Non-stationary Risk-sensitive Reinforcement Learning: Near-optimal Dynamic Regret, Adaptive Detection, and Separation Design

Yuhao Ding
UC Berkeley
Berkeley, CA 94709
yuhao_ding@berkeley.edu

Ming Jin
Virginia Tech
Blacksburg, VA 24061
jinming@vt.edu

Javad Lavaei
UC Berkeley
Berkeley, CA 94709
lavaei@berkeley.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 [32], autonomous driving [24] and human behavior modeling [34].

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 [32], 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 [30], 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.

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 [26] under episodic Markov decision processes with unknown and time-varying reward functions and state transition kernels.

The challenge of non-stationary RL with an entropic risk measure lies mainly in the non-linearity of the value function (see Equation (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

Preprint. Under review.

β is small. Furthermore, the non-linearity of the objective function 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 [40]. To address these difficulties, we first transform the standard Bellman equations to the exponential Bellman equation (see Equation (3)) which associates the instantaneous reward and value function of the next step in a multiplicative way [18], rather than in an additive way as in the risk-neutral non-stationary RL. 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 difficult to gauge the bounds. To this end, 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.
- 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.

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	$\Omega\left(\frac{e^{\frac{2\lvert\beta\rvert H}{3}}-1}{\lvert\beta\rvert}\lvert\mathcal{S}\rvert^{\frac{1}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$			

Table 1: We summarize the dynamic regrets and lower bound obtained in this paper. Here, β 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.

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 [31, 43, 41, 16], exponential decayed weights [39], or sliding window [10, 42]. In this work, we focus on the restart method mainly due to its advantage of the simplicity of the the memory efficiency [41] and generalize it to the risk-sensitive RL setting. However, the prior knowledge of the variation budget is often unavailable in practice. The work [10] develop a Bandit-over-Reinforcement-Learning framework to relax this assumption, but it leads to the suboptimal regret. To achieve a nearly-optimal regret, some adaptive algorithms with a non-stationary detection are developed in [3, 9] for bandit problems and in [40] for general RL problems. 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) [33, 11, 13, 28, 14, 38, 37, 26]. Our work is closely related to the entropic risk measure. Following the seminal paper [26], this line of work includes [4, 5, 7, 6, 8, 12, 15, 21, 23, 25, 35, 22, 36, 19, 20, 18]. 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 [19] and the function approximation setting in [20]. Then,

a simple transformation of the risk-sensitive Bellman equations is proposed in [18], 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 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 π as the expected value of the cumulative rewards the agent receives under a risk measure of exponential utility by executing π starting from an arbitrary state at step h. Specifically, we have

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

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 π afterwards; that is,

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

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Under some mild regularity conditions [4], 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.

2.2 Exponential Bellman equation

For all $(s,a,h,m) \in \mathcal{S} \times \mathcal{A} \times [H] \times [M]$, the Bellman equation associated with π 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 [18] 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 π 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 $\left(Q_h^{\pi,m}\right.$ and $\left.V_{h+1}^{\pi,m}\right)$ in a multiplicative way, rather than in an additive way as in the standard Bellman equations (2).

2.3 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:

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

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 $B > 0$.

2.4 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 π^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 π^m over M episodes:

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

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

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

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

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

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 $\beta < 0$, $N_h^m(s,a) \leftarrow 0, N_h^m(s,a,s') \leftarrow 0$ for all $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$;
6: **end if**
7: **for** $h = H, \ldots, 1$ **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 Γ_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** $h = 1, 2, \ldots, H$ **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**

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

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

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 [20] 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 [18]. 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 α_t is defined as $\frac{H+1}{H+t}$, which is motivated by [27] 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

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

for some constant $C_2 > 1$.

3.3 Theoretical results and discussions

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

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

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

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 B and C, respectively. 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 [17] and restart Q-learning [31].

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 [40] 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:

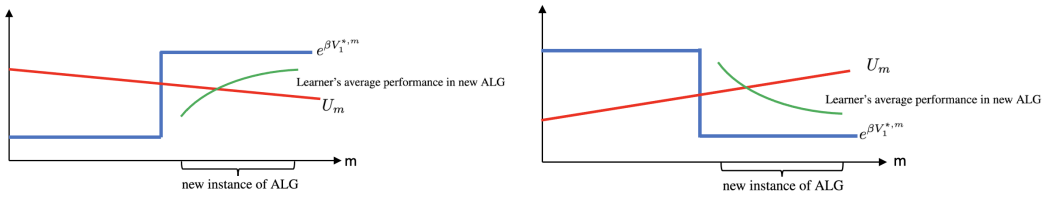
$$\text{D-Regret}(M) \leq \underbrace{\frac{1}{\beta} \sum_{m=1}^{M} \left(e^{\beta V_1^{*,m}} - e^{\beta V_1^m} \right)}_{\textbf{R1}} + \underbrace{\frac{1}{\beta} \sum_{m=1}^{M} \left(e^{\beta V_1^m} - e^{\beta V_1^{\pi^m, m}} \right)}_{\textbf{R2}} \tag{8}$$

where 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

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 (Algorithm 4) 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 π_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 $\begin{cases} \frac{1}{m-m_n+1} \sum_{\tau=m_n}^{m} (g_\tau - R_\tau) \geq 3\hat{\rho}(m - m_n + 1), & \text{if } \beta > 0, \\ \frac{1}{m-m_n+1} \sum_{\tau=m_n}^{m} (R_\tau - g_\tau) \geq 3\hat{\rho}(m - m_n + 1), & \text{if } \beta < 0, \end{cases}$

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 $e^{\beta V_1^m}$ 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 $e^{\beta V_1^m}$ is a UCB-based optimistic estimator to facilitate non-stationary detection.



(a) $\beta > 0$. (b) $\beta < 0$.

Figure 1: An illustration of the risk-sensitive non-stationarity 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 β) this quantity, we can infer that the environment has changed.

[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.

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 (Algorithm 4 in Appendix D.1), 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 ρ is a non-increasing function associated with the bound on **R2** for ALG in a stationary environment (see Appendix D.3). 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 read to Appendix D.2 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 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_τ and R_τ 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*

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

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 D.4.

5 Lower bound

We now present a lower bound on the dynamic regret.

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*

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

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. The proof is given in Appendix D.4.

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 [17, 16] 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 $\beta \to 0$.

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 a more careful analysis is needed.

Non-stationarity detection on the exponential value functions. Different from non-stationarity detection for risk-neutral RL [40], 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 β 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 β because the promised regret bound, the optimistic estimator, and the unbiased sample of the exponential value functions all depend on β.

7 Conclusion and future work

In this paper, we provide strong theoretical analyses for the non-stationary risk-sensitive RL problem, which is motivated by various risk-sensitive applications. We propose two restart-based algorithms that require the knowledge of the variation budget, as well as a black-box approach to turn a certain risk-sensitive RL algorithm in a (near-)stationary environment into another algorithm in a non-stationary environment without requiring the knowledge of the variation budge. The dynamic regret bounds of these algorithms are obtained and a lower bound is established to verify the near-optimality of the proposed upper bounds. Our results also reveal the condition under which the risk control and the handling of the non-stationarity can be separately designed in the algorithm.

One important future direction lies in extending our results to other notions of risk, such as the general coherent risk measures [2]. Furthermore, it is useful to study how to adjust the risk sensitivity parameter adaptively in a non-stationary environment.

Acknowledgement

We thank Kaiqing Zhang for the fruitful discussions.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

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

[3] Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158. PMLR, 2019.

