To see this, note that the superdifferential of $u(H-h-\lambda)-u(-\lambda)$ is $\partial u(-\lambda) - \partial u(H-h-\lambda)$ and for any $z \in \partial u(-\lambda) - \partial u(H-h-\lambda)$, we have $z \ge 0$, because the utility function u is concave. This implies that the function $u(H-h-\lambda)-u(-\lambda)$ is nondecreasing. In addition, this function is non-negative because u is nondcreasing and thus $u(H-h)-u(0) \ge 0$ for $h \in [H]$. Thus, we can deduce that $\sup_{\lambda\in[0,H-h]}|u(H-h-\lambda)-u(-\lambda)| \le u(0) - u(-H+h) = |u(-H+h)|$ since $u(0) = 0$. The proof is then complete. □

Lemma B.2 immediately implies the following result. To facilitate the presentation, we define the following event from Lemma B.2:

$$\mathcal{G}_1 = \Bigg\{OCE^u_{s'\sim P_h(\cdot|s,a)}(V^*_{h+1}(s')) - OCE^u_{s'\sim \hat{P}^k_h(\cdot|s,a)}(V^*_{h+1}(s')) \le |u(-H+h)|\sqrt{\frac{2\log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N^k_h(s^k_h, a^k_h)\}}},$$
$$V^*_{h+1}: \mathcal{S}\to[0,H-h], \forall(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]\Bigg\}. \tag{21}$$

Lemma B.3 (Optimism)**.** *Conditional on the event $\mathcal{G}_1$, we have $\hat{V}^k_h(s) \ge V^*_h(s)$ for any $k \in [K], s \in \mathcal{S}, h \in [H]$.*

Proof. We prove the result by induction. Set $\hat{V}^k_{H+1}(s) = V^*_{H+1}(s) = 0, \forall s \in \mathcal{S}$. Conditional on the occurrence of the event $\mathcal{G}_1$, assume $\hat{V}^k_{h+1}(s') \ge V^*_{h+1}(s'), \forall s' \in \mathcal{S}$. Then, under event $\mathcal{G}_1$, for step h, we have

$$\begin{aligned}
&b^k_h(s,a) + r_h(s,a) + OCE^\phi_{s'\sim \hat{P}^k_h(\cdot|s,a)}(\hat{V}^k_{h+1}(s')) - r_h(s,a) - OCE^\phi_{s'\sim P_h(\cdot|s,a)}(V^*_{h+1}(s')) \\
&\overset{(1)}{\ge} b^k_h(s,a) + OCE^\phi_{s'\sim \hat{P}^k_h(\cdot|s,a)}(V^*_{h+1}(s')) - OCE^\phi_{s'\sim P_h(\cdot|s,a)}(V^*_{h+1}(s')) \\
&\overset{(2)}{\ge} b^k_h(s,a) - |u(-H+h)|\sqrt{\frac{2\log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N^k_h(s,a)\}}} \\
&= 0,
\end{aligned} \tag{22}$$

where inequality (1) follows from the assumption $\hat{V}_{h+1}^k(s') \geq V_{h+1}^*(s'), \forall s' \in \mathcal{S}$ and property (c) in Lemma A.1, and inequality (2) holds due to Lemma B.2. Recall that

$$\hat{Q}_h^k(s,a) = \min\{b_h^k(s,a) + r_h(s,a) + OCE^{\phi}_{s' \sim \hat{P}_h^k(\cdot|s,a)}(\hat{V}_{h+1}^k(s')), H-h+1\},$$
$$Q_h^*(s,a) = r_h(s,a) + OCE^{\phi}_{s' \sim P_h(\cdot|s,a)}(V_{h+1}^*(s')).$$

By (22) and Lemma A.2, we can immediately obtain

$$\hat{Q}_h^k(s,a) - Q_h^*(s,a) \geq 0.$$

Because $\hat{V}_h^k(s) = \max_{a' \in \mathcal{A}} \hat{Q}_h^k(s,a')$, we have $\hat{V}_h^k(s) \geq V_h^*(s)$. The result then follows by induction. □

We next state a concentration bound (Lemma B.5) for the OCE of the estimated next-state value function $\hat{V}_{h+1}^k$ under the estimated transition distribution $\hat{P}_h^k$. This is different from Lemma B.2 in that $\hat{V}_{h+1}^k$ is a random quantity depending on the data while the optimal value function is deterministic. The proof of Lemma B.5 relies on the following well-known result on the L^1 concentration bound for the empirical transition probabilities (see, e.g., Lemma 17 in Jaksch et al. (2010)):

Lemma B.4. *For any $\delta \in (0,1)$, we have*

$$P\left(\left\| \hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a) \right\|_1 \leq \sqrt{\frac{2S\log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N_h^k(s,a)\}}}, \forall (s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K] \right) \geq 1-\delta.$$

Lemma B.5. *For any $\delta \in (0,1)$, we have*

$$\begin{aligned} P\Bigg(&\left| OCE^u_{s' \sim P_h(\cdot|s,a)}(\hat{V}_{h+1}^k(s')) - OCE^u_{s' \sim \hat{P}_h^k(\cdot|s,a)}(\hat{V}_{h+1}^k(s')) \right| \leq |u(-H+h)| \sqrt{\frac{2S\log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N_h^k(s,a)\}}}, \\ &\forall (s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K] \Bigg) \geq 1-\delta. \end{aligned} \tag{23}$$

Proof. With probability at least $1-\delta$, we have that for any $k \in [K], s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$,

$$\begin{aligned} &\left| OCE^u_{s' \sim P_h(\cdot|s,a)}(\hat{V}_{h+1}^k(s')) - OCE^u_{s' \sim \hat{P}_h^k(\cdot|s,a)}(\hat{V}_{h+1}^k(s')) \right| \\ &= \left| \max_{\lambda \in [0,H-h]} \{\lambda + E_{s' \sim P_h(\cdot|s,a)}[u(\hat{V}_{h+1}^k(s') - \lambda)]\} - \max_{\lambda \in [0,H-h]} \{\lambda + E_{s' \sim \hat{P}_h^k(\cdot|s,a)} u(\hat{V}_{h+1}^k(s') - \lambda)]\} \right| \\ &\leq \max_{\lambda \in [0,H-h]} \left| E_{s' \sim P_h(\cdot|s,a)}[u(\hat{V}_{h+1}^k(s') - \lambda)] - E_{s' \sim \hat{P}_h^k(\cdot|s,a)}[u(\hat{V}_{h+1}^k(s') - \lambda) \right| \\ &= \max_{\lambda \in [0,H-h]} \left| \sum_{s' \in \mathcal{S}} \left(\hat{P}_h^k(s'|s,a) - P_h(s'|s,a) \right) \cdot u(\hat{V}_{h+1}^k(s') - \lambda) \right| \\ &\overset{(1)}{\leq} \max_{\lambda \in [0,H-h]} \left\| \hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a) \right\|_1 \cdot \left\| u(\hat{V}_{h+1}^k(\cdot) - \lambda) \right\|_\infty \\ &\overset{(2)}{\leq} \sqrt{\frac{2S\log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N_h^k(s,a)\}}} \cdot \max_{\lambda \in [0,H-h]} \left\| u(\hat{V}_{h+1}^k(\cdot) - \lambda) \right\|_\infty, \end{aligned}$$

where inequality (1) follows from Hölder's inequality and inequality (2) follows from Lemma B.4. Because $\hat{V}_{h+1}^k(s') \in [0, H-h]$ for any s' by the design of the OCE-VI algorithm and because $\lambda \in [0, H-h]$, we can immediately obtain that

$$\max_{\lambda \in [0,H-h]} \left\| u(\hat{V}_{h+1}^k(\cdot) - \lambda) \right\|_\infty \leq |u(-H+h)|,$$

where we use the fact that u is nondecreasing and concave. The proof is then completed. □

In the next lemma, we will bound the following difference

$$OCE^u_{s'\sim P_h(\cdot|s,a)}\left(\hat{V}^k_{h+1}(s')\right) - OCE^u_{s'\sim P_h(\cdot|s,a)}\left(V^{\pi^k}_{h+1}(s')\right) \tag{24}$$

for any $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$, which is the key step in the recursion of the regret analysis.

