# Non-stationary Risk-Sensitive Reinforcement Learning: Near-Optimal Dynamic Regret, Adaptive Detection, and Separation Design

**Yuhao Ding**[1], **Ming Jin**[2], **Javad Lavaei**[1]

[1] UC Berkeley, Department of Industrial Engineering and Operations Research
[2] Virginia Tech, Department of Electrical and Computer Engineering
[1]{yuhao_ding, lavaei}@berkeley.edu, [2]jinming@vt.edu

**Abstract**

We study risk-sensitive reinforcement learning (RL) based on an entropic risk measure in episodic non-stationary Markov decision processes (MDPs). Both the reward functions and the state transition kernels are unknown and allowed to vary arbitrarily over time with a budget on their cumulative variations. When this variation budget is known a prior, we propose two restart-based algorithms, namely Restart-RSMB and Restart-RSQ, and establish their dynamic regrets. Based on these results, we further present a meta-algorithm that does not require any prior knowledge of the variation budget and can adaptively detect the non-stationarity on the exponential value functions. A dynamic regret lower bound is then established for non-stationary risk-sensitive RL to certify the near-optimality of the proposed algorithms. Our results also show that the risk control and the handling of the non-stationarity can be separately designed in the algorithm if the variation budget is known a prior, while the non-stationary detection mechanism in the adaptive algorithm depends on the risk parameter. This work offers the first non-asymptotic theoretical analyses for the non-stationary risk-sensitive RL in the literature.

## 1 Introduction

Risk-sensitive RL considers problems in which the objective takes into account risks that arise during the learning process, in contrast to the typical expected accumulated reward objective. Effective management of the variability of the return in RL is essential in various applications in finance (Markowitz 1968), autonomous driving (Garcıa and Fernández 2015) and human behavior modeling (Niv et al. 2012).

While classical risk-sensitive RL assumes that an agent interacts with a time-invariant (stationary) environment, both the reward functions and the transition kernels can be time-varying for many risk-sensitive applications. For example, in finance (Markowitz 1968), the federal reserve adjusts the interest rate or the balance sheet in a non-stationary way and the market participants should adjust their trading policies accordingly. In the medical treatments (Man et al. 2014), the patient's health condition and the sensitivity of the patient's internal body organs to the medicine vary over time. This non-stationarity should be accounted for to minimize the risk of any potential side effects of the treatment. A similar requirement holds for the power grid control (Ding, Lavaei, and Arcak 2021) where the power grid contingency needs to be prepared with the time-varying electricity loads.

Despite the importance and ubiquity of non-stationary risk-sensitive RL problems, the literature lacks provably efficient algorithms and theoretical results. In this work, we study risk-sensitive RL with an entropic risk measure (Howard and Matheson 1972) under episodic Markov decision processes with unknown and time-varying reward functions and state transition kernels.

The non-stationary RL problem with an entropic risk measure has the following technical challenges. (1) Due to the non-stationarity of the model, any estimation error of the expectation operator may be tremendously amplified in the value function when the risk parameter $\beta$ is small. (2) In addition, the exponential Bellman equation (see Equation (3)) used in our risk-sensitive analysis associates the instantaneous reward and value function of the next step in a multiplicative way (Fei et al. 2021). However, this multiplicative feature of the exponential Bellman equation will also involve the policy evaluation errors due to the non-stationary drifting as multiplicative terms, which makes it more difficult to gauge the bounds than the risk-neural non-stationary setting in which all policy evaluation errors are in an additive way. (3) Furthermore, the non-linearity of the objective function (see Equation (1)) makes it difficult to obtain an unbiased estimation of the value function, which is needed in the design of a non-stationary detection mechanism in risk-neutral non-stationary RL (Wei and Luo 2021). (4) It is unclear whether the risk control and the handling of the non-stationarity can be separately designed when achieving the optimal dynamic regret. To address these difficulties, we develop a novel analysis to carefully quantify the effect of the non-stationarity in risk-sensitive RL. Our main theoretical contributions, summarized in Table 1, are as follows

- When the variation budget is known a prior, we propose two provably efficient restart algorithms, namely Restart-RSMB and Restart-RSQ, and establish their dynamic regrets. The stationary version of the model-based method Restart-RSMB is also the first model-based risk-sensitive algorithm in the stationary setting in the literature.
- When the variation budget is unknown (parameter-free), we propose a meta-algorithm that adaptively detects the non-stationarity of the exponential value functions. The

proposed adaptive algorithms, namely Adaptive-RSMB and Adaptive-RSQ, can achieve the (almost) same dynamic regret as the algorithms requiring the knowledge of the variation budget.

- We establish a lower bound result for non-stationary RL with entropic risk measure that certifies the near-optimality of our upper bounds.
- Our results also show that the risk control and the handling of the non-stationarity can be separately designed if the variation budget is known a prior, while the non-stationary detection mechanism in the adaptive algorithms depends on the risk parameter.

### 1.1 Related Work

**Non-stationary RL.** Non-stationary RL has been mostly studied in the risk-neutral setting. When the variation budget is known a prior, a common strategy for adapting to the non-stationarity is to follow the forgetting principle, such as the restart strategy (Mao et al. 2020; Zhou et al. 2020; Zhao et al. 2020; Ding and Lavaei 2022), exponential decayed weights (Touati and Vincent 2020), or sliding window (Cheung, Simchi-Levi, and Zhu 2020; Zhong, Yang, and Szepesvári 2021). In this work, we focus on the restart method mainly due to its advantage of the simplicity of the the memory efficiency (Zhao et al. 2020) and generalize it to the risk-sensitive RL setting. However, the prior knowledge of the variation budget is often unavailable in practice. The work (Cheung, Simchi-Levi, and Zhu 2020) develop a Bandit-over-Reinforcement-Learning framework to relax this assumption, but it leads to the suboptimal regret. To achieve a nearly-optimal regret without the prior knowledge of the variation budget, (Auer, Gajane, and Ortner 2019) and (Chen et al. 2019) maintain a distribution over bandit arms with properly controlled variance for all reward estimators. For RL problems, the seminar work (Wei and Luo 2021) proposes a black-box reduction approach that turns a certain RL algorithm with optimal regret in a (near-)stationary environment into another algorithm with optimal dynamic regret in a non-stationary environment. However, the above works only consider risk-neutral RL and may not apply to the more general risk-sensitive RL problems.