[4] Nicole Bäuerle and Ulrich Rieder. More risk-sensitive markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

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

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

[7] Vivek S Borkar and Sean P Meyn. Risk-sensitive optimal control for markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

[8] Rolando Cavazos-Cadena and Emmanuel Fernández-Gaucherand. The vanishing discount approach in markov chains with risk-sensitive criteria. *IEEE Transactions on Automatic Control*, 45(10):1800–1816, 2000.

[9] Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pages 696–726. PMLR, 2019.

[10] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, pages 1843–1854. PMLR, 2020.

[11] Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in mdps. *Advances in neural information processing systems*, 27, 2014.

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

[13] Erick Delage and Shie Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

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

[15] Giovanni B Di Masi and Lukasz Stettner. Risk-sensitive control of discrete-time markov processes with infinite horizon. *SIAM Journal on Control and Optimization*, 38(1):61–78, 1999.

[16] Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for CMDPs with non-stationary objectives and constraints. *arXiv preprint arXiv:2201.11965*, 2022.

[17] Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 3538–3546. PMLR, 2021.

[18] Yingjie Fei, Zhuoran Yang, Yudong Chen, and Zhaoran Wang. Exponential Bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *arXiv preprint arXiv:2006.13827*, 2020.

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

[21] Emmanuel Fernández-Gaucherand and Steven I Marcus. Risk-sensitive optimal control of hidden markov models: Structural results. *IEEE Transactions on Automatic Control*, 42(10):1418–1422, 1997.

[22] Wendell H Fleming and William M McEneaney. Risk sensitive optimal control and differential games. In *Stochastic theory and adaptive control*, pages 185–197. Springer, 1992.

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

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

[25] Daniel Hernández-Hernández and Steven I Marcus. Risk sensitive control of Markov processes in countable state space. *Systems & control letters*, 29(3):147–155, 1996.

[26] Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

[27] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

[28] Prashanth La and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive mdps. *Advances in neural information processing systems*, 26, 2013.

[29] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[30] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA type 1 diabetes simulator: new features. *Journal of diabetes science and technology*, 8(1):26–34, 2014.

[31] Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Başar. Model-free non-stationary RL: Near-optimal regret and applications in multi-agent rl and inventory control. *arXiv preprint arXiv:2010.03161*, 2020.

[32] Harry M Markowitz. Portfolio selection. In *Portfolio selection*. Yale university press, 1968.

[33] John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE transactions on neural Networks*, 12(4):875–889, 2001.

[34] Yael Niv, Jeffrey A Edlund, Peter Dayan, and John P O'Doherty. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2):551–562, 2012.

[35] Takayuki Osogami. Robustness and risk-sensitivity in markov decision processes. *Advances in Neural Information Processing Systems*, 25, 2012.

[36] Yun Shen, Wilhelm Stannat, and Klaus Obermayer. Risk-sensitive markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

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

[38] Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the CVaR via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[39] Ahmed Touati and Pascal Vincent. Efficient learning in non-stationary linear Markov decision processes. *arXiv preprint arXiv:2010.12870*, 2020.

[40] Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on Learning Theory*, pages 4300–4354. PMLR, 2021.

[41] Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 746–755. PMLR, 2020.

[42] Han Zhong, Zhuoran Yang, and Zhaoran Wang Csaba Szepesvári. Optimistic policy optimization is provably efficient in non-stationary MDPs. *arXiv preprint arXiv:2110.08984*, 2021.

[43] Huozhi Zhou, Jinglin Chen, Lav R Varshney, and Ashish Jagmohan. Nonstationary reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.04244*, 2020.

A 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.

B Proof of Theorem 3.1

B.1 Preliminaries

First, we set some notations and definitions. Define $\iota := \log(6H|\mathcal{S}||\mathcal{A}|W/p)$ for a given $p \in (0, 1]$. We adopt the shorthand notations $\mathbb{1}_h^m(s,a) := \mathbb{1}\{(s_h^m, a_h^m) = (s,a)\}$ and $r_h^m := r_h(s_h^m, a_h^m)$ for $(m,h) \in [M] \times [H]$. The epoch is defined as an interval that starts at the first episode after a restart and ends at the first time when the restart is triggered. In Algorithm 1, the restart mechanism divides M episodes into $\left\lceil \frac{M}{W} \right\rceil$ epochs.

For every $(m,h) \in [M] \times [H]$, and $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we define two visitation counters $N_h^m(s,a,s')$ and $N_h^m(x,a)$ at step h in episode m as follows:

$$\begin{aligned} N_h^m(s,a,s') &= \sum_{\tau=\ell^m}^{m-1} 1\{(s,a,s') = (s_h^\tau, a_h^\tau, s_{h+1}^\tau)\}, \\ N_h^m(s,a) &= \sum_{\tau=\ell^m}^{m-1} 1\{(s,a) = (s_h^\tau, a_h^\tau)\}. \end{aligned} \tag{9a}$$

This allows us to estimate the transition kernel $\mathcal{P}_h^m$ and reward function r^m for episode m using only the data from the episode $\ell^m = (\left\lceil \frac{m}{W} \right\rceil - 1)W + 1$ to the episode m by

$$\widehat{\mathcal{P}}_h^m(s' \mid s,a) = \frac{N_h^m(s,a,s') + \frac{\lambda}{|\mathcal{S}|}}{N_h^m(s,a) + \lambda}, \text{ for all } (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \tag{10a}$$

$$\widehat{r}_h^m(s,a) = \frac{1}{N_h^m(s,a) + \lambda} \sum_{\tau=\ell^m}^{m-1} 1\{(s,a) = (s_h^\tau, a_h^\tau)\} r_h^\tau(s_h^\tau, a_h^\tau), \text{ for all } (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{10b}$$

where $\lambda > 0$ is the regularization parameter. We denote by V_h^m, G_h^m, Γ_h^m the values of V_h, G_h, Γ_h after the updates in step h of episode m, respectively. We also set $Q_h^m = \frac{1}{\beta} \log\{G_h^m\}$.

Let us fix a pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Recall from Algorithm 1 that

$$w_h^m(s,a) = \sum_{s'} \hat{\mathcal{P}}_h^m(s' \mid s,a) \left[e^{\beta[\hat{r}_h^m(s,a) + V_{h+1}^m(s')]} \right].$$

We define

$$q_{h,1}^{m,+}(s,a) := \begin{cases} w_h^m(s,a) + \Gamma_h^m(s,a), & \text{if } \beta > 0 \\ w_h^m(s,a) - \Gamma_h^m(s,a), & \text{if } \beta < 0 \end{cases}$$

$$q_{h,1}^{m}(s,a) := \begin{cases} \min\left\{ q_{h,1}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta > 0 \\ \max\left\{ q_{h,1}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta < 0 \end{cases}$$

and

$$q_{h,2}^m(s,a) := \mathbb{E}_{s' \sim \mathcal{P}_h^m(\cdot|s,a)} \left[e^{\beta[r_h^m(s,a) + V_{h+1}^m(s')]} \right], \tag{11}$$

as well as the following for a policy π,

$$q_{h,3}^{m,\pi}(s,a) := \mathbb{E}_{s' \sim \mathcal{P}_h^m(\cdot|s,a)} \left[e^{\beta[r_h^m(s,a) + V_{h+1}^{\pi,m}(s')]} \right] \tag{12}$$

B.2 Model prediction errors

Lemma B.1 *Define* $\overline{\mathcal{V}}_{h+1} := \left\{\bar{V}_{h+1} : \mathcal{S} \to \mathbb{R} \mid \forall s \in \mathcal{S}, \bar{V}_{h+1}(s) \in [0, H-h]\right\}$. *For any* $p \in (0,1]$, *with probability* $1 - p/2$, *we have*

$$\begin{aligned}
&\left|\sum_{s' \in \mathcal{S}} \left(\widehat{\mathcal{P}}_h^m\left(s' \mid s,a\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]} - \mathcal{P}_h^m\left(s' \mid s,a\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right| \\
\leq &\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}}
\end{aligned}$$

for every $(s,a,m,h) \in \mathcal{S} \times \mathcal{A} \times [M] \times [H]$ *and* $\bar{V} \in \overline{\mathcal{V}}_{h+1}$, *where* Γ_h^m *is defined in* (6).