We first introduce some notations. Pick any $\lambda^k_{h+1} \in [0, H-h]$ such that

$$\lambda^k_{h+1} \in \arg\max_{\lambda\in[0,H-h]}\{\lambda + E_{s'\sim P_h(\cdot|s,a)}[u(V^{\pi^k}_{h+1}(s') - \lambda)]\}. \tag{25}$$

By the first order optimality condition of the above optimization problem and the fact that the state space $\mathcal{S}$ is finite, we have

$$1 \in E_{s'\sim P_h(\cdot|s,a)}[\partial u(V^{\pi^k}_{h+1}(s') - \lambda^k_{h+1})]. \tag{26}$$

Thus, we can find $\Lambda^k_{h+1}(s') \in \partial u(V^{\pi^k}_{h+1}(s') - \lambda^k_{h+1}), s' \in \mathcal{S}$ such that $E_{s'\sim P_h(\cdot|s,a)}\left[\Lambda^k_{h+1}(s')\right] = 1$. In addition, because the utility function u is nondecreasing, we have $\Lambda^k_{h+1}(s') \geq 0, \forall s' \in \mathcal{S}$. Now we can define the following new probability measure $B_h(\cdot|s,a)$: for any $(s,a,s',h,k) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]\times[K]$, define

$$B_h(s'|s,a) := P_h(s'|s,a)\Lambda^k_{h+1}(s'), \tag{27}$$

where $\sum_{s'\in\mathcal{S}} B_h(s'|s,a) = 1$ because $E_{s'\sim P_h(\cdot|s,a)}\left[\Lambda^k_{h+1}(s')\right] = 1$. Now we can state the following important result.

Lemma B.6. *For any $(h,k) \in [H]\times[K]$ and functions $\hat{V}^k_{h+1}, V^{\pi^k}_{h+1}, V^*_{h+1} : \mathcal{S} \to [0, H-h]$, we have*

$$\begin{aligned} &OCE^u_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}\left(\hat{V}^k_{h+1}(s^k_{h+1})\right) - OCE^u_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}\left(V^{\pi^k}_{h+1}(s^k_{h+1})\right) \\ &\leq E_{s^k_{h+1}\sim B_h(\cdot|s^k_h,a^k_h)}\left[\hat{V}^k_{h+1}(s^k_{h+1}) - V^{\pi^k}_{h+1}(s^k_{h+1})\right], \end{aligned} \tag{28}$$

where $B_h(\cdot|s^k_h,a^k_h)$ is the new probability measure given in (27).

Proof. Pick any $\mu^k_{h+1} \in [0, H-h]$ such that

$$\mu^k_{h+1} \in \arg\max_{\lambda\in[0,H-h]}\{\lambda + E_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}[u(\hat{V}^k_{h+1}(s^k_{h+1}) - \lambda)]\},$$

and recall $\lambda^k_{h+1} \in [0, H-h]$ given in (25). We can then compute

$$\begin{aligned} &OCE^u_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}\left(\hat{V}^k_{h+1}(s^k_{h+1})\right) - OCE^u_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}\left(V^{\pi^k}_{h+1}(s^k_{h+1})\right) \\ &\overset{(1)}{=} \max_{\lambda\in[0,H-h]}\left\{\lambda + E_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}[u(\hat{V}^k_{h+1}(s^k_{h+1}) - \lambda)]\right\} \\ &\qquad - \max_{\lambda\in[0,H-h]}\left\{\lambda + E_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}[u(V^{\pi^k}_{h+1}(s^k_{h+1}) - \lambda)]\right\} \\ &= \mu^k_{h+1} + E_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}[u(\hat{V}^k_{h+1}(s^k_{h+1}) - \mu^k_{h+1})] - \lambda^k_{h+1} - E_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}[u(V^{\pi^k}_{h+1}(s^k_{h+1}) - \lambda^k_{h+1})] \\ &\overset{(2)}{\leq} \mu^k_{h+1} - \lambda^k_{h+1} + E_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}\left[\Lambda^k_{h+1}(s^k_{h+1})\cdot(\hat{V}^k_{h+1}(s^k_{h+1}) - V^{\pi^k}_{h+1}(s^k_{h+1}) - (\mu^k_{h+1} - \lambda^k_{h+1}))\right] \\ &= \left(1 - E_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}\left[\Lambda^k_{h+1}(s^k_{h+1})\right]\right)(\mu^k_{h+1} - \lambda^k_{h+1}) \\ &\qquad + E_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}\left[\Lambda^k_{h+1}(s^k_{h+1})\cdot(\hat{V}^k_{h+1}(s^k_{h+1}) - V^{\pi^k}_{h+1}(s^k_{h+1}))\right] \\ &\overset{(3)}{=} E_{s^k_{h+1}\sim B_h(\cdot|s^k_h,a^k_h)}[\hat{V}^k_{h+1}(s^k_{h+1}) - V^{\pi^k}_{h+1}(s^k_{h+1})], \end{aligned}$$

where equality (1) holds due to Lemma A.3, inequality (2) holds due to the fact that $u(y) \leq u(x) + z(y-x)$ for any $x, y \in [-H+h, H-h], z \in \partial u(x)$ when $u(x)$ is a concave function, and equality (3) follows from (27) and the fact that $E_{s^k_{h+1}\sim P_h(\cdot|s^k_h,a^k_h)}[\Lambda^k_{h+1}(s^k_{h+1})] = 1$. The proof is therefore completed. □

In the next lemma, we bound the term $\hat{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)$ by using a recursive procedure. Lemma B.7 below is an extension of the so-called simulation lemma in the risk-neutral RL (see, e.g, Agarwal et al. (2021)) to our risk-sensitive RL setting. The key to overcome the difficulty in the recursion setting due to the nonlinearity of the OCE is Lemma B.6. To facilitate the presentation, we first introduce the following notations.

For any $k \in [K], h \in [H]$, let $w_{hk}(s_h^k, a_h^k)$ be the state-action distribution induced by π^k at time step h starting from s_1^k, i.e., the probability of π^k visiting (s_h^k, a_h^k) at time step h starting from s_1^k. Mathematically, the formula of $w_{hk}(s_h^k, a_h^k)$ is given by

$$w_{hk}(s_h^k, a_h^k) = \begin{cases} 1, & h = 1, \\ P_1(s_2^k|s_1^k, a_1^k), & h = 2, \\ \sum_{s_2^k \in \mathcal{S}} \cdots \sum_{s_{h-1}^k \in \mathcal{S}} P_1(s_2^k|s_1^k, a_1^k) \cdots P_{h-1}(s_h^k|s_{h-1}^k, a_{h-1}^k), & h \geq 3. \end{cases} \tag{29}$$

Similarly, let $w_{hk}^B(s_h^k, a_h^k)$ be the probability of π^k visiting (s_h^k, a_h^k) at step h starting from s_1^k under probability measures $B_i(\cdot|s_i^k, a_i^k), i = 1, \cdots, h$. The explicit formula of $w_{hk}^B(s_h^k, a_h^k)$ is given by

$$w_{hk}^B(s_h^k, a_h^k) = \begin{cases} 1, & h = 1, \\ B_1(s_2^k|s_1^k, a_1^k), & h = 2, \\ \sum_{s_2^k \in \mathcal{S}} \cdots \sum_{s_{h-1}^k \in \mathcal{S}} B_1(s_2^k|s_1^k, a_1^k) \cdots B_{h-1}(s_h^k|s_{h-1}^k, a_{h-1}^k), & h \geq 3. \end{cases} \tag{30}$$

Equivalently, by (27) we have

$$w_{hk}^B(s_h^k, a_h^k) = \begin{cases} 1, & h = 1, \\ P_1(s_2^k|s_1^k, a_1^k)\Lambda_2^k(s_2^k), & h = 2, \\ \sum_{s_2^k \in \mathcal{S}} \cdots \sum_{s_{h-1}^k \in \mathcal{S}} P_1(s_2^k|s_1^k, a_1^k) \cdots P_{h-1}(s_h^k|s_{h-1}^k, a_{h-1}^k)\Lambda_2^k(s_2^k) \cdots \Lambda_h^k(s_h^k), & h \geq 3. \end{cases} \tag{31}$$

Finally, given (s_1^k, a_1^k), we define

$$E_{(s_h^k, a_h^k) \sim w_{hk}^B}[\cdot] := \begin{cases} 1, & h = 1, \\ E_{s_2^k \sim B_1(\cdot|s_1^k, a_1^k)}\left[E_{s_3^k \sim B_2(\cdot|s_2^k, a_2^k)}\left[\cdots E_{s_h^k \sim B_{h-1}(\cdot|s_{h-1}^k, a_{h-1}^k)}[\cdot]\right]\right], & h \geq 2. \end{cases} \tag{32}$$