**Risk-sensitive RL.** Many risk-sensitive objectives have been investigated in the literature and applied to RL, such as the entropic risk measure, Markowitz mean-variance model, Value-at-Risk (VaR), and Conditional Value at Risk (CVaR) (Moody and Saffell 2001; Chow and Ghavamzadeh 2014; Delage and Mannor 2010; La and Ghavamzadeh 2013; Di Castro, Tamar, and Mannor 2012; Tamar, Glassner, and Mannor 2015; Tamar et al. 2015; Howard and Matheson 1972). Our work is closely related to the entropic risk measure. Following the seminal paper (Howard and Matheson 1972), this line of work includes (Bäuerle and Rieder 2014; Borkar 2001; Borkar and Meyn 2002; Borkar 2002; Cavazos-Cadena and Fernández-Gaucherand 2000; Coraluppi and Marcus 1999; Di Masi and Stettner 1999; Fernández-Gaucherand and Marcus 1997; Fleming and McEneaney 1995; Hernández-Hernández and Marcus 1996; Osogami 2012; Fleming and McEneaney 1992; Shen, Stannat, and Obermayer 2013; Fei et al. 2020; Fei, Yang, and Wang 2021; Fei et al. 2021). In particular, when transitions are unknown and simulators of the environment are unavailable, the first non-asymptotic regret guarantees are established under the tabular setting in (Fei et al. 2020) and the function approximation setting in (Fei, Yang, and Wang 2021). Then, a simple transformation of the risk-sensitive Bellman equations is proposed in (Fei et al. 2021), which leads to improved regret upper bounds. However, the above papers all assume that the environment is stationary, and therefore their results may quickly collapse in a non-stationary environment.

## 2 Problem Formulation

### 2.1 Notations

For a positive integer $n$, let $[n] := \{1, 2, \ldots, n\}$. Given a variable $x$, the notation $a = \mathcal{O}(b(x))$ means that $a \le C \cdot b(x)$ for some constant $C > 0$ that is independent of $x$. Similarly, $a = \widetilde{\mathcal{O}}(b(x))$ indicates that the previous inequality may also depend on the function $\log(x)$, where $C > 0$ is again independent of $x$. In addition, the notation $a = \Omega(b(x))$ means that $a \ge C \cdot b(x)$ for some constant $C > 0$ that is independent of $x$.

### 2.2 Episodic MDP and Risk-Sensitive Objective

In this paper, we study risk-sensitive RL in non-stationary environments via episodic MDPs with adversarial bandit-information reward feedback and unknown adversarial transition dynamics. At each episode $m$, an episodic MDP is defined by the finite state space $\mathcal{S}$, the finite action space $\mathcal{A}$, a collection of transition probability measure $\{\mathcal{P}_h^m\}_{h=1}^H$ specifying the transition probability $\mathcal{P}_h^m(s' \mid s, a)$ from state $s$ to the next state $s'$ under action $a \in \mathcal{A}$, a collection of reward functions $\{r_h^m\}_{h=1}^H$ where $r_h^m : \mathcal{S} \times \mathcal{A} \to [0, 1]$ , and $H > 0$ as the length of episodes. In this paper, we focus on a bandit setting where the agent only observes the values of reward functions, i.e., $r_h^m(s_h^m, a_h^m)$ at the visited state-action pair $(s_h^m, a_h^m)$. We also assume that reward functions are deterministic to streamline the presentation, while our analysis readily generalizes to the setting where reward functions are random.

For simplicity, we assume the initial state $s_1^m$ to be fixed as $s_1$ in different episodes. We use the convention that the episode terminates when a state $s_{H+1}$ at step $H+1$ is reached, at which the agent does not take any further action and receives no reward.

A policy $\pi^m = \{\pi_h^m\}_{h \in [H]}$ of an agent is a sequence of functions $\pi_h^m : \mathcal{S} \to \mathcal{A}$, where $\pi_h^m(s)$ is the action that the agent takes in state $s$ at step $h$ at episode $m$. For each $h \in [H]$ and $m \in [M]$, we define the value function $V_h^{\pi,m} : \mathcal{S} \to \mathbb{R}$ of a policy $\pi$ as the expected value of the cumulative rewards the agent receives under a risk measure of exponential utility by executing $\pi$ starting from an arbitrary state at step $h$. Specifically, we have

$$
\begin{aligned}
&V_h^{\pi,m}(s) \\
:=&\frac{1}{\beta} \log \left\{ \mathbb{E}_{\pi, \mathcal{P}^m} \left[ \exp\left( \beta \sum_{i=h}^{H} r_i^m(s_i, a_i) \right) \mid s_h = s \right] \right\} \qquad (1)
\end{aligned}
$$

| Algorithm | D-Regret | Parameter-free | Model-free | Separation |
|---|---|---|---|---|
| Restart-RSMB | $\widetilde{\mathcal{O}}\left(e^{\lvert\beta\rvert H}\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^2M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$ | ✗ | ✗ | ✓ |
| Restart-RSQ | $\widetilde{\mathcal{O}}\left(e^{\lvert\beta\rvert H}\lvert\mathcal{S}\rvert^{\frac{1}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{9}{4}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$ | ✗ | ✓ | ✓ |
| Adaptive-RSMB | $\widetilde{\mathcal{O}}\left(e^{\lvert\beta\rvert H}\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^2M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$ | ✓ | ✗ | ✗ |
| Adaptive-RSQ | $\widetilde{\mathcal{O}}\left(e^{\lvert\beta\rvert H}\lvert\mathcal{S}\rvert^{\frac{1}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$ | ✓ | ✓ | ✗ |
| Lower bound | $\Omega\left(\frac{e^{\frac{2\lvert\beta\rvert H}{3}}-1}{\lvert\beta\rvert}\lvert\mathcal{S}\rvert^{\frac{1}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$ | N/A | N/A | N/A |

Table 1: We summarize the dynamic regrets and lower bound obtained in this paper. Here, $\beta$ is the risk parameter, $H$ is the horizon of each episode, $M$ is the total number of episodes, $B$ is the total variation measurement, and $|\mathcal{S}|$ and $|\mathcal{A}|$ are the cardinalities of the state and action spaces.

where the expectation $\mathbb{E}_{\pi,\mathcal{P}^m}$ is taken over the random state-action sequence $\{(x_i^m, a_i^m)\}_{i=h}^H$, the action $a_i^m$ follows the policy $\pi_i^m(\cdot \mid x_i^m)$, and the next state $x_{i+1}$ follows the transition dynamics $\mathcal{P}_i^m(\cdot \mid x_i^m, a_i^m)$. Here $\beta \neq 0$ is the risk parameter of the exponential utility: $\beta > 0$ corresponds to a risk-seeking value function, $\beta < 0$ corresponds to a risk-averse value function, and as $\beta \to 0$ the agent tends to be risk-neutral and we recover the classical value function $V_h^{\pi,m}(s) = \mathbb{E}_{\pi,\mathcal{P}^m}\left[\sum_{t=1}^H r_h^m(s_t, a_t) \mid s_0 = s\right]$ in standard RL.