Proof. For the ease of notation, we denote $\sum_{s' \in \mathcal{S}} \mathcal{P}_h^m\left(s' \mid s,a\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}$ as $\left(\mathcal{P}_h^m e^{\beta\left[r_h^m+\bar{V}\right]}\right)(s,a)$. Then, for every $\bar{V} \in \mathcal{V}_{h+1}$, we consider the difference between $\sum_{s' \in \mathcal{S}} \widehat{\mathcal{P}}_h^m\left(s' \mid \cdot,\cdot\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}$ and $\sum_{s' \in \mathcal{S}} \mathcal{P}_h^m\left(s' \mid \cdot,\cdot\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}$ as follows:

$$\begin{aligned}
&\left(N_h^m(s,a)+\lambda\right)\left|\sum_{s' \in \mathcal{S}} \left(\widehat{\mathcal{P}}_h^m\left(s' \mid s,a\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]} - \mathcal{P}_h^m\left(s' \mid s,a\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right| \qquad (13) \\
= &\left|\sum_{s' \in \mathcal{S}} \left(N_h^m\left(s,a,s'\right) + \frac{\lambda}{|\mathcal{S}|}\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]} - \left(N_h^m(s,a)+\lambda\right)\left(\mathcal{P}_h^m e^{\beta\left[r_h^m+\bar{V}\right]}\right)(s,a)\right| \\
\leq &\left|\sum_{s' \in \mathcal{S}} N_h^m\left(s,a,s'\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]} - N_h^m(s,a)\left(\mathcal{P}_h^m e^{\beta\left[r_h^m+\bar{V}\right]}\right)(s,a)\right| \\
&+ \lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s' \in \mathcal{S}} e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]} - \mathcal{P}_h^m e^{\beta\left[r_h^m+\bar{V}\right]}(s,a)\right| \\
= &\left|\sum_{\tau=\ell_Q^m}^{m-1} \mathbb{1}\left\{(s,a) = (s_h^\tau, a_h^\tau)\right\}\left(e^{\beta\left[r_h^m(s_h^\tau,a_h^\tau)+\bar{V}(s_{h+1}^\tau)\right]} - \left(\mathcal{P}_h^m e^{\beta\left[r_h^m+\bar{V}\right]}\right)(s,a)\right)\right| \\
&+ \lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s' \in \mathcal{S}} e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]} - \mathcal{P}_h^m e^{\beta\left[r_h^m+\bar{V}\right]}(s,a)\right| \\
= &\left|\sum_{\tau=\ell_Q^m}^{m-1} \mathbb{1}\left\{(s,a) = (s_h^\tau, a_h^\tau)\right\} e^{\beta r_h^m(s_h^\tau,a_h^\tau)}\left(e^{\beta\bar{V}(s_{h+1}^\tau)} - \left(\mathcal{P}_h^m e^{\beta\bar{V}}\right)(s,a)\right)\right| \\
&+ \lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s' \in \mathcal{S}} e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]} - \mathcal{P}_h^m e^{\beta\left[r_h^m+\bar{V}\right]}(s,a)\right| \\
\leq &\left|\sum_{\tau=\ell_Q^m}^{m-1} \mathbb{1}\left\{(s,a) = (s_h^\tau, a_h^\tau)\right\} e^{\beta r_h^m(s_h^\tau,a_h^\tau)}\left(e^{\beta\bar{V}(s_{h+1}^\tau)} - \left(\mathcal{P}_h^\tau e^{\beta\bar{V}}\right)(s,a)\right)\right| \qquad (14) \\
&+ \left|\sum_{\tau=\ell_Q^m}^{m-1} \mathbb{1}\left\{(s,a) = (s_h^\tau, a_h^\tau)\right\} e^{\beta r_h^m(s_h^\tau,a_h^\tau)}\left(\left(\mathcal{P}_h^\tau e^{\beta\bar{V}}\right)(s,a) - \left(\mathcal{P}_h^m e^{\beta\bar{V}}\right)(s,a)\right)\right| \qquad (15) \\
&+ \lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s' \in \mathcal{S}} e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]} - \mathcal{P}_h^m e^{\beta\left[r_h^m+\bar{V}\right]}(s,a)\right| \qquad (16)
\end{aligned}$$

for every $(m,h) \in [M] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$.

To analyze the term in (14), we let $\eta_h^\tau := e^{\beta\left[r_h^m(s_h^\tau,a_h^\tau)+\bar{V}(s_{h+1}^\tau)\right]} - \left(\mathcal{P}_h^\tau e^{\beta\left[r_h^m+\bar{V}\right]}\right)(s_h^\tau, a_h^\tau)$. Conditioning on the filtration $\mathcal{F}_{h,1}^m$, the term η_h^τ is a zero-mean and $\left|e^{\beta(H-h+1)} - 1\right|$-sub-Gaussian random variable. By Lemma F.2, we use $Y = \lambda I$ and $X_\tau = \mathbb{1}\left\{(s,a) = (s_h^\tau, a_h^\tau)\right\}$ and thus with probability at

least $1-\delta$ it holds for every $m \in [M]$ that

$$
\begin{aligned}
&(N_h^m(s,a)+\lambda)^{-1/2}\left|\sum_{\tau=\ell_Q^m}^{m-1}\mathbb{1}\left\{(s,a)=(s_h^\tau,a_h^\tau)\right\}\left(e^{\beta[r_h^m(s_h^\tau,a_h^\tau)+\bar{V}](s_{h+1}^\tau)}-\left(\mathcal{P}_h^\tau e^{\beta[r_h^m+\bar{V}]}\right)(s_h^\tau,a_h^\tau)\right)\right| \\
\leq&\sqrt{\frac{(e^{\beta(H-h+1)}-1)^2}{2}\log\left(\frac{\left(N_h^m(s,a)+\lambda\right)^{1/2}\lambda^{-1/2}}{\delta}\right)} \\
\leq&\sqrt{\frac{(e^{\beta(H-h+1)}-1)^2}{2}\log\left(\frac{W}{\delta}\right)}
\end{aligned}
$$

where W is the restart period.