Lemma B.7. *For each episode $k \in [K]$, we have*

$$\begin{aligned} &\hat{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) \\ &\leq \sum_{h=1}^{H} E_{(s_h^k, a_h^k) \sim w_{hk}^B}\left[b_h^k(s_h^k, a_h^k) + OCE^u_{s_{h+1}^k \sim \hat{P}_h^k(\cdot|s_h^k, a_h^k)}(\hat{V}_{h+1}^k(s_{h+1}^k)) - OCE^u_{s_{h+1}^k \sim P_h(\cdot|s_h^k, a_h^k)}(\hat{V}_{h+1}^k(s_{h+1}^k))\right]. \end{aligned} \tag{33}$$

Proof. For any $k \in [K]$, let $a_h^k = \arg\max_{a\in\mathcal{A}} \hat{Q}_h^k(s_h^k, a),\ h \in [H]$. Then, we can compute

$$\begin{aligned}
&\hat{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\\
&\overset{(1)}{\le} \hat{Q}_1^k(s_1^k, a_1^k) - Q_1^{\pi^k}(s_1^k, a_1^k)\\
&\le b_1^k(s_1^k, a_1^k) + OCE^u_{s_2^k\sim \hat{P}_1^k(\cdot|s_1^k,a_1^k)}(\hat{V}_2^k(s_2^k)) - OCE^u_{s_2^k\sim P_1(\cdot|s_1^k,a_1^k)}(V_2^{\pi^k}(s_2^k))\\
&= b_1^k(s_1^k, a_1^k) + OCE^u_{s_2^k\sim \hat{P}_1^k(\cdot|s_1^k,a_1^k)}(\hat{V}_2^k(s_2^k)) - OCE^u_{s_2^k\sim P_1(\cdot|s_1^k,a_1^k)}(\hat{V}_2^k(s_2^k))\\
&+ OCE^u_{s_2^k\sim P_1(\cdot|s_1^k,a_1^k)}(\hat{V}_2^k(s_2^k)) - OCE^u_{s_2^k\sim P_1(\cdot|s_1^k,a_1^k)}(V_2^{\pi^k}(s_2^k))\\
&\overset{(2)}{\le} b_1^k(s_1^k, a_1^k) + OCE^u_{s_2^k\sim \hat{P}_1^k(\cdot|s_1^k,a_1^k)}(\hat{V}_2^k(s_2^k)) - OCE^u_{s_2^k\sim P_1(\cdot|s_1^k,a_1^k)}(\hat{V}_2^k(s_2^k))\\
&+ E_{s_2^k\sim B_1(\cdot|s_1^k,a_1^k)}\left[\hat{V}_2^k(s_2^k) - V_2^{\pi^k}(s_2^k)\right]\\
&\le b_1^k(s_1^k, a_1^k) + OCE^u_{s_2^k\sim \hat{P}_1^k(\cdot|s_1^k,a_1^k)}(\hat{V}_2^k(s_2^k)) - OCE^u_{s_2^k\sim P_1(\cdot|s_1^k,a_1^k)}(\hat{V}_2^k(s_2^k))\\
&+ E_{s_2^k\sim B_1(\cdot|s_1^k,a_1^k)}\Big[b_2^k(s_2^k, a_2^k) + OCE^u_{s_3^k\sim \hat{P}_2^k(\cdot|s_2^k,a_2^k)}(\hat{V}_3^k(s_3^k)) - OCE^u_{s_3^k\sim P_2(\cdot|s_2^k,a_2^k)}(\hat{V}_3^k(s_3^k))\\
&\qquad\qquad + E_{s_3^k\sim B_2(\cdot|s_2^k,a_2^k)}\left[\hat{V}_3^k(s_3^k) - V_3^{\pi^k}(s_3^k)\right]\Big],
\end{aligned}$$

where inequality (1) holds because $\hat{V}_1^k(s_1^k) = \max_{a\in\mathcal{A}} \hat{Q}_1^k(s_1^k, a) = \hat{Q}_1^k(s_1^k, a_1^k)$ and inequality (2) holds due to Lemma B.6. Applying the above procedure recursively and using the fact that $\hat{V}_{H+1}^k(s) = V_{H+1}^*(s) = 0$ for any $s \in \mathcal{S}$, we immediately obtain (33). □

From Lemma B.7, it is clear that we need to bound the sum of bonuses in order to bound the regret. We present such a bound in Lemma B.9. To this end, we first state Lemma B.8, which is adapted from a well-known result heavily used in the risk-neutral setting (see page 24-25 of Azar et al. (2017) or page 21 of Jin et al. (2018)). Lemma B.9 is a nontrivial extension of Lemma B.8 due to the new probability measure w_{hk}^B involved in the expectation.

Lemma B.8. *Consider arbitrary K sequences of trajectories $\tau^k = \{s_h^k, a_h^k\}_{h=1}^H$ for $k = 1, \cdots, K$, we have*

$$\sum_{k=1}^K \frac{1}{\max\{1, N_h^k(s_h^k, a_h^k)\}} \le SA\log(3K).$$

Lemma B.9. *We have*

$$E\left[\sum_{k=1}^K\sum_{h=1}^H E_{(s_h^k,a_h^k)\sim w_{hk}^B}\left[\frac{|u(-H+h)|}{\sqrt{\max\{1, N_h^k(s_h^k, a_h^k)\}}}\right]\right] \le \sum_{h=1}^H |u(-H+h)|\sqrt{\prod_{i=1}^{h-1} u'_-(-H+i)SAK\log(3K)}, \tag{34}$$

where $E_{(s_h^k,a_h^k)\sim w_{hk}^B}[\cdot]$ given in (32) is taken over (s_h^k, a_h^k) conditional on (s_1^k, a_1^k) and $u'_-(\cdot)$ is the left derivative of u.

Proof. By (31) and (32), we have

$$\begin{aligned}
&E_{(s_h^k,a_h^k)\sim w_{hk}^B}\left[\frac{|u(-H+h)|}{\sqrt{\max\{1, N_h^k(s_h^k, a_h^k)\}}}\right]\\
&= \sum_{s_2^k\in\mathcal{S}}\cdots\sum_{s_h^k\in\mathcal{S}} P_1(s_2^k|s_1^k, a_1^k)\cdots P_{h-1}(s_h^k|s_{h-1}^k, a_{h-1}^k)\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k)\frac{|u(-H+h)|}{\sqrt{\max\{1, N_h^k(s_h^k, a_h^k)\}}}.
\end{aligned}\tag{35}$$

This implies

$$
\begin{aligned}
&E\left[E_{(s_h^k,a_h^k)\sim w_{hk}^B}\left[\frac{|u(-H+h)|}{\sqrt{\max\{1,N_h^k(s_h^k,a_h^k)\}}}\right]\right] \\
&= E\left[E\left[\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k)\frac{|u(-H+h)|}{\sqrt{\max\{1,N_h^k(s_h^k,a_h^k)\}}}\middle| s_1^k,a_1^k\right]\right] \\
&= E\left[\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k)\frac{|u(-H+h)|}{\sqrt{\max\{1,N_h^k(s_h^k,a_h^k)\}}}\right].
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&E\left[\sum_{k=1}^K\sum_{h=1}^H E_{(s_h^k,a_h^k)\sim w_{hk}^B}\left[\frac{|u(-H+h)|}{\sqrt{\max\{1,N_h^k(s_h^k,a_h^k)\}}}\right]\right] \\
&= \sum_{h=1}^H\sum_{k=1}^K E\left[E_{(s_h^k,a_h^k)\sim w_{hk}^B}\left[\frac{|u(-H+h)|}{\sqrt{\max\{1,N_h^k(s_h^k,a_h^k)\}}}\right]\right] \\
&= \sum_{h=1}^H |u(-H+h)|\sum_{k=1}^K E\left[\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k)\frac{1}{\sqrt{\max\{1,N_h^k(s_h^k,a_h^k)\}}}\right] \\
&\overset{(1)}{\leq} \sum_{h=1}^H |u(-H+h)|\sum_{k=1}^K \sqrt{E\left[\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k)\right]^2}\cdot\sqrt{E\left[\frac{1}{\max\{1,N_h^k(s_h^k,a_h^k)\}}\right]} \\
&\overset{(2)}{\leq} \sum_{h=1}^H |u(-H+h)|\sqrt{\sum_{k=1}^K E\left[\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k)\right]^2}\cdot\sqrt{\sum_{k=1}^K E\left[\frac{1}{\max\{1,N_h^k(s_h^k,a_h^k)\}}\right]}
\end{aligned}
$$

where the inequalities (1) and (2) follow from Cauchy–Schwarz inequality.