We further define the action-value function $Q_h^{\pi,m}: \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, for each $h \in [H]$ and $m \in [M]$, which gives the expected value of the risk measured by the exponential utility when the agent starts from an arbitrary state-action pair and follows the policy $\pi$ afterwards; that is,

$$\begin{aligned}
&Q_h^{\pi,m}\\
:=&\frac{1}{\beta}\log\left\{\exp\left(\beta\cdot r_h^m(s,a)\right)\mathbb{E}\left[\exp\left(\beta\sum_{i=h}^H r_i^m(s_t,a_t)\right)\right.\right.\\
&\left.\left.\qquad\qquad\Big| s_h = s, a_h = a\right]\right\}\\
=&r_h^m(s,a) + \frac{1}{\beta}\log\left\{\mathbb{E}\left[\exp\left(\beta\sum_{i=h+1}^H r_i^m(s_t,a_t)\right)\right.\right.\\
&\left.\left.\qquad\qquad\Big| s_h = s, a_h = a\right]\right\}
\end{aligned}$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Under some mild regularity conditions (Bäuerle and Rieder 2014), for each episode $m$, there always exists an optimal policy, denoted as $\pi^{*,m}$, that yields the optimal value $V_h^{\pi^{*,m},m}(s) := \sup_\pi V_h^{\pi,m}(s)$ for all $(h,s) \in [H]\times\mathcal{S}$. For convenience, we denote $V_h^{\pi^{*,m},m}(s)$ as $V_h^{*,m}(s)$ when it is clear from the context.

## 2.3 Exponential Bellman Equation

For all $(s,a,h,m) \in \mathcal{S}\times\mathcal{A}\times[H]\times[M]$, the Bellman equation associated with $\pi$ is given by

$$Q_h^{\pi,m}(s,a) = r_h^m(s,a) + \frac{1}{\beta}\log\left\{\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\cdot V_{h+1}^{\pi,m}(s')}\right]\right\}, \tag{2a}$$

$$V_h^{\pi,m}(s) = Q_h^{\pi,m}(s,\pi(s)), \quad V_{H+1}^{\pi,m}(s) = 0. \tag{2b}$$

In Equation (2), it can be seen that the action value $Q_h^{\pi,m}$ of step $h$ is a non-linear function of the value function $V_{h+1}^{\pi,m}$ of the later step. Based on Equation (2), for $h \in [H]$ and $m \in [M]$, the Bellman optimality equation is given by

$$Q_h^{*,m}(s,a) = r_h^m(s,a) + \frac{1}{\beta}\log\left\{\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\cdot V_{h+1}^{*,m}(s')}\right]\right\},$$

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

It has been recently shown in (Fei et al. 2021) that under the risk-sensitive measurement, it is easier to analyze a simple transformation of the Bellman equation (by taking exponential on both sides of (2)), which is called *exponential Bellman equation*: for every policy $\pi$ and tuple $(s,a,h,m)$, we have

$$e^{\beta\cdot Q_h^{\pi,m}(s,a)} = \mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left(r_h^m(s,a)+V_{h+1}^{\pi,m}(s')\right)}\right]. \tag{3}$$

When $\pi = \pi^{*,m}$, we obtain the corresponding optimality equation

$$e^{\beta\cdot Q_h^{*,m}(s,a)} = \mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left(r_h^m(s,a)+V_{h+1}^{*,m}(s')\right)}\right]. \tag{4}$$

Note that Equation (3) associates the current and future cumulative utilities $\left(Q_h^{\pi,m}\text{ and }V_{h+1}^{\pi,m}\right)$ in a multiplicative way, rather than in an additive way as in the standard Bellman equations (2).

## 2.4 Non-stationarity and Variation Budget

In this work, we focus on a non-stationary environment where the transition function $P_h^m$ and reward functions $r_h^m$ can vary over the episodes. We measure the non-stationarity of the MDP over an interval $\mathcal{I}$ in terms of its variation in the reward functions and transition kernels:

$$B_{r,\mathcal{I}} := \sum_{m\in\mathcal{I}}\sum_{h=1}^H \sup_{s,a}\left|r_h^m(s,a) - r_h^{m+1}(s,a)\right|,$$

$$B_{\mathcal{P},\mathcal{I}} := \sum_{m\in\mathcal{I}}\sum_{h=1}^H \sup_{s,a}\left\|\mathcal{P}_h^m(\cdot \mid s,a) - \mathcal{P}_h^{m+1}(\cdot \mid s,a)\right\|_1.$$

Note that our definition of variation only imposes restrictions on the summation of non-stationarity across different episodes, and does not put any restriction on the difference between two steps in the same episode. We further let $B_r \coloneqq B_{r,[1,M]}$, $B_p \coloneqq B_{p,[1,M]}$, and $B \coloneqq B_r + B_p$, and assume $B > 0$.

### 2.5 Performance Metrics

Since both the reward and the transition dynamics vary over the episodes and are revealed only after a policy is decided, the agent aims to ensure the long-term optimality guarantee over some given period of episodes $M$. Suppose that the agent executes policy $\pi^m$ in episode $m$. We now define the dynamic regret as the difference between the total reward value of policy $\{\pi^{\star,m}\}_{m=1}^{M}$ and that of the agent's policy $\pi^m$ over $M$ episodes:

$$\text{D-Regret}(M) \coloneqq \sum_{m=1}^{M} \left( V_1^{\star,m} - V_1^{\pi^m,m} \right).$$

## 3 Restart Algorithms with The Knowledge of Variation Budget

### 3.1 Periodically Restarted Risk-Sensitive Model-Based Method

We first present the Periodically Restarted Risk-sensitive Model-based method (Restart-RSMB) in Algorithm 1. It consists of two main stages: estimation of value function (line 7-13) with the periodical restart (line 5) and the policy execution (line 15).