For the term in (15), by the definition of $B_{\mathcal{P},\mathcal{E}}$ and N_h^m, we have

$$
\begin{aligned}
&\left|\sum_{\tau=\ell_Q^m}^{m-1}\mathbb{1}\left\{(s,a)=(s_h^\tau,a_h^\tau)\right\}\left(\left(\mathcal{P}_h^\tau e^{\beta[r_h^m+\bar{V}]}\right)(s,a)-\left(\mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}\right)(s,a)\right)\right| \\
=&\left|\sum_{\tau=\ell_Q^m}^{m-1}\mathbb{1}\left\{(s,a)=(s_h^\tau,a_h^\tau)\right\}\left(\left(\mathcal{P}_h^\tau\left(e^{\beta[r_h^m+\bar{V}]}-1\right)\right)(s,a)-\left(\mathcal{P}_h^m\left(e^{\beta[r_h^m+\bar{V}]}-1\right)\right)(s,a)\right)\right| \\
\leq&\left|\sum_{\tau=\ell_Q^m}^{m-1}\mathbb{1}\left\{(s,a)=(s_h^\tau,a_h^\tau)\right\}\right|\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}} \\
\leq&\left(N_h^m(s,a)+\lambda\right)\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}.
\end{aligned}
$$

where the first equality is due to $\mathcal{P}_h^m 1=\mathcal{P}_h^\tau 1$ for all $\tau\in[\ell^m,m-1]$. For the term in (16), we have

$$
\begin{aligned}
\lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s'\in\mathcal{S}}e^{\beta[r_h^m(s,a)+\bar{V}(s')]}-\mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}(s,a)\right|\leq&\frac{\lambda}{|\mathcal{S}|}\sum_{s'\in\mathcal{S}}\left|e^{\beta[r_h^m(s,a)+\bar{V}(s')]}-\mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}(s,a)\right| \\
\leq&\lambda\left|e^{\beta(H-h+1)}-1\right|.
\end{aligned}
$$

By returning to (13) and setting $\lambda=1$, with probability at least $1-\delta$ it holds that

$$
\begin{aligned}
&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta[r_h^m(s,a)+\bar{V}(s')]}-\mathcal{P}_h^m\left(s'\mid s,a\right)e^{\beta[r_h^m(s,a)+\bar{V}(s')]}\right)\right| \\
\leq&\left(N_h^m(s,a)+\lambda\right)^{-\frac{1}{2}}\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{1}{2}\left(\log\left(\frac{W}{\delta}\right)\right)}+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}+\left|e^{\beta(H-h+1)}-1\right| \\
\leq&C_1\left(N_h^m(s,a)+\lambda\right)^{-\frac{1}{2}}\left|e^{\beta(H-h+1)}-1\right|\sqrt{\left(\log\left(\frac{W}{\delta}\right)\right)}+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}
\end{aligned}
$$

for all $m\in[M]$ and for some constant $C_1>1$.

Furthermore, let $d(V,V')=\max_{s\in\mathcal{S}}|V(s)-V'(s)|$ be a distance on $\mathcal{V}_{h+1}$. For every ϵ, an ϵ-covering $\mathcal{V}_{h+1}^\epsilon$ of $\mathcal{V}_{h+1}$ with respect to distance $d(\cdot,\cdot)$ satisfies $|\mathcal{V}_{h+1}^\epsilon|\leq\left(\frac{1}{\epsilon}\right)^{|\mathcal{S}|}$. Then, for every $V\in\mathcal{V}_{h+1}$, there exists $V'\in\mathcal{V}_{h+1}^\epsilon$ such that $\max_{s\in\mathcal{S}}|V(s)-V'(s)|\leq\epsilon$, which further implies that

$$
\max_{s,a,s'}\left|e^{\beta[r_h^m(s,a)+V(s')]}-e^{\beta[r_h^m(s,a)+V'(s')]}\right|\leq g_h(\beta)\epsilon,
$$

where

$$
g_h(\beta)=\begin{cases}e^{\beta(H-h+1)}\beta, & \text{if } \beta>0,\\ -\beta, & \text{if } \beta<0.\end{cases} \tag{17}
$$

Thus, by the triangle inequality and (13), we have

$$\begin{aligned}
&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+V\left(s'\right)\right]}-\mathcal{P}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+V\left(s'\right)\right]}\right)\right| \\
\leq&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+V'\left(s'\right)\right]}-\mathcal{P}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+V'\left(s'\right)\right]}\right)\right|+2g_h(\beta)\beta\epsilon \\
\leq&C_1\left(N_h^m(s,a)+\lambda\right)^{-1/2}\left|e^{\beta(H-h+1)}-1\right|\sqrt{\log\left(\frac{W}{\delta}\right)}+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}+2g_h(\beta)\epsilon.
\end{aligned}$$

Then, by choosing $\delta=(p/2)/\left(\left|\mathcal{V}_{h+1}^{\epsilon}\right|H\left|\mathcal{S}\right|\left|\mathcal{A}\right|\right)$, $\epsilon=\frac{1}{4\sqrt{W}}$, and taking a union bound over $V\in\mathcal{V}_{h+1}^{\epsilon}$ and $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, it holds with probability at least $1-p/2$ that

$$\begin{aligned}
&\sup_{V\in\mathcal{V}_{h+1}}\left\{\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+V\left(s'\right)\right]}-\mathcal{P}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+V\left(s'\right)\right]}\right)\right|\right\} \\
\leq&C_1\left(N_h^m(s,a)+\lambda\right)^{-1/2}\left|e^{\beta(H-h+1)}-1\right|\sqrt{\left(\log\left(\frac{6W\left|\mathcal{V}_{h+1}^{\epsilon}\right|H\left|\mathcal{S}\right|\left|\mathcal{A}\right|}{p}\right)\right)}+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}} \\
&+2g_h(\beta)\epsilon \\
\leq&C_1\left(N_h^m(s,a)+\lambda\right)^{-1/2}\left|e^{\beta(H-h+1)}-1\right|\sqrt{\left|\mathcal{S}\right|\left(\log\left(\frac{6WH\left|\mathcal{S}\right|\left|\mathcal{A}\right|}{p}\right)\right)}+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}} \\
&+g_h(\beta)W^{-1/2} \\
\leq&\left(C_1\left|e^{\beta(H-h+1)}-1\right|+g_h(\beta)\right)\left(N_h^m(s,a)+\lambda\right)^{-1/2}\sqrt{\left|\mathcal{S}\right|\left(\log\left(\frac{6WH\left|\mathcal{S}\right|\left|\mathcal{A}\right|}{p}\right)\right)}+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}} \\
\leq&C_1\left(\left|e^{\beta(H-h+1)}-1\right|+g_h(\beta)\right)\left(N_h^m(s,a)+\lambda\right)^{-1/2}\sqrt{\left|\mathcal{S}\right|\left(\log\left(\frac{6WH\left|\mathcal{S}\right|\left|\mathcal{A}\right|}{p}\right)\right)}+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}
\end{aligned}$$

for every $(s,a,m,h)\in\mathcal{S}\times\mathcal{A}\times[M]\times[H]$. By our choice of Γ_h^m, with probability at least $1-p/2$ it holds that

$$\begin{aligned}
&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}\left(s'\right)\right]}-\mathcal{P}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}\left(s'\right)\right]}\right)\right| \\
&\leq\Gamma_h^m+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}
\end{aligned}$$

for every $(s,a,m,h)\in\mathcal{S}\times\mathcal{A}\times[M]\times[H]$. □

Lemma B.2 *For every $(s,a,m,h)\in\mathcal{S}\times\mathcal{A}\times[M]\times[H]$ and $\bar{V}\in\overline{\mathcal{V}}_{h+1}$, we have*

$$\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[\hat{r}_h^m(s,a)+\bar{V}\left(s'\right)\right]}-\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}\left(s'\right)\right]}\right)\right|\leq\Gamma_h^m+g_h(\beta)B_{r,\mathcal{E}}$$

where $g_h(\beta)$ is defined in (17).