Recall that $\Lambda_{h+1}^k(s') \in \partial u(V_{h+1}^{\pi^k}(s') - \lambda_{h+1}^k), s' \in \mathcal{S}$ and it satisfies $E_{s'\sim P_h(\cdot|s,a)}\left[\Lambda_{h+1}^k(s')\right] = 1$, where λ_{h+1}^k is defined in (25). By Lemma A.2 and Lemma A.3 and the concavity of the utility function u, we have $0 \leq \Lambda_{h+1}^k \leq u'_-(-H+h)$. Because $E_{(s_h^k,a_h^k)\sim w_{hk}^B}[1] = \sum_{s_2^k\in\mathcal{S}}\cdots\sum_{s_h^k\in\mathcal{S}} P_1(s_2^k|s_1^k,a_1^k)\cdots P_{h-1}(s_h^k|s_{h-1}^k,a_{h-1}^k)\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k) = 1$, taking the expectation on both sides, we have $E\left[\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k)\right] = 1$. Then, we have

$$
\begin{aligned}
&\sum_{k=1}^K E\left[\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k)\right]^2 \\
&\leq \sum_{k=1}^K E\left[\Lambda_2^k(s_2^k)\cdots\Lambda_h^k(s_h^k)\right]\cdot\prod_{i=1}^{h-1} u'_-(-H+i) \\
&= K\cdot\prod_{i=1}^{h-1} u'_-(-H+i).
\end{aligned}
$$

Combining this inequality and Lemma B.8, we have

$$
\begin{aligned}
&E\left[\sum_{k=1}^K\sum_{h=1}^H E_{(s_h^k,a_h^k)\sim w_{hk}^B}\left[\frac{|u(-H+h)|}{\sqrt{\max\{1,N_h^k(s_h^k,a_h^k)\}}}\right]\right] \\
&\leq \sum_{h=1}^H |u(-H+h)|\sqrt{\prod_{i=1}^{h-1} u'_-(-H+i) SAK\log(3K)},
\end{aligned}
$$

which completes the proof. □

B.2. Proof of Theorem 4.1

Now we are ready to prove Theorem 4.1. Recall $\mathcal{G}_1$ in (21) and we define

$$\mathcal{G}_2 = \left\{ \left| OCE^u_{s' \sim P_h(\cdot|s,a)}(\hat{V}^k_{h+1}(s')) - OCE^u_{s' \sim \hat{P}^k_h(\cdot|s,a)}(\hat{V}^k_{h+1}(s')) \right| \leq |u(-H+h)| \sqrt{\frac{2S \log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N^k_h(s^k_h, a^k_h)\}}}, \right.$$
$$\left. \forall (s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K] \right\}.$$

We also define $\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2$. From Lemmas B.2 and B.5, we know that $P(\mathcal{G}_1) \geq 1-\delta$ and $P(\mathcal{G}_2) \geq 1-\delta$, which implies that $P(\mathcal{G}) \geq 1 - 2\delta$.

Proof of Theorem 4.1. For any $k \in [K]$, let $a^k_h = \arg\max_{a \in \mathcal{A}} \hat{Q}^k_h(s^k_1, a),\ h \in [H]$. Then, when the event $\mathcal{G}$ holds, we can compute

$$\begin{aligned}
&V^*_1(s^k_1) - V^{\pi^k}_1(s^k_1) \\
&\overset{(1)}{\leq} \hat{V}^k_1(s^k_1) - V^{\pi^k}_1(s^k_1) \\
&\overset{(2)}{\leq} \sum_{h=1}^H E_{(s^k_h, a^k_h) \sim w^B_{hk}} \left[b^k_h(s^k_h, a^k_h) + OCE^u_{s^k_{h+1} \sim \hat{P}^k_h(\cdot|s^k_h, a^k_h)}(\hat{V}^k_{h+1}(s^k_{h+1})) - OCE^u_{s^k_{h+1} \sim P_h(\cdot|s^k_h, a^k_h)}(\hat{V}^k_{h+1}(s^k_{h+1})) \right] \\
&\overset{(3)}{\leq} \sum_{h=1}^H E_{(s^k_h, a^k_h) \sim w^B_{hk}} \left[2\sqrt{2}|u(-H+h)| \sqrt{\frac{S \log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N^k_h(s^k_h, a^k_h)\}}} \right],
\end{aligned} \tag{36}$$

where inequality (1) follows from Lemma B.3, inequality (2) holds due to Lemma B.7, inequality (3) holds due to Lemma B.5 and the fact that $b^k_h(s^k_h, a^k_h) \leq |u(-H+h)| \sqrt{\frac{2S \log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N^k_h(s^k_h, a^k_h)\}}}$.

We can write the expected regret as follows:

$$\begin{aligned}
&Regret(\mathcal{M}, \textbf{OCE-VI}, K) \\
&= E\left[\sum_{k=1}^K \left(V^*_1(s^k_1) - V^{\pi^k}_1(s^k_1) \right) \right] \\
&= E\left[1_{\mathcal{G}} \cdot \sum_{k=1}^K \left(V^*_1(s^k_1) - V^{\pi^k}_1(s^k_1) \right) \right] + E\left[1_{\mathcal{G}^c} \cdot \sum_{k=1}^K \left(V^*_1(s^k_1) - V^{\pi^k}_1(s^k_1) \right) \right] \\
&\overset{(4)}{\leq} E\left[1_{\mathcal{G}} \cdot \sum_{k=1}^K \left(V^*_1(s^k_1) - V^{\pi^k}_1(s^k_1) \right) \right] + 2\delta K H,
\end{aligned}$$

where inequality (4) holds because $P(\mathcal{G}^c) \leq 2\delta$ and $0 \leq V^{\pi^k}_1(s^k_1) \leq V^*_1(s^k_1) \leq H$ by Lemma A.2. Using (36) and Lemma B.9, we deduce that

$$\begin{aligned}
&E\left[1_{\mathcal{G}} \cdot \sum_{k=1}^K \left(V^*_1(s^k_1) - V^{\pi^k}_1(s^k_1) \right) \right] \\
&\leq E\left[\sum_{k=1}^K \sum_{h=1}^H E_{(s^k_h, a^k_h) \sim w^B_{hk}} \left[2\sqrt{2}|u(-H+h)| \sqrt{\frac{S \log\left(\frac{SAHK}{\delta}\right)}{\max\{1, N^k_h(s^k_h, a^k_h)\}}} \right] \right] \\
&\leq 2\sqrt{2} \sum_{h=1}^H |u(-H+h)| S \sqrt{\prod_{i=1}^{h-1} u'_{-}(-H+i) A K \log(3K) \log\left(\frac{SAHK}{\delta}\right)}.
\end{aligned}$$

Then, we have

$$\begin{aligned}
&Regret(\mathcal{M}, \textbf{OCE-VI}, K) \\
&\leq 2\sqrt{2}\sum_{h=1}^{H}|u(-H+h)|S\sqrt{\prod_{i=1}^{h-1}u'_{-}(-H+i)AK\log(3K)\log\left(\frac{SAHK}{\delta}\right)}+2\delta KH \\
&\leq 2\sqrt{2}\sum_{h=1}^{H}|u(-H+h)|S\sqrt{\prod_{i=1}^{h-1}u'_{-}(-H+i)AK\log(3K)\log\left(2SAH^2K^2\right)}+1,
\end{aligned}$$

where the last inequality follows by choosing $\delta = \frac{1}{2KH}$. The proof is then completed. □

C. Proof of Theorem 4.3

Proof. We adapt the proof of Theorem 9 in Domingues et al. (2021) to our risk-sensitive setting. The proof of Theorem 4.3 is long, so we divide it into a few steps.

- **Step 1: Constructing the hard MDP instances.**

  We first construct hard MDP instances, which are almost the same as the ones in Domingues et al. (2021) except one small yet important difference: the transition probabilities in (37).