To estimate the value function under the unknown non-stationarity, we take the optimistic value evaluation to properly handle the exploration-exploitation trade-off and apply the restart strategy to adapt to the unknown non-stationarity. In particular, we reset the visitation counters $N_h^m(s,a,s')$ and $N_h^m(x,a)$ to zero every $W$ episodes (line 5). Then, the reward and transition dynamics are estimated using only the data from the episode $\ell^m = (\lceil \frac{m}{W} \rceil - 1)W + 1$ to the episode $m$ by

$$\widehat{\mathcal{P}}_h^m(s' \mid s,a) = \frac{N_h^m(s,a,s') + \frac{\lambda}{|\mathcal{S}|}}{N_h^m(s,a)+\lambda}, \quad \text{for all } (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}, \tag{5a}$$

$$\widehat{r}_h^m(s,a) = \frac{\sum_{\tau=\ell^m}^{m-1} \mathbb{1}\{(s,a)=(s_h^\tau,a_h^\tau)\} r_h^\tau(s_h^\tau,a_h^\tau)}{N_h^m(s,a)+\lambda}, \quad \text{for all } (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{5b}$$

which are used to compute the estimated cumulative rewards at step $h$ (line 9). To encourage a sufficient exploration in the uncertain environment, Algorithm 1 applies the counter-based Upper Confidence Bound (UCB). Under the entropic risk measure, this bonus term takes the form

$$\begin{cases} C_1\left(\left(e^{\beta(H-h+1)}-1\right)+e^{\beta(H-h+1)}\beta\right)\sqrt{\frac{|\mathcal{S}|\log(6WH|\mathcal{S}||\mathcal{A}|/p)}{N_h^m(s,a)+1}}, & \text{if } \beta>0, \\ C_1\left(\left(1-e^{\beta(H-h+1)}\right)-\beta\right)\sqrt{\frac{|\mathcal{S}|\log(6WH|\mathcal{S}||\mathcal{A}|/p)}{N_h^m(s,a)+1}}, & \text{if } \beta<0, \end{cases} \tag{6}$$

Algorithm 1: Periodically Restarted Risk-sensitive Model-based RL (Restart-RSMB)

1: **Inputs:** Time horizon $M$, restart period $W$;
2: **for** $m = 1,\dots,M$ **do**
3: Set the initial state $x_1^m = x_1$ and $\ell^m = (\lceil \frac{m}{W} \rceil - 1)W+1$;
4: **if** $m = \ell^m$ **then**
5: $Q_h^m(s,a), V_h^m(s) \leftarrow H-h+1$ if $\beta > 0$, $Q_h^m(s,a), V_h^m(s) \leftarrow 0$ if $\beta < 0$, $N_h^m(s,a) \leftarrow 0, N_h^m(s,a,s') \leftarrow 0$ for all $(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$;
6: **end if**
7: **for** $h = H,\dots,1$ **do**
8: **for** $(s,a) \in \mathcal{S}\times\mathcal{A}$ **do**
9: $w_h^m(s,a) = \sum_{s'} \hat{\mathcal{P}}_h^m(s' \mid s,a)\left[e^{\beta[\hat{r}_h^m(s,a)+V_{h+1}^m(s')]}\right]$ where $\hat{\mathcal{P}}_h^m, \hat{r}_h^m$ are defined in (5);
10: $G_h^m(s,a) \leftarrow \begin{cases} \min\left\{e^{\beta(H-h+1)}, w_h^m(s,a)+\Gamma_h^m(s,a)\right\}, & \text{if } \beta>0; \\ \max\left\{e^{\beta(H-h+1)}, w_h^m(s,a)-\Gamma_h^m(s,a)\right\}, & \text{if } \beta<0; \end{cases}$ where $\Gamma_h^m$ is defined in (6);
11: $V_h^m(s) \leftarrow \max_{a'\in\mathcal{A}} \frac{1}{\beta}\log G_h^m(s,a')$;
12: **end for**
13: **end for**
14: **for** $h = 1,2,\dots,H$ **do**
15: Take an action $a_h^m \leftarrow \operatorname{argmax}_{a'\in\mathcal{A}} \frac{1}{\beta}\log\{G_h^m(s_h^m,a')\}$, and observe $r_h(s_h^m,a_h^m)$ and $s_{h+1}^m$;
16: $N_h^m(s_h^m,a_h^m) \leftarrow N_h^m(s_h^m,a_h^m)+1$; $N_h^m(s_h^m,a_h^m,s_{h+1}^m) \leftarrow N_h^m(s_h^m,a_h^m,s_{h+1}^m,)+1$;
17: **end for**
18: **end for**

for some constant $C_1 > 1$. Bonus terms of the form (6) are called "doubly decaying bonus" since they shrink deterministically and exponentially across the horizon steps due to the term $e^{\beta(H-h+1)}$, apart from decreasing in the visit count. We refer the reader to (Fei, Yang, and Wang 2021) for more discussion.

### 3.2 Periodically Restarted Risk-Sensitive Q-Learning

Next, we introduce Periodically Restarted Risk-sensitive Q-learning (Restart-RSQ) in Algorithm 2, which is model-free and inspired by RSQ2 in (Fei et al. 2021). Similar to Algorithm 1, we use the optimistic value evaluation to handle the exploration-exploitation trade-off and apply the restart strategy to adapt to the unknown non-stationarity. In particular, we re-initialize the value functions $Q_h^m(s,a), V_h^m(s)$ and reset the visitation counter $N_h^m(x,a)$ to zero every $W$ episodes (line 5). The algorithm then updates the exponential Q values using the Q-learning style update (line 11-12) for the state action pair that just visited (line 8). The learning rate $\alpha_t$ is defined as $\frac{H+1}{H+t}$, which is motivated by (Jin et al. 2018)