Proof. Since

$$\left|e^{\beta x}-e^{\beta y}\right|\leq\begin{cases}\beta e^{\beta u}|x-y|, & \text{if } \beta>0,\\ -\beta|x-y|, & \text{if } \beta<0\end{cases}$$

for every $0\leq x\leq u$ and $0\leq y\leq u$ where $u>0$ is some constant, it holds that

$$\begin{aligned}
&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[\hat{r}_h^m(s,a)+\bar{V}\left(s'\right)\right]}-\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}\left(s'\right)\right]}\right)\right| \\
\leq&g_h(\beta)\left|\hat{r}_h^m(s,a)-r_h^m(s,a)\right|. \qquad (18)
\end{aligned}$$

Furthermore, by our estimation $\hat{r}_h^m(x,a)$, we have

$$
\begin{aligned}
&|\hat{r}_h^m(x,a) - r_h^m(x,a)| \\
=&|\hat{r}_h^m(x,a) - r_h^m(x,a)| \\
=&\left(n_h^m(x,a)+\lambda\right)^{-1}\left|\sum_{\tau=\ell^m}^{m-1} 1\left\{(x,a)=(x_h^\tau,a_h^\tau)\right\}\left(r_h^\tau(x_h^\tau,a_h^\tau) - r_h^m(x,a)\right) - \lambda r_h^m(x,a)\right| \\
\leq& B_{r,\mathcal{E}} + \left(n_h^m(x,a)+\lambda\right)^{-1}|\lambda r_h^m(x,a)| \\
\leq& B_{r,\mathcal{E}} + \left(n_h^m(x,a)+\lambda\right)^{-1}\lambda \\
\leq& B_{r,\mathcal{E}} + \left(n_h^m(x,a)+\lambda\right)^{-1/2}\lambda
\end{aligned}
$$

By substituting the above inequality into (18) and setting $\lambda = 1$, we obtain the desired results. □

Lemma B.3 *For every $p \in (0,1]$, with probability $1-p/2$, we have*

$$
\begin{aligned}
&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s' \mid s,a\right)e^{\beta\left[\hat{r}_h^m(s,a)+\bar{V}(s')\right]} - \mathcal{P}_h^m\left(s' \mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right| \\
\leq& 2\Gamma_h^m + \left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}
\end{aligned}
$$

where $g_h(\beta)$ is defined in (17)*, for every $(s,a,m,h) \in \mathcal{S}\times\mathcal{A}\times[M]\times[H]$ and $\bar{V} \in \overline{\mathcal{V}}_{h+1}$.*

Proof. The proof follows from Lemma B.1, Lemma B.2 and Cauchy-Schwartz inequality. □

B.3 Value difference bounds

Lemma B.4 *Recall the definition of Γ_h^m from Algorithm 1. For all $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$, the following statement holds with probability at least $1-p/2$:*

- *If $\beta > 0$:*

$$
\begin{aligned}
-\left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} - g_h(\beta)B_{r,\mathcal{E}} \leq& \left(q_{h,1}^m - q_{h,2}^m\right)(s,a) \\
\leq& 4\Gamma_h^m + \left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}.
\end{aligned}
$$

- *If $\beta < 0$:*

$$
\begin{aligned}
-\left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} - g_h(\beta)B_{r,\mathcal{E}} \leq& \left(q_{h,2}^m - q_{h,1}^m\right)(s,a) \\
\leq& 4\Gamma_h^m + \left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}.
\end{aligned}
$$

(Note that $g_h(\beta)$ is defined in (17)*).*

Proof. We focus on the case of $\beta > 0$ since the proof for $\beta < 0$ is similar. We first fix a tuple $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$. By the definitions of $q_{h,1}^{m,+}$ and $q_{h,2}^m$, one can compute

$$
\begin{aligned}
&\left|\left(q_{h,1}^{m,+} - 2\Gamma_h^m - q_{h,2}^m\right)(s,a)\right| \\
=&\left|\left(w_h^m - q_{h,2}^m\right)(s,a)\right| \\
=&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s' \mid s,a\right)e^{\beta\left[\hat{r}_h^m(s,a)+\bar{V}(s')\right]} - \mathcal{P}_h^m\left(s' \mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right| \\
\leq& 2\Gamma_h^m + \left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}
\end{aligned}
$$

where the last step holds by Lemma B.1. Then, we have

$$
\begin{aligned}
-\left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} - g_h(\beta)B_{r,\mathcal{E}} \leq& \left(q_{h,1}^{m,+} - q_{h,2}^m\right)(s,a) \\
\leq& 4\Gamma_h^m + \left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}.
\end{aligned}
$$

Furthermore, if $q_{h,1}^{m,+} \le e^{\beta(H-h+1)}$, one can write

$$q_{h,1}^{m,+} - q_{h,2}^{m} = q_{h,1}^{m} - q_{h,2}^{m} \ge -\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} - g_h(\beta) B_{r,\mathcal{E}}.$$

If $q_{h,1}^{m,+} \ge e^{\beta(H-h+1)}$, we have $q_{h,1}^{m,+} - q_{h,2}^{m} = e^{\beta(H-h+1)} - q_{h,2}^{m} \ge 0$. In addition, since $q_{h,1}^{m,+} \ge q_{h,1}^{m}$, it holds that $q_{h,1}^{m} - q_{h,2}^{m} \le q_{h,1}^{m,+} - q_{h,2}^{m}$. This completes the proof.

□

Lemma B.5 *On the event of Lemma B.4, for all $(m,h,s,a) \in [M] \times [H] \times \mathcal{S} \times \mathcal{A}$ and every policy π:*

- *If $\beta > 0$:*

$$e^{\beta \cdot Q_h^m(s,a)} - e^{\beta \cdot Q_h^{\pi,m}(s,a)} \ge -(H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

- *If $\beta < 0$:*

$$e^{\beta \cdot Q_h^m(s,a)} - e^{\beta \cdot Q_h^{\pi,m}(s,a)} \le (H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

Proof. We focus on the case of $\beta > 0$ since the proof for $\beta < 0$ is similar. For the purpose of the proof, we set $Q_{H+1}^{\pi,m}(s,a) = Q_{H+1}^{*,m}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. We fix a tuple $(m,s,a) \in [M] \times \mathcal{S} \times \mathcal{A}$ and use strong induction on h. The base case for $h = H+1$ is satisfied since $e^{\beta \cdot Q_{H+1}^m(s,a)} = e^{\beta \cdot Q_{H+1}^{\pi,m}(s,a)} = 1$ for all $m \in [M]$ by definition. Now, we fix an index $h \in [H]$ and assume that

$$e^{\beta \cdot Q_{h+1}^m(s,a)} - e^{\beta \cdot Q_{h+1}^{\pi,m}(s,a)} \ge -(H-h)\left[\left|e^{\beta(H-h)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

Moreover, by the induction assumption, we have

$$\begin{aligned} e^{\beta \cdot V_{h+1}^m(s)} &= \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^m(s,a')} \\ &\ge \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^{\pi,m}(s,a')} - (H-h)\left[\left|e^{\beta(H-h)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right] \\ &\ge e^{\beta \cdot V_{h+1}^{\pi,m}(s)} - (H-h)\left[\left|e^{\beta(H-h)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right]. \end{aligned} \tag{19}$$

By the definitions of $q_{h,2}^m$ and $q_{h,3}^{m,\pi}$, it follows from (19) that

$$q_{h,2}^{m} - q_{h,3}^{m,\pi} \ge -(H-h)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

In addition, on the event of Lemma B.4, we also have

$$q_{h,1}^{m} - q_{h,2}^{m} \ge -\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

Therefore, it follows that

$$\begin{aligned} \left(e^{\beta \cdot Q_h^m} - e^{\beta \cdot Q_h^{\pi,m}}\right)(s,a) &= \left(q_{h,1}^{m} - q_{h,3}^{m,\pi}\right)(s,a) \\ &= \left(q_{h,1}^{m} - q_{h,2}^{m}\right)(s,a) + \left(q_{h,2}^{m} - q_{h,3}^{m,\pi}\right)(s,a) \\ &\ge -(H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right] \end{aligned}$$

which completes the induction. □

Lemma B.6 *For all $(m,h,s) \in [M] \times [H] \times \mathcal{S}$, policy π and $\delta \in (0,1]$, with probability at least $1 - \delta/2$:*

- *If $\beta > 0$:*

$$e^{\beta \cdot V_h^m(s,a)} - e^{\beta \cdot V_h^{\pi,m}(s,a)} \ge -(H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

- *If $\beta < 0$:*

$$e^{\beta \cdot V_h^m(s,a)} - e^{\beta \cdot V_h^{\pi,m}(s,a)} \le (H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

Proof. The result follows from Lemma B.5 and Equation (19).