  Based on assumption 4.2, we can construct a full A-ary tree of depth $d-1$ with root $\tilde{s}_{root}$, which has $S-3$ states. In this rooted tree, each node has exactly A children and the total number of nodes is given by $\sum_{i=0}^{d-1} A^i = S-3$. We add three special states to the tree: a "waiting" state $\tilde{s}_w$ where the agent starts and can choose action $\tilde{a}_w$ to stay up to a stage $\bar{H} < H-d$, a "good" state $\tilde{s}_g$ where the agent obtains rewards, and a "bad" state $\tilde{s}_b$ that gives no reward. Note that $\bar{H}$ is a parameter to be chosen later. Both $\tilde{s}_g$ and $\tilde{s}_b$ are absorbing states. For any state in the tree, the transitions are deterministic, the a-th action in a node leads to the a-th child of that node. The agent stays or leaves $\tilde{s}_w$ with probability

  $$P_h(\tilde{s}_w|\tilde{s}_w, a) := 1\{a = \tilde{a}_w, h \leq \bar{H}\},\ P_h(\tilde{s}_{root}|\tilde{s}_w, a) := 1 - P_h(\tilde{s}_w|\tilde{s}_w, a).$$

  Then, the agent transverses the tree until she arrives at the leaves. Let L be the number of leaves, and the set of the leaves is $\mathcal{L} = \{\tilde{s}_1, \cdots, \tilde{s}_L\}$. For any $\tilde{s}_i \in \mathcal{L}$, any action will lead to a transition to either $\tilde{s}_g$ or $\tilde{s}_b$ with the transition probability

  $$P_h(\tilde{s}_g|\tilde{s}_i, a) = p + \Delta_{(h^*,s^*,a^*)}(h, \tilde{s}_i, a),\ P_h(\tilde{s}_b|\tilde{s}_i, a) = 1 - p - \Delta_{(h^*,s^*,a^*)}(h, \tilde{s}_i, a), \tag{37}$$

  where the parameter p and the function Δ will be specified later. In Domingues et al. (2021), p is set to be 0.5 in the risk-neutral setting, whereas we will tune p in our risk-sensitive setting to obtain a tighter regret lower bound.

  The reward function is defined as

  $$r_h(s, a) := 1\{s = \tilde{s}_g, h \geq \bar{H} + d + 1\},\ \forall a \in \mathcal{A}.$$

  For each

  $$(h^*, s^*, a^*) \in \{1+d, \cdots, \bar{H}+d\} \times \mathcal{L} \times \mathcal{A} =: \mathcal{Z},$$

  we define an MDP $\mathcal{M}_{(h^*,s^*,a^*)}$, where $\Delta_{(h^*,s^*,a^*)}(h, \tilde{s}_i, a) = 1\{h = h^*, \tilde{s}_i = s^*, a = a^*\}\epsilon$ and ϵ is a parameter to be chosen later. Denote by $P_{(h^*,s^*,a^*)}$ and $E_{(h^*,s^*,a^*)}$ the probability measure and expectation, respectively, in the MDP $\mathcal{M}_{(h^*,s^*,a^*)}$. Let $\mathcal{M}_0$ be the MDP with $\Delta_0(h, \tilde{s}_i, a) = 0$ for all $(h, \tilde{s}_i, a) \in [H] \times \mathcal{L} \times \mathcal{A}$. Denote by P_0 and E_0 the probability measure and expectation, respectively, in the MDP $\mathcal{M}_0$.

- **Step 2: Computing the Expected Regret of an Algorithm in $\mathcal{M}_{(h^*,s^*,a^*)}$.**

  We now compute $Regret(\mathcal{M}_{(h^*,s^*,a^*)}, \textbf{algo}, K)$ for a learning algorithm **algo**, which executes policy π^k in episode $k \in [K]$. By (7), we need to compute the optimal value function, $V_1^*(s_1^k)$, and the value function under policy π^k, $V_1^{\pi^k}(s_1^k)$, for $k \in [K]$. Unlike Domingues et al. (2021), these quantities can not be computed explicitly in general in our risk-sensitive setting because the OCE is defined by an optimization problem. Hence, in the following, we will bound $V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)$ in order to lower bound the regret.

We first compute the value function $V_1^{\pi^k}(s_1^k)$ under policy π^k. For notational simplicity, we denote $\pi_h^k(s_h^k)$ as a_h^k for all $h \in [H], k \in [K]$. Under policy π^k, we use $\hat{H}$ to denote the number of time steps in which the agent stays in the "waiting" state, which is no larger than $\bar{H}$. Because there is no reward collected before step $\hat{H} + d$, we can obtain

$$V_1^{\pi^k}(s_1^k) = V_{\hat{H}+d}^{\pi^k}(s_{\hat{H}+d}^k). \tag{38}$$

Next, we compute $V_{\hat{H}+d}^{\pi^k}(s_{\hat{H}+d}^k)$. To this end, we first show

$$V_{\bar{H}+d+1}^{\pi^k}(s_{\bar{H}+d+1}^k) = (H - \bar{H} - d) \times 1\{s_{\bar{H}+d+1}^k = \tilde{s}_g\}. \tag{39}$$

We prove it by recursion. It is clear that

$$V_H^{\pi^k}(s_H^k) = r_H(s_H^k, a_H^k) + OCE_{s_{H+1}^k \sim P_H(\cdot|s_H^k, a_H^k)}(V_{H+1}^{\pi^k}(s_{H+1}^k)) \overset{(1)}{=} 1\{s_H^k = \tilde{s}_g\} \overset{(2)}{=} 1\{s_{\bar{H}+d+1}^k = \tilde{s}_g\},$$

where equality (1) holds because $V_{H+1}^{\pi^k}(s_{H+1}^k) = 0$, and equality (2) follows from the fact that the agent is in the absorbing states when $h \geq \bar{H} + d + 1$. Then, we can compute

$$V_{H-1}^{\pi^k}(s_{H-1}^k) = r_{H-1}(s_{H-1}^k, a_{H-1}^k) + OCE_{s_H^k \sim P_{H-1}(\cdot|s_{H-1}^k, a_{H-1}^k)}(V_H^{\pi^k}(s_H^k)) \overset{(3)}{=} 2 \times 1\{s_{\bar{H}+d+1}^k = \tilde{s}_g\},$$

where equality (3) holds because the random variable $1\{s_{\bar{H}+d+1}^k = \tilde{s}_g\}$ is known at step $H - 1$, and we use property (b) in Lemma A.1. Repeating this procedure until time step $h = \bar{H} + d + 1$, we obtain (39).

Given (39), we next compute the value function under policy π^k at time $\hat{H} + d + 1$. Note that at time step $\hat{H} + d$, the agent is at the leaf of the rooted tree, where $\hat{H} \leq \bar{H}$. Hence, the probability that the agent is at good state $\tilde{s}_g$ at step $h = \hat{H} + d + 1$ is the same as that at step $h = \bar{H} + d + 1$. In addition, the reward function is given by $r_h(s, a) = 1\{s = \tilde{s}_g, h \geq \bar{H} + d + 1\}, \forall a \in \mathcal{A}$. Hence, we obtain

$$\begin{aligned} V_{\hat{H}+d+1}^{\pi^k}(s_{\hat{H}+d+1}^k) = V_{\bar{H}+d+1}^{\pi^k}(s_{\bar{H}+d+1}^k) &= (H - \bar{H} - d) \times 1\{s_{\bar{H}+d+1}^k = \tilde{s}_g\} \\ &= (H - \bar{H} - d) \times 1\{s_{\hat{H}+d+1}^k = \tilde{s}_g\}. \end{aligned}$$

It then follows that

$$\begin{aligned} & V_1^{\pi^k}(s_1^k) = V_{\hat{H}+d}^{\pi^k}(s_{\hat{H}+d}^k) \\ & = r_{\hat{H}+d}(s_{\hat{H}+d}^k, a_{\hat{H}+d}^k) + OCE_{s_{\hat{H}+d+1}^k \sim P_{\hat{H}+d}(\cdot|s_{\hat{H}+d}^k, a_{\hat{H}+d}^k)}(V_{\hat{H}+d+1}^{\pi^k}(s_{\hat{H}+d+1}^k)) \\ & = OCE_{s_{\hat{H}+d+1}^k \sim P_{\hat{H}+d}(\cdot|s_{\hat{H}+d}^k, a_{\hat{H}+d}^k)}\Big((H - \bar{H} - d) \times 1\{s_{\hat{H}+d+1}^k = \tilde{s}_g\}\Big) \\ & = \sup_{\lambda \in [0, H - \bar{H} - d]} \{\lambda + P_{(h^*, s^*, a^*)}(s_{\hat{H}+d+1}^k = \tilde{s}_g) u(H - \bar{H} - d - \lambda) + (1 - P_{(h^*, s^*, a^*)}(s_{\hat{H}+d+1}^k = \tilde{s}_g)) u(-\lambda)\} \\ & = \sup_{\lambda \in [0, H - \bar{H} - d]} \{\lambda + P_{(h^*, s^*, a^*)}(s_{\bar{H}+d+1}^k = \tilde{s}_g) u(H - \bar{H} - d - \lambda) + (1 - P_{(h^*, s^*, a^*)}(s_{\bar{H}+d+1}^k = \tilde{s}_g)) u(-\lambda)\}. \end{aligned} \tag{40}$$

Similar to Equation (7) in Domingues et al. (2021), we can derive

$$\begin{aligned} & P_{(h^*, s^*, a^*)}(s_{\bar{H}+d+1}^k = \tilde{s}_g) \\ & = \sum_{h=1+d}^{\bar{H}+d} p P_{(h^*, s^*, a^*)}(s_h^k \in \mathcal{L}) + 1\{h = h^*\} P_{(h^*, s^*, a^*)}(s_h^k = s^*, a_h^k = a^*) \epsilon \\ & = p + \epsilon \cdot P_{(h^*, s^*, a^*)}(s_{h^*}^k = s^*, a_{h^*}^k = a^*). \end{aligned} \tag{41}$$

Together with (40), we obtain an expression of the value function $V_1^{\pi^k}(s_1^k)$.