**Algorithm 2**: Periodically Restarted Risk-sensitive Q-learning (Restart-RSQ)

1: **Inputs:** Time horizon $M$, restart period $W$;
2: **for** $m = 1, \ldots, M$ **do**
3: Set the initial state $x_1^m = x_1$ and $\ell^m = (\lceil \frac{m}{W} \rceil - 1)W + 1$;
4: **if** $m = \ell^m$ **then**
5: $Q_h^m(s,a), V_h^m(s) \leftarrow H - h + 1$ if $\beta > 0$, $Q_h^m(s,a), V_h^m(s) \leftarrow 0$ if $\beta < 0$, $N_h^m(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ ;
6: **end if**
7: **for** $h = 1, 2, \ldots, H$ **do**
8: Take an action $a_h^m \leftarrow \text{argmax}_{a' \in \mathcal{A}} \frac{1}{\beta} \log\{G_h^m(s_h^m, a')\}$, and observe $r_h^m(s_h^m, a_h^m)$ and $s_{h+1}^m$;
9: $N_h^m(s_h^m, a_h^m) \leftarrow N_h^m(s_h^m, a_h^m) + 1$; $t \leftarrow N_h^m(s_h^m, a_h^m)$;
10: Set $\alpha_t = \frac{H+1}{H+t}$ and define $\Gamma_{h,t}^m(s_h^m, a_h^m)$ as in (7);
11: $w_h^m(s_h^m, a_h^m) = (1 - \alpha_t) \cdot G_h(s_h^m, a_h^m) + \alpha_t \cdot \left[ e^{\beta [r_h^m(s_h^m, a_h^m) + V_{h+1}^m(s')]} \right]$;
12: $G_h^m(s_h^m, a_h^m) \leftarrow$
$$\begin{cases} \min\left\{ e^{\beta(H-h+1)}, w_h^m(s_h^m, a_h^m) + \alpha_t \Gamma_{h,t}^m(s_h^m, a_h^m) \right\}, & \text{if } \beta > 0; \\ \max\left\{ e^{\beta(H-h+1)}, w_h^m(s_h^m, a_h^m) - \alpha_t \Gamma_{h,t}^m(s_h^m, a_h^m) \right\}, & \text{if } \beta < 0; \end{cases}$$
13: $V_h^m(s_h^m) \leftarrow \max_{a' \in \mathcal{A}} \frac{1}{\beta} \log G_h^m(s_h^m, a')$;
14: **end for**
15: **end for**

**Algorithm 3**: Risk-sensitive MALG with Stationary Tests and Restarts (Adaptive-ALG)

1: **Inputs:** ALG and its associated $\rho(\cdot)$, $\hat{n} = \log_2 M + 1$, $\hat{\rho}(m) = 6\hat{n} \log(\frac{M}{\delta}) \rho(m)$;
2: **for** $n = 0, 1, \ldots,$ **do**
3: Set $m_n \leftarrow m$ and run MALG-Initialization (see Appendix) for the block $[m_n, m_n + 2^n - 1]$;
4: **while** $m < m_n + 2^n$ **do**
5: Identify the unique active instance covering the episode $m$ and denote it as $alg$;
6: Construct the optimistic estimator $g_m$ for the active instance $alg$;
7: Follow $alg$'s decision $\pi_m$, receive estimated value $R_m = e^{\beta \sum_{h=1}^H r_h^m}$, and update $alg$;
8: Set $U_m = \begin{cases} \min_{\tau \in [m_n, m]} g_\tau, & \text{if } \beta > 0, \\ \max_{\tau \in [m_n, m]} g_\tau, & \text{if } \beta < 0; \end{cases}$
9: Perform **Test1** and **Test2**; Increment $t \leftarrow t + 1$;
10: **If** either test returns *fail*, **then** restart from Line 2.
11: **end while**
12: **end for**
13: **Test1**: Return *fail* if $m = alg.e$ for some order-k $alg$ and
$$\begin{cases} \frac{1}{2^k} \sum_{\tau = alg.s}^{alg.e} R_\tau - U_t \geq 9\hat{\rho}(2^k), & \text{if } \beta > 0, \\ U_t - \frac{1}{2^k} \sum_{\tau = alg.s}^{alg.e} R_\tau \geq 9\hat{\rho}(2^k), & \text{if } \beta < 0; \end{cases}$$
14: **Test2**: Return *fail* if
$$\begin{cases} \frac{1}{m - m_n + 1} \sum_{\tau = m_n}^{m} (g_\tau - R_\tau) \geq 3\hat{\rho}(m - m_n + 1), & \text{if } \beta > 0, \\ \frac{1}{m - m_n + 1} \sum_{\tau = m_n}^{m} (R_\tau - g_\tau) \geq 3\hat{\rho}(m - m_n + 1), & \text{if } \beta < 0, \end{cases}$$

and ensures that only the last $\mathcal{O}(\frac{1}{H})$ fraction of samples in each epoch is given non-negligible weights when used to estimate the optimistic Q-values under the non-stationarity. Algorithm 2 also applies the UCB by incorporating a "doubly decaying bonus" term that takes the form

$$\Gamma_{h,t}^m(s_h^m, a_h^m) \leftarrow C_2 \left| e^{\beta(H-h+1)} - 1 \right| \sqrt{\frac{|\mathcal{S}| \log(MH|\mathcal{S}||\mathcal{A}|/\delta)}{t}} \quad (7)$$

for some constant $C_2 > 1$.

### 3.3 Theoretical Results and Discussions

We now present our main theoretical results for Algorithms 1 and 2.

**Theorem 3.1** *For every $\delta \in (0,1]$, with probability at least $1 - \delta$ there exists a universal constant $c_1 > 0$ (used in Algorithm 1) such that the dynamic regret of Algorithm 1 with $W = M^{\frac{2}{3}} B^{-\frac{2}{3}} |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}}$ is bounded by*

$$\text{D-Regret(M)} \leq \widetilde{\mathcal{O}}\left( e^{|\beta|H} |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}} H^2 M^{\frac{2}{3}} B^{\frac{1}{3}} \right).$$

**Theorem 3.2** *For every $\delta \in (0,1]$, with probability at least $1 - \delta$ there exists a universal constant $c_2 > 0$ (used in Algorithm 2) such that the dynamic regret of Algorithm 2 with $W = M^{\frac{2}{3}} H^{-\frac{3}{4}} B^{-\frac{2}{3}} |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}}$ is bounded by*

$$\text{D-Regret(M)} \leq \widetilde{\mathcal{O}}\left( e^{|\beta|H} |\mathcal{S}|^{\frac{1}{3}} |\mathcal{A}|^{\frac{1}{3}} H^{\frac{9}{4}} M^{\frac{2}{3}} B^{\frac{1}{3}} \right).$$

The proofs of the two theorems are provided in Appendices. Note that the above results generalize those in the literature of risk-neutral non-stationary RL. In particular, when $\beta \to 0$, we recover the regret bounds with the same dependence on $M$ and $B$ for the restart model-based RL (Domingues et al. 2021) and restart Q-learning (Mao et al. 2020).