B.4 Proof of Theorem 3.1

We first consider $\beta > 0$. For $h \in [H]$, we define

$$\delta_h^m := e^{\beta V_h^m(s_h^m)} - e^{\beta V_h^{\pi^m,m}(s_h^m)}, \tag{20a}$$

$$\begin{aligned}\zeta_{h+1}^m &:= q_{h,2}^m - q_{h,3}^m - e^{\beta r_h^m(s_h^m,a_h^m)}\delta_{h+1}^m \\ &= \left[P_h^m\left(e^{\beta\left[r_h^m(s_h^m,a_h^m)+V_{h+1}^m(s')\right]} - e^{\beta\left[r_h^m(s_h^m,a_h^m)+V_{h+1}^{\pi^m,m}(s')\right]}\right)\right](s_h^m,a_h^m) - e^{\beta r_h^m(s_h^m,a_h^m)}\delta_{h+1}^m,\end{aligned} \tag{20b}$$

where $[P_h^m f](s,a) := \mathbb{E}_{s'\sim P_h^m(|\cdot|s,a)}[f(s')]$ for every $f : \mathcal{S} \to \mathbb{R}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$. Then, for every $(m,h) \in [M]\times[H]$, we have

$$\begin{aligned}\delta_h^m &\overset{(i)}{=} \left(e^{\beta\cdot Q_h^m} - e^{\beta\cdot Q_h^{\pi^m,m}}\right)(s_h^m,a_h^m) \\ &\overset{(ii)}{=} q_{h,1}^m(s_h^m,a_h^m) - q_{h,2}^m(s_h^m,a_h^m) + q_{h,2}^m(s_h^m,a_h^m) - q_{h,3}^m(s_h^m,a_h^m) \\ &\overset{(iii)}{\le} 4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}} + q_{h,2}^m(s_h^m,a_h^m) - q_{h,3}^m(s_h^m,a_h^m) \\ &= 4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}} + e^{\beta\cdot r_h^m(s_h^m,a_h^m)}\delta_{h+1}^m + \zeta_{h+1}^m.\end{aligned} \tag{21}$$

In the above equation, step (i) holds by the construction of Algorithm 1 and the definition of $V_h^{\pi^m}$ in Equation (2b); step (ii) holds by Equations (11) and (12); step (iii) holds on the event of Lemma B.4; the last step follows from the definition of δ_h^m and ζ_h^m in Equations 20a and 20b.

Using the fact that $V_{H+1}^m(s) = V_{H+1}^{\pi^m}(s) = 0$, we can expand the recursion in Equation (21) to obtain

$$\begin{aligned}\delta_1^m &\le \sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + \sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\left(4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}\right) \\ &\le \sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + \sum_{h\in[H]} e^{\beta(h-1)}\left(4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}\right).\end{aligned}$$

where the last step follows from $r_h^m(\cdot,\cdot) \in [0,1]$. Summing the above display over $m \in [M]$ gives

$$\begin{aligned}&\sum_{m\in[M]}\delta_1^m \\ &\le \sum_{m\in[M]}\sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + \sum_{m\in[M]}\sum_{h\in[H]} e^{\beta(h-1)}\left(4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}\right) \\ &= \sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{h\in[H]}\left(e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + e^{\beta(h-1)}\left(4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}\right)\right) \\ &= \sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{h\in[H]}\left(e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + 4e^{\beta(h-1)}\Gamma_h^m\right) + WH\left(\left|e^{\beta H} - 1\right| B_{\mathcal{P}} + g_1(\beta)B_r\right).\end{aligned} \tag{22}$$

We aim to control the terms in (22). Since $\left\{e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m\right\}$ is a martingale difference sequence satisfying $\left|e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m\right| \le 2|e^{\beta H} - 1|$ for all $(m,h) \in [M]\times[H]$, by the Azuma-Hoeffding inequality, we have:

$$\mathcal{P}\left(\sum_{m\in[M]}\sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m \ge t\right) \le \exp\left(-\frac{t^2}{8HM\left(e^{\beta H}-1\right)^2}\right), \quad \forall t > 0.$$

Hence, with probability $1-\delta/2$, it holds that

$$\sum_{k\in[K]}\sum_{h\in[H]} e^{\beta(h-1)}\zeta_{h+1}^m \le \left(e^{\beta H}-1\right)\sqrt{2HM\log(2/\delta)} \le 2\left(e^{\beta H}-1\right)\sqrt{2HM\iota}, \tag{23}$$

where $\iota = \log(6H|\mathcal{S}||\mathcal{A}|W/\delta)$. Furthermore, recall the definition of Γ_h^m, we can derive

$$
\begin{aligned}
&\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{h\in[H]} e^{\beta(h-1)}\Gamma_h^m \\
\leq& \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{h\in[H]} \left(C_1\left|e^{\beta(H-h+1)}-1\right|+g_h(\beta)\right)\sqrt{|\mathcal{S}|\iota}\sqrt{\frac{1}{N_h^m\left(s_h^m,a_h^m\right)+1}} \\
\leq& \left(C_1\left|e^{\beta H}-1\right|+g_1(\beta)\right)\sqrt{|\mathcal{S}|\iota}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{h\in[H]}\sqrt{\frac{1}{N_h^m\left(s_h^m,a_h^m\right)+1}} \\
\overset{(i)}{\leq}& \left(C_1\left|e^{\beta H}-1\right|+g_1(\beta)\right)\sqrt{|\mathcal{S}|\iota}\sum_{h\in[H]}\sqrt{W}\sqrt{\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\frac{1}{N_h^m\left(s_h^m,a_h^m\right)+1}} \\
\leq& \left(C_1\left|e^{\beta H}-1\right|+g_1(\beta)\right)\sqrt{|\mathcal{S}|\iota}\sqrt{2H^2|S||A|W\iota}
\end{aligned}
$$

where step (i) follows the Cauchy-Schwarz inequality and the last step holds by the pigeonhole principle. Thus, it holds that

$$
\sum_{m\in[M]}\sum_{h\in[H]} e^{\beta(h-1)}\Gamma_h^m \leq \left(C_1\left|e^{\beta H}-1\right|+e^{\beta H}|\beta|\right)\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}}. \tag{24}
$$

Substituting (23) and (24) into (22) yields that

$$
\begin{aligned}
\sum_{m\in[M]}\delta_1^m \leq& 2\left|e^{\beta H}-1\right|\sqrt{2HM\iota}+\left(C_1\left|e^{\beta H}-1\right|+g_1(\beta)\right)\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}} \\
&+WH\left(\left|e^{\beta H}-1\right|B_{\mathcal{P}}+g_1(\beta)B_r\right)
\end{aligned} \tag{25}
$$