We next compute the optimal value function $V_1^*(s_1^k)$. Based on (40), one can easily show that the optimal policy is to let $P_{(h^*,s^*,a^*)}(s_{h^*}^k = s^*, a_{h^*}^k = a^*) = 1$. Specifically, in the MDP $\mathcal{M}_{(h^*,s^*,a^*)}$, the optimal policy is to traverse the tree at step $h^* - d$, so that the agent visits the leaf state s^* at time step h^* and takes the action a^* at this leaf state. Thus, the optimal value function is given by

$$V_1^*(s_1^k) = \sup_{\lambda \in [0, H-\bar{H}-d]} \{\lambda + (p+\epsilon)u(H - \bar{H} - d - \lambda) + (1-p-\epsilon)u(-\lambda)\}.$$

Now, we can compute that for each episode $k \in [K]$,

$$\begin{aligned}
&V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \\
&= \sup_{\lambda \in [0, H-\bar{H}-d]} \{\lambda + (p+\epsilon)u(H - \bar{H} - d - \lambda) + (1-p-\epsilon)u(-\lambda)\} \\
&- \sup_{\lambda \in [0, H-\bar{H}-d]} \{\lambda + P_{(h^*,s^*,a^*)}(s_{\bar{H}+d+1}^k = \tilde{s}_g)u(H - \bar{H} - d - \lambda) + (1 - P_{(h^*,s^*,a^*)}(s_{\bar{H}+d+1}^k = \tilde{s}_g))u(-\lambda)\} \\
&\overset{(1)}{\geq} \rho + (p+\epsilon)u(H - \bar{H} - d - \rho) + (1-p-\epsilon)u(-\rho) \\
&- \rho - P_{(h^*,s^*,a^*)}(s_{\bar{H}+d+1}^k = \tilde{s}_g)u(H - \bar{H} - d - \rho) - (1 - P_{(h^*,s^*,a^*)}(s_{\bar{H}+d+1}^k = \tilde{s}_g))u(-\rho) \\
&\overset{(2)}{=} \epsilon[u(H - \bar{H} - d - \rho) - u(-\rho)] \times [1 - P_{(h^*,s^*,a^*)}(s_{h^*}^k = s^*, a_{h^*}^k = a^*)],
\end{aligned}$$

where inequality (1) holds by setting

$$\rho \in \operatorname*{arg\,max}_{\lambda \in [0, H-\bar{H}-d]} \{\lambda + P_{(h^*,s^*,a^*)}(s_{\bar{H}+d+1}^k = \tilde{s}_g)u(H - \bar{H} - d - \lambda) + (1 - P_{(h^*,s^*,a^*)}(s_{\bar{H}+d+1}^k = \tilde{s}_g))u(-\lambda)\}, \quad (42)$$

and equality (2) holds by applying (41).

Therefore, the regret of a learning algorithm **algo** in $\mathcal{M}_{(h^*,s^*,a^*)}$ can be lower bounded as follow:

$$\begin{aligned}
&Regret(\mathcal{M}_{(h^*,s^*,a^*)}, \textbf{algo}, K) \\
&= \sum_{k=1}^{K} E_{(h^*,s^*,a^*)}[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)] \\
&\geq \epsilon[u(H - \bar{H} - d - \rho) - u(-\rho)] \sum_{k=1}^{K} \left(1 - P_{(h^*,s^*,a^*)}(s_{h^*}^k = s^*, a_{h^*}^k = a^*)\right) \\
&= \epsilon[u(H - \bar{H} - d - \rho) - u(-\rho)] \left(K - E_{(h^*,s^*,a^*)}\left[N_{(h^*,s^*,a^*)}^K\right]\right), \qquad (43)
\end{aligned}$$

where

$$N_{(h^*,s^*,a^*)}^K = \sum_{k=1}^{K} 1\{s_{h^*} = s^*, a_{h^*} = a^*\}. \qquad (44)$$

- **Step 3: Bounding Maximum Regret over all possible $\mathcal{M}_{(h^*,s^*,a^*)}$.**

  We can deduce from (43) that the maximum regret of an algorithm **algo** over all possible $\mathcal{M}_{(h^*,s^*,a^*)}$ is lower bounded by

$$\begin{aligned}
&\max_{(h^*,s^*,a^*)\in\mathcal{Z}} Regret(\mathcal{M}_{(h^*,s^*,a^*)}, \textbf{algo}, K) \\
&\geq \frac{1}{\bar{H}LA} \sum_{(h^*,s^*,a^*)\in\mathcal{Z}} Regret(\mathcal{M}_{(h^*,s^*,a^*)}, \textbf{algo}, K) \\
&\geq K[u(H - \bar{H} - d - \rho) - u(-\rho)]\epsilon \left(1 - \frac{1}{K\bar{H}LA} \sum_{(h^*,s^*,a^*)} E_{(h^*,s^*,a^*)}\left[N_{(h^*,s^*,a^*)}^K\right]\right).
\end{aligned} \qquad (45)$$

So to lower bound the regret, we have to upper bound $\sum_{(h^*,s^*,a^*)\in\mathcal{Z}} E_{(h^*,s^*,a^*)}\left[N_{(h^*,s^*,a^*)}^K\right]$.

- **Step 4: Bounding $\sum_{(h^*,s^*,a^*)\in\mathcal{Z}} E_{(h^*,s^*,a^*)}\left[N^K_{(h^*,s^*,a^*)}\right]$.**

  For this step, we use similar arguments to those used in Domingues et al. (2021); see page 13 therein. Fix $(h^*, s^*, a^*) \in [H] \times \mathcal{S} \times \mathcal{A}$. Because $\frac{1}{K} N^K_{(h^*,s^*,a^*)} \in [0,1]$, one can obtain from Lemma 1 of Garivier et al. (2019) that

$$kl\left(\frac{1}{K}E_0\left[N^K_{(h^*,s^*,a^*)}\right], \frac{1}{K}E_{(h^*,s^*,a^*)}\left[N^K_{(h^*,s^*,a^*)}\right]\right) \le KL(P_0, P_{(h^*,s^*,a^*)}),$$

  where KL denotes the Kullback-Leibler divergence between two probability measures and $kl(p,q)$ denotes the KL divergence between two Bernoulli distributions with success probabilities p and q respectively; see Definition 4 in Domingues et al. (2021). It then follows from Pinsker's inequality, $(p-q)^2 \le \frac{1}{2}kl(p,q)$, that

$$\frac{1}{K}E_{(h^*,s^*,a^*)}\left[N^K_{(h^*,s^*,a^*)}\right] \le \frac{1}{K}E_0\left[N^K_{(h^*,s^*,a^*)}\right] + \sqrt{\frac{1}{2}KL(P_0, P_{(h^*,s^*,a^*)})}.$$

  Because $\mathcal{M}_0$ and $\mathcal{M}_{(h^*,s^*,a^*)}$ differ at stage h^* when $(s_{h^*}, a_{h^*}) = (s^*, a^*)$, by Lemma 5 of Domingues et al. (2021) and Lemma C.1 in Appendix C.1, we can prove that

$$KL\left(P_0, P_{(h^*,s^*,a^*)}\right) = E_0\left[N^K_{(h^*,s^*,a^*)}\right] kl(p, p+\epsilon) \le E_0\left[N^K_{(h^*,s^*,a^*)}\right]\frac{c_1\epsilon^2}{p},$$

  where $c_1 \ge 2$ is a certain positive constant, $p \in [0, 1-\frac{1}{c_1}]$ and ϵ satisfies

$$\epsilon \in \left[0, \frac{(1-2p) + \sqrt{1-\frac{4p}{c_1}}}{2}\right]. \tag{46}$$