## 4 Adaptive Algorithm without The Knowledge of Variation Budget

In Theorems 3.1 and 3.2, we need to set the restart period to $W = \mathcal{O}(B^{-\frac{2}{3}} M^{\frac{2}{3}})$, which clearly requires the variation budget $B$ in advance. To overcome this limitation, we propose a meta-algorithm that adaptively detects the non-stationarity without the knowledge of $B$, while still achieving the similar dynamic regret as in Theorems 3.1 and 3.2. In particular, we generalize the black-box approach (Wei and Luo 2021) to the risk-sensitive RL setting and design a non-stationarity detection based on the exponential Bellman equations (3).

### 4.1 Risk-Sensitive Non-Stationary Detection

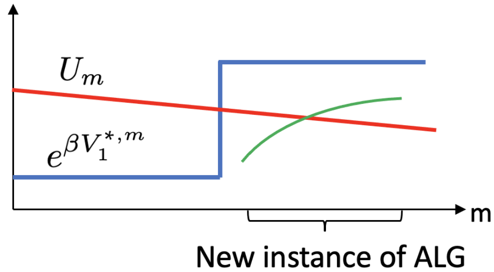
We first sketch the high-level idea of the black-box reduction approach for risk-sensitive non-stationary RL with $\beta > 0$.

Note that the dynamic regret can be bounded and decomposed as follows:

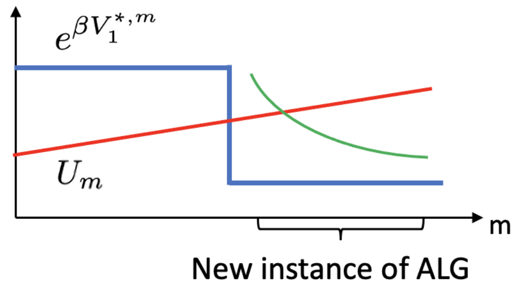
$$\text{D-Regret}(M) \le \underbrace{\frac{1}{\beta}\sum_{m=1}^{M}\left(e^{\beta V_1^{*,m}} - e^{\beta \widetilde{V}_1^{m}}\right)}_{\mathbf{R1}} + \underbrace{\frac{1}{\beta}\sum_{m=1}^{M}\left(e^{\beta \widetilde{V}_1^{m}} - e^{\beta V_1^{\pi^m,m}}\right)}_{\mathbf{R2}} \quad (8)$$

where $\widetilde{V}_1^m$ is an UCB-based optimistic estimator of the value function as constructed in Algorithms 1 and 2. In a stationary environment with $\beta > 0$, the base algorithms, such as Algorithms 1 and 2 without the restart mechanism (that is, $W = M$), ensure that **R1** is simply non-positive and **R2** is bounded by $\widetilde{\mathcal{O}}(M^{\frac{1}{2}})$. However, in a non-stationary environment, both terms can be substantially larger. Thus, if we can detect the event that either of the two terms is abnormally larger than the promised bound for a stationary environment, we learn that the environment has changed substantially and should restart the base algorithm. This detection can be easily performed for **R2** since both $e^{\beta \widetilde{V}_1^m}$ and $e^{\beta V_1^{\pi^m,m}}$ are observable [1], but not for **R1** since $V_1^{*,m}$ is unknown. To address this issue, we fully utilize the fact that $e^{\beta \widetilde{V}_1^m}$ is a UCB-based optimistic estimator to facilitate non-stationary detection.

We illustrate the idea of non-stationary detection for risk-sensitive RL in Figure 1. Here, the value of $V_1^{*,m}$ drastically increases which results to an increase in $e^{\beta V_1^{*,m}}$ for $\beta > 0$ and an decrease in $e^{\beta V_1^{*,m}}$ for $\beta < 0$. If we start running another instance of base algorithm after this environment change, then its performance will gradually approach due to its regret guarantee in a stationary environment. Since the optimistic estimators should always be an upper bound of the learner's average performance in a stationary environment for $\beta > 0$ or a lower bound of the learner's average performance in a stationary environment for $\beta < 0$, if, at some point, we find that the new instance of the base algorithm significantly outperformances/underperformances (depending on the value of $\beta$) this quantity, we can infer that the environment has changed.




(a) $\beta > 0$.




(b) $\beta < 0$.

Figure 1: An illustration of the risk-sensitive non-stationarity detection. The green curves represent the learner's average performance in new ALG. Since both $U_m$ and learner's average performance depend on the risk-sensitive parameter $\beta$ in a non-linear way. The non-stationarity detection relies on the choice of $\beta$ and thus the risk control and the handling of the non-stationarity can not be separately designed.

## 4.2 Multi-Scale ALG (MALG) and Non-Stationarity Tests

To detect the non-stationarity at different scales, we schedule and run instances of the base algorithm ALG in a randomized and multi-scale manner. In particular, Adaptive-ALG runs MALG in a sequence of blocks with doubling lengths. Within each block, Adaptive-ALG first initializes a MALG schedule (see Appendix), and then interacts the unique active instance at each episode with the environment (lines 5-7 in Algorithm 3). At the end of each episode, Adaptive-ALG performs two non-stationarity tests (line 10 in Algorithm 3), and if either of them returns *fail*, the restart is triggered. We now describe these three parts in detail below.

**MALG-initialization.** MALG is run for an interval of length $2^n$ (unless it is terminated by the non-stationarity detection), which is called a *block*. During the initialization, MALG partions the block equally into $2^{n-k}$ sub-intervals of length $2^k$ for $k = 0, 1, \ldots, n$, and an instance of based algorithm (denoted by ALG) is scheduled for each of these sub-intervals with probability $\frac{\rho(2^n)}{\rho(2^k)}$, where $\rho$ is a non-increasing function associated with the bound on **R2** for ALG in a stationary environment (see Appendix). We refer to these instances of length $2^k$ as order-$k$ instances.