For $\beta > 0$, we have that $g_1(\beta) = e^{\beta H}\beta$ and the dynamic regret can be decomposed based on Lemma F.1:

$$
\begin{aligned}
&\text{D-Regret}(M) \\
\leq& \sum_{m\in[M]}\frac{1}{\beta}\left[e^{\beta\cdot V_1^{*,m}(s_1^m)}-e^{\beta\cdot V_1^m(s_1^m)}\right]+\sum_{m\in[M]}\frac{1}{\beta}\left[e^{\beta\cdot V_1^m(s_1^m)}-e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right] \\
\leq& \frac{1}{\beta}\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} H\left(\left|e^{\beta H}-1\right|B_{\mathcal{P},\mathcal{E}}+g_1(\beta)B_{r,\mathcal{E}}\right)+\sum_{m\in[M]}\frac{1}{\beta}\left[e^{\beta\cdot V_1^m(s_1^m)}-e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right] \\
\leq& \frac{1}{\beta}WH\left(\left|e^{\beta H}-1\right|B_{\mathcal{P}}+g_1(\beta)B_r\right)+\frac{1}{\beta}\sum_{m\in[M]}\delta_1^m \\
\leq& \frac{1}{\beta}\Bigg(2\left(e^{\beta H}-1\right)\sqrt{2HM\iota}+\left(C_1\left(e^{\beta H}-1\right)+e^{\beta H}\beta\right)\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}} \\
&+WH\left(\left(e^{\beta H}-1\right)B_{\mathcal{P}}+e^{\beta H}\beta B_r\right)\Bigg) \\
\leq& 2e^{\beta H}H\sqrt{2HM\iota}+e^{\beta H}\left(C_1H+1\right)\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}}+WHe^{\beta H}\left(HB_{\mathcal{P}}+B_r\right) \\
\leq& 2e^{\beta H}H\sqrt{2HM\iota}+(C_1+1)e^{\beta H}H\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}}+WH^2e^{\beta H}\left(B_{\mathcal{P}}+B_r\right)
\end{aligned} \tag{26}
$$

where the second inequality follows from Lemma B.6, the third inequality holds because of the definition of $B_{\mathcal{P}}$, B_r and δ_1^m, the forth inequality is due to (25), and the fifth inequality follows from $e^{\beta H}-1 \leq \beta H e^{\beta H}$ for $\beta > 0$.

Finally, by setting $W = M^{\frac{2}{3}}\left(B_{\mathcal{P}}+B_r\right)^{-\frac{2}{3}}|S|^{\frac{2}{3}}|A|^{\frac{1}{3}}$, we conclude that

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

The proof of $\beta < 0$ follows a similar procedure and is therefore omitted.

C Proof of Theorem 3.2

C.1 Preliminaries

We first lay out some additional notations to facilitate our proof. Let N_h^m, G_h^m, V_h^m be N_h, G_h, V_h at the beginning of episode m, before t is updated. We also set $Q_h^m := \frac{1}{\beta} G_h^m$. Let $\widehat{P}_h^m(\cdot \mid s,a)$ denote the delta function centered at s_{h+1}^m for all $(m,h,s,a) \in [M] \times [H] \times \mathcal{S} \times \mathcal{A}$. This means that $\mathbb{E}_{s' \sim \widehat{P}_h^m(\cdot|s,a)} [f(s')] = f(s_{h+1}^m)$ for every $f : \mathcal{S} \to \mathbb{R}$. Denote by $n_h^m := N_h^m(s_h^m, a_h^m)$. Recall from Algorithm 2 that the learning rate is defined as

$$\alpha_t := \frac{H+1}{H+t}$$

for $t \in \mathbb{Z}$. We also define

$$\alpha_t^0 := \prod_{j=1}^{t} (1-\alpha_j), \quad \alpha_t^i := \alpha_i \prod_{j=i+1}^{t} (1-\alpha_j) \tag{27}$$

for integers $i, t \geq 1$. We set $\alpha_t^0 = 1$ and $\sum_{i \in [t]} \alpha_t^i = 0$ if $t = 0$, and $\alpha_t^i = \alpha_i$ if $t < i+1$.

The epoch is defined as an interval that starts at the first episode after a restart and ends at the first time when the restart is triggered. In Algorithm 2, the restart mechanism divides M episodes into $\left\lceil \frac{M}{W} \right\rceil$ epochs.

Define the shorthand notation $\iota := \log(|\mathcal{S}||\mathcal{A}|MH/\delta)$ for $\delta \in (0,1]$. We fix a tuple $(m,h,s,a) \in [M] \times [H] \times \mathcal{S} \times \mathcal{A}$ with $m_i^{\mathcal{E}} \leq M$ being the episode in which (s,a) is visited the i-th time at step h in epoch $\mathcal{E}$. Let us define

$$q_{h,1}^{m,+}(s,a) := \alpha_t^0 e^{\beta(H-h+1)} + \begin{cases} \sum_{i \in [t]} \alpha_t^i \left[e^{\beta \left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{m_i^{\mathcal{E}}} \left(s_{h+1}^{m_i^{\mathcal{E}}} \right) \right]} + \Gamma_{h,i} \right], & \text{if } \beta > 0, \\ \sum_{i \in [t]} \alpha_t^i \left[e^{\beta \left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{m_i^{\mathcal{E}}} \left(s_{h+1}^{m_i^{\mathcal{E}}} \right) \right]} - \Gamma_{h,i} \right], & \text{if } \beta < 0, \end{cases}$$

$$q_{h,1}^{m}(s,a) := \begin{cases} \min \left\{ q_{h,1}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta > 0, \\ \max \left\{ q_{h,1}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta < 0, \end{cases}$$

and

$$q_{h,2}^{m,\circ}(s,a) := \alpha_t^0 e^{\beta(H-h+1)} + \sum_{i \in [t]} \alpha_t^i \left[e^{\beta \left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}} \left(s_{h+1}^{m_i^{\mathcal{E}}} \right) \right]} \right]$$

$$q_{h,2}^{m,+}(s,a) := \alpha_t^0 e^{\beta(H-h+1)} + \begin{cases} \sum_{i \in [t]} \alpha_t^i \left[e^{\beta \left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}} \left(s_{h+1}^{m_i^{\mathcal{E}}} \right) \right]} + \Gamma_{h,i} \right], & \text{if } \beta > 0 \\ \sum_{i \in [t]} \alpha_t^i \left[e^{\beta \left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}} \left(s_{h+1}^{m_i^{\mathcal{E}}} \right) \right]} - \Gamma_{h,i} \right], & \text{if } \beta < 0 \end{cases}$$

$$q_{h,2}^{m}(s,a) := \begin{cases} \min \left\{ q_{h,2}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta > 0 \\ \max \left\{ q_{h,2}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta < 0 \end{cases}$$

and

$$q_{h,3}^{m}(s,a) := \alpha_t^0 e^{\beta \cdot Q_h^{*,m}(s,a)} + \sum_{i \in [t]} \alpha_t^i \left[\mathbb{E}_{s' \sim P_h^m(\cdot|s,a)} e^{\beta \left[r_h^m(s,a) + V_{h+1}^{*,m}(s') \right]} \right].$$

By the definition of $q_{h,2}^{m,\circ}$, $q_{h,2}^{m,+}$ and $q_{h,2}^{m}$, it can be seen that $q_{h,2}^{m,\circ} \leq q_{h,2}^{m}$ if $\beta > 0$, and $q_{h,2}^{m,\circ} \geq q_{h,2}^{m}$ if $\beta < 0$. In addition, by definition, we have $\left(e^{\beta \cdot Q_h^m} - e^{\beta \cdot Q_h^{*,m}} \right)(s,a) = \left(q_{h,1}^m - q_{h,3}^m \right)(s,a)$.