  Thus,

$$\frac{1}{K}E_{(h^*,s^*,a^*)}\left[N^K_{(h^*,s^*,a^*)}\right] \le \frac{1}{K}E_0\left[N^K_{(h^*,s^*,a^*)}\right] + \sqrt{\frac{c_1}{2p}}\epsilon\sqrt{E_0\left[N^K_{(h^*,s^*,a^*)}\right]}.$$

  According to the definition of $N^K_{(h^*,s^*,a^*)}$ in (44), we know that $\sum_{(h^*,s^*,a^*)\in\mathcal{Z}} N^K_{(h^*,s^*,a^*)} \le K$. Then, by Cauchy-Schwarz inequality, we have

$$\frac{1}{K}\sum_{(h^*,s^*,a^*)\in\mathcal{Z}} E_{(h^*,s^*,a^*)}\left[N^K_{(h^*,s^*,a^*)}\right] \le 1 + \sqrt{\frac{c_1}{2p}}\epsilon\sqrt{\bar{H}LAK}. \tag{47}$$

- **Step 5: Optimizing ϵ and Choosing $\bar{H}$ and p.**

  By combining (45) with (47), we have

$$\begin{aligned}&\max_{(h^*,s^*,a^*)\in\mathcal{Z}} Regret(\mathcal{M}_{(h^*,s^*,a^*)}, \textbf{algo}, K)\\ &\ge K[u(H-\bar{H}-d-\rho) - u(-\rho)]\epsilon\left(1 - \frac{1}{\bar{H}LA} - \sqrt{\frac{c_1}{2p}}\epsilon\frac{\sqrt{\bar{H}LAK}}{\bar{H}LA}\right),\end{aligned} \tag{48}$$

  where the right-hand side of the inequality is a quadratic function of ϵ. Maximizing this function by taking

$$\epsilon = \sqrt{\frac{p}{2c_1}}\left(1 - \frac{1}{\bar{H}LA}\right)\sqrt{\frac{\bar{H}LA}{K}}, \tag{49}$$

  we derive

$$\begin{aligned}&\max_{(h^*,s^*,a^*)\in\mathcal{Z}} Regret(\mathcal{M}_{(h^*,s^*,a^*)}, \textbf{algo}, K)\\ &\ge \frac{1}{2\sqrt{2}}\sqrt{\frac{p}{c_1}}[u(H-\bar{H}-d-\rho) - u(-\rho)]\sqrt{\bar{H}LAK}(1 - \frac{1}{\bar{H}LA})^2.\end{aligned} \tag{50}$$

According to Assumption 4.2, we have $A \geq 2$, $S \geq 6$, and thus $L = (1 - \frac{1}{A})(S-3) + \frac{1}{A} \geq \frac{S}{4}$. Then, we can deduce from (50) that

$$\begin{aligned}
&\max_{(h^*,s^*,a^*)\in\mathcal{Z}} Regret(\mathcal{M}_{(h^*,s^*,a^*)}, \textbf{algo}, K) \\
&\geq \frac{1}{2\sqrt{2}} \cdot \sqrt{\frac{p}{c_1}}[u(H - \bar{H} - d - \rho) - u(-\rho)]\sqrt{\bar{H} \cdot \frac{S}{4}AK} \cdot \frac{4}{9} \\
&= \frac{1}{9\sqrt{2}} \cdot \sqrt{\frac{p}{c_1}}[u(H - \bar{H} - d - \rho) - u(-\rho)]\sqrt{SA\bar{H}K}. \qquad (51)
\end{aligned}$$

The bound in (51) is not explicit in the sense that the quantity ρ defined in (42) depends on the unknown probability $P_{(h^*,s^*,a^*)}(s^k_{\bar{H}+d+1} = \tilde{s}_g)$, which equals $p + \epsilon \cdot P_{(h^*,s^*,a^*)}(s^k_{h^*} = s^*, a^k_{h^*} = a^*)$. We next lower bound the term on the right-hand-side of (51) in order to derive the explicit bound given in Theorem 4.3.

Because ρ satisfies (42), by the first-order optimality condition, we have

$$1 \in P_{(h^*,s^*,a^*)}(s^k_{\bar{H}+d+1} = \tilde{s}_g)\partial u(H - \bar{H} - d - \rho) + (1 - P_{(h^*,s^*,a^*)}(s^k_{\bar{H}+d+1} = \tilde{s}_g))\partial u(-\rho). \qquad (52)$$

According to Assumption 4.2, we have $H \geq c_2 d$ with $c_2 > 2$. We choose

$$\bar{H} = \frac{H}{c_2}. \qquad (53)$$

Then, by the monotonicity of the supergradients of the concave function u, we obtain from (52) that

$$1 \leq P_{(h^*,s^*,a^*)}(s^k_{\bar{H}+d+1} = \tilde{s}_g)\partial u\left(\left(1 - \frac{2}{c_2}\right)H - \rho\right) + (1 - P_{(h^*,s^*,a^*)}(s^k_{\bar{H}+d+1} = \tilde{s}_g))\partial u(-\rho), \qquad (54)$$

where the inequality means every element in the set on the right-hand-side is greater than one. Recall that $P_{(h^*,s^*,a^*)}(s^k_{\bar{H}+d+1} = \tilde{s}_g) = p + \epsilon \cdot P_{(h^*,s^*,a^*)}(s^k_{h^*} = s^*, a^k_{h^*} = a^*)$. Then, we can deduce from (52) that

$$1 \leq p \cdot \partial u\left(\left(1 - \frac{2}{c_2}\right)H - \rho\right) + (1-p) \cdot \partial u(-\rho), \qquad (55)$$

where we use the fact that ∂u is monotone so that all elements in the set $\partial u((1 - \frac{2}{c_2})H - \rho) - \partial u(-\rho)$ are all non-negative. Now consider the function $p \cdot \partial u((1 - \frac{2}{c_2})H - \lambda) + (1-p) \cdot \partial u(-\lambda)$ for $\lambda \in [0, \rho]$. When $\lambda = 0$, it is clear that $p \cdot \partial u((1 - \frac{2}{c_2})H) + (1-p) \cdot \partial u(0)$ contains an element that is smaller than one, because $1 \in \partial u(0)$ and the elements in $\partial u((1 - \frac{2}{c_2})H)$ are smaller than one. Together with (55) and the continuity of the superdifferential mapping, we then deduce that there exists some $\lambda^* \in [0, \rho]$ such that

$$1 \in p \cdot \partial u\left(\left(1 - \frac{2}{c_2}\right)H - \lambda^*\right) + (1-p) \cdot \partial u(-\lambda^*). \qquad (56)$$

Now, we are ready to lower bound the right-hand-side of (51). Note that $u(H - \bar{H} - d - \lambda) - u(-\lambda)$ is nondecreasing in $\lambda \in [0, H - \bar{H} - d]$. Using (53) and the assumption $H \geq 2c_2 d$, we then have

$$u(H - \bar{H} - d - \rho) - u(-\rho) \geq u\left(\left(1 - \frac{2}{c_2}\right)H - \lambda^*\right) - u(-\lambda^*). \qquad (57)$$

For fixed $c_1 \geq 4$, we can choose

$$p = 1 - \frac{2}{c_1} \geq \frac{1}{2}. \qquad (58)$$

It follows from (51) and (53) that

$$\begin{aligned}
&\max_{(h^*,s^*,a^*)\in\mathcal{Z}} Regret(\mathcal{M}_{(h^*,s^*,a^*)}, \textbf{algo}, K) \\
&\geq \frac{1}{9\sqrt{2}} \cdot \sqrt{\frac{p}{c_1}}[u(H - \bar{H} - d - \rho) - u(-\rho)]\sqrt{SA\bar{H}K} \\
&\geq \frac{1}{18\sqrt{2c_1c_2}} \cdot \left[u\left(\left(1 - \frac{2}{c_2}\right)H - \lambda^*\right) - u(-\lambda^*)\right]\sqrt{SAHK}. \qquad (59)
\end{aligned}$$

Finally, we need to make ϵ in (49) satisfy the constraint (46). It is easy to check that $\epsilon \leq \sqrt{\frac{HSA}{2c_1c_2K}}$. Moreover, we have $\frac{(1-2p)+\sqrt{1-\frac{4p}{c_1}}}{2} \geq \frac{1}{c_1}$. Hence, we can choose $K \geq \frac{c_1 HSA}{2c_2}$ to make ϵ in (49) feasible. The proof is therefore completed.