**MALG-interaction.** After the initialization, MALG starts interacting with the environment as follows. In each episode $m$, the unique instance $alg$ that covers this episode with the shortest length is considered as active, while all others are regarded as inactive. MALG follows the decision of the active instance $alg$ and updates it after receiving the feedback from the environment. All inactive instances do not make any decisions or updates, that is, they are paused but may be resumed at some future episode. We refer the reader to Appendix for an illustrative example for MALG procedure.

**Non-stationarity detection** For $\beta > 0$, two non-stationarity tests are performed for the two terms in the decomposition (8). In particular, **Test1** prevents **R1** from growing too large by testing if there is some order$-k$ instance's interval during which the learner's average performance $\frac{1}{2^k}\sum_{\tau=alg.s}^{alg.e} R_\tau$ is larger than the promised optimistic estimator $U_m = \min_{\tau\in[m_n,m]} g_\tau$ (for a stationary environment) by

[1] More precisely, $\sum_{m=1}^{M} e^{\beta V_1^{\pi^m,m}}$ can be estimated from $\sum_{m=1}^{M} e^{\beta \sum_{h=1}^{H} r_h^m}$ using the Azuma's inequality.

a certain amount. On the other hand, **Test2** prevents **R2** from growing too large by directly testing if its average is large than the promised regret bound. The two non-stationarity tests for $\beta < 0$ are similar but with $\frac{1}{2^k}\sum_{\tau=alg.s}^{alg.e} R_\tau$ and $U_m$ exchanged in **TEST1**, as well as with $g_\tau$ and $R_\tau$ exchanged in **TEST2**.

### 4.3 Theoretical Results and Discussions

For simplicity, we denote the revised Algorithms 1 and 2 without the restart mechanism (that is, $W = M$) as RSMB and RSQ, respectively. We now present our main theoretical result for Algorithm 3 when the base algorithms are RSMB and RSQ, respectively.

**Theorem 4.1** *For every $\delta \in (0,1]$, with probability at least $1-\delta$ it holds for Algorithm 3 that*

$$\text{D-Regret}(\text{M}) \leq \begin{cases} \widetilde{\mathcal{O}}\left(e^{|\beta|H}|S|^{\frac{2}{3}}|A|^{\frac{1}{3}}H^2M^{\frac{2}{3}}B^{\frac{1}{3}}\right), & \textit{if ALG is RSMB}, \\ \widetilde{\mathcal{O}}\left(e^{|\beta|H}|S|^{\frac{2}{3}}|A|^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right), & \textit{if ALG is RSQ}. \end{cases}$$

The above results show that the dynamic regret bound of the adaptive Algorithm 3 (almost) matches that of the restart Algorithms 1-2 that require the knowledge of the variation budget. The proof of Theorem 4.1 relies on the results in Theorems 3.1-2 and is provided in Appendix.

## 5 Lower Bound

We now present a lower bound on the dynamic regret which complements the upper bounds in Theorems 3.1, 3.2 and 4.1.

**Theorem 5.1** *For sufficiently large $M$, there exists an instance of non-stationary MDP with $H$ horizons, state space $\mathcal{S}$, action space $\mathcal{A}$ and variation budget $B$ such that*

$$\text{D-Regret}(\text{M}) \geq \Omega\left(\frac{e^{\frac{2|\beta|H}{3}}-1}{|\beta|}|\mathcal{S}|^{\frac{1}{3}}|\mathcal{A}|^{\frac{1}{3}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right).$$

Theorem 5.1 shows that the exponential dependence on $|\beta|$ and $H$ in Theorems 3.1, 3.2 and 4.1 is essentially indispensable and that the results in Theorems 3.1, 3.2 and 4.1 are nearly optimal in their dependence on $|\mathcal{A}|$, $M$ and $B$. When $\beta \to 0$, we recover the existing lower bound for the non-stationary risk-neutral episodic MDP problems (Mao et al. 2020).

The proof is given in Appendix. In the proof, the hard instance we construct is a non-stationary MDP with piecewise constant dynamics on each segment of the horizon, and its dynamics experience an abrupt change at the beginning of each new segment. In each segment, we construct a $|\mathcal{S}||\mathcal{A}|$-arm bandit model with Bernoulli reward for each arm. This bandit model can be seen as a special case of our episodic MDP problem, and then we show the expected regret, in terms of the logarithmic-exponential objective, that any bandit algorithm has to incur.

## 6 Risk Control Under the Non-stationarity

Risk control in non-stationary RL is more challenging since the rewards and dynamics are time-varying and unknown. In this section, we discuss some key ideas behind our methods and proofs.

**Normalized dynamics estimation in model-based algorithm.** In model-based algorithms for non-stationary risk-neutral RL, the un-normalized dynamics estimation (Domingues et al. 2021; Ding and Lavaei 2022) is sufficient for achieving a near-optimal regret because the effect of the model estimation error due to the "unnormalization" on the dynamic regret is little. However, it is critical to use the normalized dynamics estimation (5a) in Algorithm 1. This is because that a small model estimation error due to the "unnormalization" may be amplified when $\beta \to 0$. We note that the stationary version of our Algorithm 1 is also the first model-based algorithm with a theoretical guarantee for *stationary* risk-sensitive RL problems in the literature.

**Multiplicative feature of the exponential Bellman equation.** The multiplicative feature of the exponential Bellman equation will involve the policy evaluation error as multiplicative terms. These terms are easy to bound in a stationary environment in light of the optimistic estimator of the exponential value function. However, due to the non-stationary drifting of the environment, the estimator $V_h^m$ may no longer be an optimistic estimator and the errors of the optimistic estimator are all in the form of a multiplicative way due to the nature of the exponential Bellman equation. We need to introduce additional terms to guarantee each multiplicative terms are non-negative as in the proof of Theorem 3.2.

**Non-stationarity detection on the exponential value functions.** Different from non-stationarity detection for risk-neutral RL (Wei and Luo 2021), we design non-stationarity detection mechanism for the exponential value functions (3) instead of the value functions (1) in Algorithm 3. This is because the non-linearity of the risk-sensitive value function makes it difficult to obtain its unbiased estimation, which is needed in the design of non-stationary detection mechanism.