□

C.1. An Auxiliary Lemma and Its Proof

Recall that for any $p, q \in (0,1)$ with $p+q=1$, $kl(p,q)$ denotes the KL divergence between two Bernoulli distributions with success probabilities p and q respectively, i.e.,

$$kl(p,q) = p \log\left(\frac{p}{q}\right) + q \log\left(\frac{1-p}{1-q}\right).$$

Lemma C.1. *Fix any constant $c_1 \geq 2$. If $p \in [0, 1-\frac{1}{c_1}]$ and $\epsilon \in \left[0, \frac{(1-2p)+\sqrt{1-\frac{4p}{c_1}}}{2}\right]$, then we have $kl(p, p+\epsilon) \leq \frac{c_1\epsilon^2}{p}$.*

Proof. Using the inequality $\log(1+x) \leq x$ for any $x > -1$, we have

$$\begin{aligned} kl(p,p+\epsilon) &= p\log\left(\frac{p}{p+\epsilon}\right) + (1-p)\log\left(\frac{1-p}{1-p-\epsilon}\right) \\ &\leq p\left(\frac{p}{p+\epsilon} - 1\right) + (1-p)\left(\frac{1-p}{1-p-\epsilon} - 1\right) \\ &= \frac{\epsilon^2}{(p+\epsilon)(1-p-\epsilon)} \\ &\overset{(1)}{\leq} \frac{c_1\epsilon^2}{p}, \end{aligned}$$

where inequality (1) holds if we have

$$\frac{p}{c_1} \leq p(1-p) + (1-2p)\epsilon - \epsilon^2.$$

One can easily verify that the above inequality holds if $p \in [0, 1-\frac{1}{c_1}]$ and $\epsilon \in \left[0, \frac{(1-2p)+\sqrt{1-\frac{4p}{c_1}}}{2}\right]$. The proof is then completed. □

D. Numerical Experiments

In this section, we conduct numerical experiments to illustrate the performance of the OCE-VI algorithm on randomly generated MDPs.

We adopt the methods in Dann (2019, Section 4.7) to randomly generate MDPs with state space $\mathcal{S} = \{1, \cdots, S\}$, action space $\mathcal{A} = \{1, \cdots, A\}$ and episode length H. For each $h = 1, 2, \ldots, H$, the transition probabilities $P_h(\cdot|s,a)$ are generated independently from the Dirichlet distribution $Dir(0.1, \cdots, 0.1)$. Reward functions $r_h(s,a)$ are set to 0 with probability 85% and generated independently from the uniform distribution $U[0,1]$ with probability 15%. In comparing the performance of different learning algorithms, we assume that the reward functions are known, but the transition probabilities are unknown.

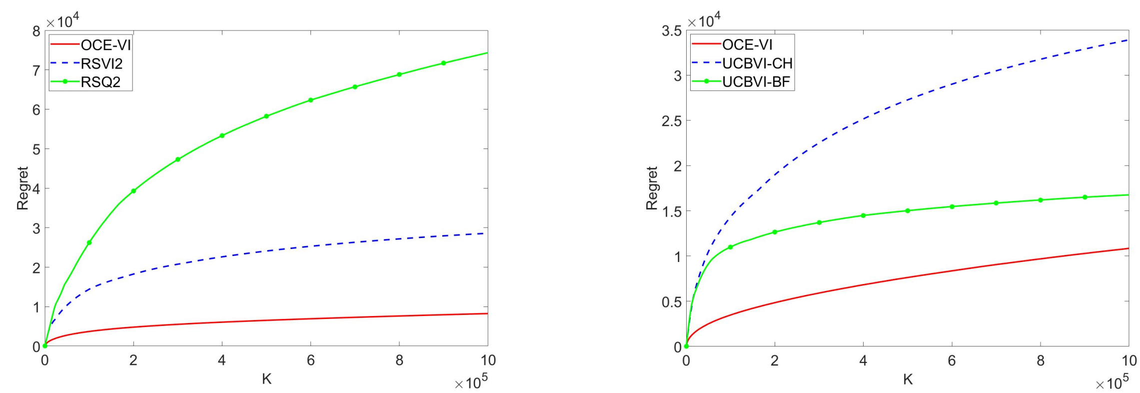
In our experiments we consider two different OCEs[3]: entropic risk and mean-variance models. For entropic risk, we compare the performance of our OCE-VI algorithm with the RSVI2 and RSQ2 algorithms in Fei et al. (2021a). For mean-variance models, because there is no existing benchmark algorithm in the episodic RL setting with recursive mean-variance criterion, we compare our OCE-VI algorithm with the UCBVI-CH (with Chernoff-Hoeffding bonus) and UCBVI-BF (with Bernstein

[3]For CVaR, our OCE-VI algorithm is essentially the ICVaR algorithm (Du et al., 2022), so we do not compare their performances in the experiments.

bonus) algorithms in Azar et al. (2017) designed for the risk-neutral episodic RL. While the original UCBVI algorithms in Azar et al. (2017) are developed for MDPs with stationary transitions, we adapt them to our non-stationary MDP setting with time-dependent transition probabilities.

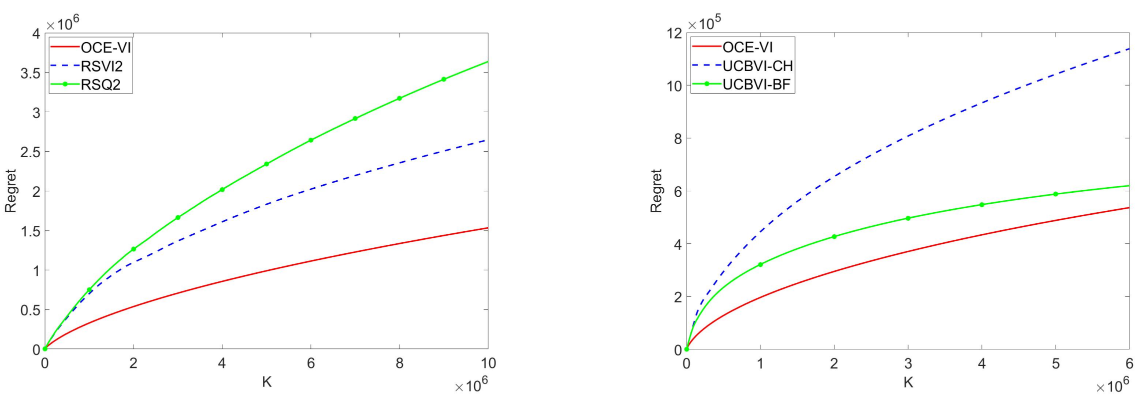
We consider two sets of parameters. The first one is $(H, S, A) = (3, 6, 3)$, and we use the risk-aversion parameter $\beta = -0.6$ for the entropic risk and $c = \frac{1}{6}$ for the mean-variance models. We set $K = 10^6$ and $\delta = \frac{1}{2KH}$ for all algorithms. The second one is $(H, S, A) = (6, 20, 3)$, and we use $\beta = -0.6$ for the entropic risk and $c = \frac{1}{12}$ for the mean-variance models. Because the size of the MDP becomes larger and learning can be more difficult in the second setting, we consider $K = 10^7$ to show the sublinear regret (in K) of algorithms.

Figures 2 and 3 illustrate the performance comparisons of the OCE-VI algorithm with other algorithms, where we plot the average regret of each algorithm as a function of the number of episodes K. We compute the expected regret of each algorithm by averaging over 30 independent runs, but we do not plot the confidence intervals since the confidence intervals estimated from the 30 samples are very narrow compared with the magnitude of the regret and are almost invisible in the figures. We can observe from Figures 2 and 3 that for episodic RL with recursive entropic risk, our algorithm can outperform the RSVI2 algorithm in Fei et al. (2021a) on randomly generated MDPs in the same risk-sensitive RL setting. For episodic RL with recursive mean-variance models, we find that our algorithm performs better than UCBVI algorithms in Azar et al. (2017), though this is not surprising given that UCBVI is designed for the risk-netural RL setting.



(a) (b)

Figure 2. Performance comparison of OCE-VI algorithm with other algorithms on a randomly generated MDP with $(H, S, A) = (3, 6, 3)$. Figure 2a is for episodic RL with recursive entropic risk and Figure 2b is for the mean-variance models.



(a) (b)

Figure 3. Performance comparison of OCE-VI algorithm with other algorithms on a randomly generated MDP with $(H, S, A) = (6, 20, 3)$. Figure 3a is for episodic RL with recursive entropic risk and Figure 3b is for the mean-variance models.