**Separation design of the risk-control and the non-stationarity.** When the variation budget is known, the risk-control and the handling of the non-stationarity can be separately designed in the algorithm, that is, the restart frequency in Algorithms 1 and 2 does not depend on the risk parameter $\beta$ and only depends on the non-stationarity of the environment $B$. If we know the environment's variation budget in advance, then we can schedule the restart frequency ahead no matter the risk-sensitivity. On the other hand, without such knowledge of the variation budget, the adaptive non-stationary detection needs to take into account the risk parameter $\beta$ because the promised regret bound, the optimistic estimator, and the unbiased sample of the exponential value functions all depend on $\beta$.

## Acknowledgements

This work was supported by grants from ARO, AFOSR, ONR and NSF.

## References


Auer, P.; Gajane, P.; and Ortner, R. 2019. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, 138–158. PMLR.

Bäuerle, N.; and Rieder, U. 2014. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1): 105–120.

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

Borkar, V. S. 2002. Q-learning for risk-sensitive control. *Mathematics of operations research*, 27(2): 294–311.

Borkar, V. S.; and Meyn, S. P. 2002. Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1): 192–209.

Cavazos-Cadena, R.; and Fernández-Gaucherand, E. 2000. The vanishing discount approach in Markov chains with risk-sensitive criteria. *IEEE Transactions on Automatic Control*, 45(10): 1800–1816.

Chen, Y.; Lee, C.-W.; Luo, H.; and Wei, C.-Y. 2019. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, 696–726. PMLR.

Cheung, W. C.; Simchi-Levi, D.; and Zhu, R. 2020. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, 1843–1854. PMLR.

Chow, Y.; and Ghavamzadeh, M. 2014. Algorithms for CVaR optimization in MDPs. *Advances in neural information processing systems*, 27.

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

Delage, E.; and Mannor, S. 2010. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations research*, 58(1): 203–213.

Di Castro, D.; Tamar, A.; and Mannor, S. 2012. Policy gradients with variance related risk criteria. *arXiv preprint arXiv:1206.6404*.

Di Masi, G. B.; and Stettner, L. 1999. Risk-sensitive control of discrete-time Markov processes with infinite horizon. *SIAM Journal on Control and Optimization*, 38(1): 61–78.

Ding, Y.; and Lavaei, J. 2022. Provably Efficient Primal-Dual Reinforcement Learning for CMDPs with Non-stationary Objectives and Constraints. *arXiv preprint arXiv:2201.11965*.

Ding, Y.; Lavaei, J.; and Arcak, M. 2021. Time-variation in online nonconvex optimization enables escaping from spurious local minima. *IEEE Transactions on Automatic Control*.

Domingues, O. D.; Ménard, P.; Pirotta, M.; Kaufmann, E.; and Valko, M. 2021. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, 3538–3546. PMLR.

Fei, Y.; Yang, Z.; Chen, Y.; and Wang, Z. 2021. Exponential Bellman Equation and Improved Regret Bounds for Risk-Sensitive Reinforcement Learning. *Advances in Neural Information Processing Systems*, 34.

Fei, Y.; Yang, Z.; Chen, Y.; Wang, Z.; and Xie, Q. 2020. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *arXiv preprint arXiv:2006.13827*.

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

Fernández-Gaucherand, E.; and Marcus, S. I. 1997. Risk-sensitive optimal control of hidden Markov models: Structural results. *IEEE Transactions on Automatic Control*, 42(10): 1418–1422.

Fleming, W. H.; and McEneaney, W. M. 1992. Risk sensitive optimal control and differential games. In *Stochastic theory and adaptive control*, 185–197. Springer.

Fleming, W. H.; and McEneaney, W. M. 1995. Risk-sensitive control on an infinite time horizon. *SIAM Journal on Control and Optimization*, 33(6): 1881–1915.

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

Hernández-Hernández, D.; and Marcus, S. I. 1996. Risk sensitive control of Markov processes in countable state space. *Systems & control letters*, 29(3): 147–155.

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

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

La, P.; and Ghavamzadeh, M. 2013. Actor-critic algorithms for risk-sensitive MDPs. *Advances in neural information processing systems*, 26.

Man, C. D.; Micheletto, F.; Lv, D.; Breton, M.; Kovatchev, B.; and Cobelli, C. 2014. The UVA/PADOVA type 1 diabetes simulator: new features. *Journal of diabetes science and technology*, 8(1): 26–34.

Mao, W.; Zhang, K.; Zhu, R.; Simchi-Levi, D.; and Başar, T. 2020. Model-Free Non-Stationary RL: Near-Optimal Regret and Applications in Multi-Agent RL and Inventory Control. *arXiv preprint arXiv:2010.03161*.

Markowitz, H. M. 1968. Portfolio selection. In *Portfolio selection*. Yale university press.

Moody, J.; and Saffell, M. 2001. Learning to trade via direct reinforcement. *IEEE transactions on neural Networks*, 12(4): 875–889.

Niv, Y.; Edlund, J. A.; Dayan, P.; and O'Doherty, J. P. 2012. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2): 551–562.

Osogami, T. 2012. Robustness and risk-sensitivity in Markov decision processes. *Advances in Neural Information Processing Systems*, 25.

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

Tamar, A.; Chow, Y.; Ghavamzadeh, M.; and Mannor, S. 2015. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28.

Tamar, A.; Glassner, Y.; and Mannor, S. 2015. Optimizing the CVaR via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Touati, A.; and Vincent, P. 2020. Efficient learning in non-stationary linear Markov decision processes. *arXiv preprint arXiv:2010.12870*.

Wei, C.-Y.; and Luo, H. 2021. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on Learning Theory*, 4300–4354. PMLR.

Zhao, P.; Zhang, L.; Jiang, Y.; and Zhou, Z.-H. 2020. A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 746–755. PMLR.

Zhong, H.; Yang, Z.; and Szepesvári, Z. W. C. 2021. Optimistic Policy Optimization is Provably Efficient in Non-stationary MDPs. *arXiv preprint arXiv:2110.08984*.

Zhou, H.; Chen, J.; Varshney, L. R.; and Jagmohan, A. 2020. Nonstationary reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.04244*.