C.2 Value difference bounds

Lemma C.1 *For every triple (s,a,h) and episodes m_1, m_2 in the epoch $\mathcal{E}$, it holds that $\left|V_h^{*,m_1}(s) - V_h^{*,m_2}(s)\right| \le B_{r,\mathcal{E}} + HB_{\mathcal{P},\mathcal{E}}$.*

Proof. Let $a_1 = \operatorname{argmax}_a Q_h^{*,m_1}(s,a)$ and $a_2 = \operatorname{argmax}_a Q_h^{*,m_2}(s,a)$, it holds that

$$\begin{aligned} V_h^{*,m_1}(s) = Q_h^{*,m_1}(s,a_1) \ge Q_h^{*,m_1}(s,a_2) \ge & Q_h^{*,m_2}(s,a_2) - B_{r,\mathcal{E}} - HB_{\mathcal{P},\mathcal{E}} \\ = & V_h^{*,m_2}(s) - B_{r,\mathcal{E}} - HB_{\mathcal{P},\mathcal{E}} \end{aligned}$$

where the second inequality follows from [31, Lemma 1]. Similarly, we have

$$V_h^{*,m_2}(s) \ge V_h^{*,m_1}(s) - B_{r,\mathcal{E}} - HB_{\mathcal{P},\mathcal{E}}.$$

This completes the proof. □

Lemma C.2 *For every $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$ and $m_1,\dots,m_t < m$ with $t = N_h^m(s,a)$, we have*

$$\begin{aligned} & \left| \sum_{i\in[t]} \alpha_t^i \left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - \mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right] \right] \right| \\ \le & \Gamma_{h,t} + 2g_h(\beta)B_{r,\mathcal{E}} + \left(g_h(\beta)H + \left|e^{\beta(H-h+1)} - 1\right|\right) B_{\mathcal{P},\mathcal{E}} \end{aligned}$$

with probability at least $1-\delta$, and

$$\sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} \in \left[\Gamma_{h,t}, 2\Gamma_{h,t}\right],$$

where $\Gamma_{h,t}$ is defined in (7)*.*

Proof. For every $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$, we have the following decomposition:

$$\begin{aligned} & e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - \mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right] \\ = & e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} && \text{(28a)} \\ & + e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} && \text{(28b)} \\ & + e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - \mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right] && \text{(28c)} \\ & + \mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right] - \mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right]. && \text{(28d)} \end{aligned}$$

For the terms in (28a), it holds that

$$\begin{aligned} \left| e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} \right| \le & g_h(\beta)\left| r_h^{m_i^{\mathcal{E}}}(s,a) - r_h^m(s,a)\right| \\ \le & g_h(\beta)B_{r,\epsilon}, \end{aligned} \tag{29}$$

where the first inequality follows from the Lipschitz continuity of $e^{\beta x}$ with respect to x and the second inequality is due to the definition of the local variation budget $B_{r,\epsilon}$.

For the terms in (28b), it holds that

$$\begin{aligned} \left| e^{\beta\left[r_h^m(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} \right| \le & g_h(\beta)\left| V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right) - V_{h+1}^{*,m}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right| \\ \le & g_h(\beta)\left(B_{r,\mathcal{E}} + HB_{\mathcal{P},\mathcal{E}}\right) \end{aligned} \tag{30}$$

where the second inequality follows from Lemma C.1.

For the terms in (28d), we have

$$\begin{aligned}
&\left|\mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right]-\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right]\right| \\
&=\left|\mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}-1\right]-\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}-1\right]\right| \\
&\le\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}} \qquad (31)
\end{aligned}$$

where the first step follows from $\mathcal{P}_h^m 1(s,a)=\mathcal{P}_h^\tau 1(s,a)$ for all $\tau\in[\ell^m,m-1]$ and the last step holds by the definition of $B_{\mathcal{P},\mathcal{E}}$.

We now analyze the terms in (28c). For every $(m,h,s,a)\in[M]\times[H]\times\mathcal{S}\times\mathcal{A}$, we define

$$\psi(i,m,h,s,a)\coloneqq e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]}-\mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right].$$

For a fix tuple $(m,h,s,a)\in[M]\times[H]\times\mathcal{S}\times\mathcal{A}$, $\{\psi(i,m,h,s,a)\}_{i\in[t]}$ with $t=N_h^m(s,a)$ is a martingale difference sequence. By the Azuma-Hoeffding inequality, with probability at least $1-\delta/(HM|\mathcal{S}||\mathcal{A}|)$, it holds that

$$\left|\sum_{i\in[t]}\alpha_t^i\cdot\psi(i,m,h,s,a)\right|\le\frac{C_2}{2}\left|e^{\beta(H-h+1)}-1\right|\sqrt{\iota\sum_{i\in[t]}\left(\alpha_t^i\right)^2}\le C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}}$$

where $C_2>0$ is some universal constant, the first step holds since $r_h(s,a)+V_{h+1}^*(s')\in[0,H-h+1]$ for $s'\in\mathcal{S}$, and the last step follows from the second property in Lemma F.3. Then, applying the union bound over $(m,h,s,a)\in[M]\times[H]\times\mathcal{S}\times\mathcal{A}$, we have that the following holds for all $(m,h,s,a)\in[M]\times[H]\times\mathcal{S}\times\mathcal{A}$ with probability at least $1-\delta$:

$$\left|\sum_{i\in[t]}\alpha_t^i\cdot\psi(i,m,h,s,a)\right|\le C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}}, \qquad (32)$$

where $t=N_h^m(s,a)$.

Finally, by combining Equations (29)-(32) and noticing that $\sum_{i\in[t]}\alpha_t^i=1$ from the forth property in Lemma F.3, we have

$$\begin{aligned}
&\left|\sum_{i\in[t]}\alpha_t^i\left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]}-\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right]\right]\right| \\
\le&C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}}+2g_h(\beta)B_{r,\mathcal{E}}+\left(g_h(\beta)H+\left|e^{\beta(H-h+1)}-1\right|\right)B_{\mathcal{P},\mathcal{E}}
\end{aligned}$$

For bounds on $\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}$, we recall the definition of $\{\Gamma_{h,t}\}$ in (7) and compute

$$\begin{aligned}
\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}&=C_2\left|e^{\beta(H-h+1)}-1\right|\sum_{i\in[t]}\alpha_t^i\sqrt{\frac{H\iota}{i}} \\
&\in\left[C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}},2C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}}\right]
\end{aligned}$$

where the last step holds by the first property in Lemma F.3. □

Lemma C.3 *For all (m,h,s,a) and $\delta\in(0,1]$, the following statements hold with probability at least $1-\delta$:*

- *If $\beta>0$:*

$$\begin{aligned}
&-2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}}-\left(e^{\beta(H-h+1)}\beta H+\left(e^{\beta(H-h+1)}-1\right)\right)B_{\mathcal{P},\mathcal{E}}\le q_{h,2}^m(s,a)-q_{h,3}^m(s,a) \\
&\le\alpha_t^0\left(e^{\beta(H-h+1)}-1\right)+2\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}+2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}}+\left(e^{\beta(H-h+1)}\beta H+\left(e^{\beta(H-h+1)}-1\right)\right)B_{\mathcal{P},\mathcal{E}}.
\end{aligned}$$

- *If $\beta < 0$:*

$$2\beta B_{r,\mathcal{E}} - \left(-\beta H + \left(1 - e^{\beta(H-h+1)}\right)\right) B_{\mathcal{P},\mathcal{E}} \le q^m_{h,3}(s,a) - q^m_{h,2}(s,a)$$
$$\le \alpha^0_t \left(1 - e^{\beta(H-h+1)}\right) + 2\sum_{i\in[t]} \alpha^i_t \Gamma_{h,i} - 2\beta B_{r,\mathcal{E}} + \left(-\beta H + \left(1 - e^{\beta(H-h+1)}\right)\right) B_{\mathcal{P},\mathcal{E}}.$$

Proof. We focus on the case where $\beta > 0$ and the case for $\beta < 0$ can be proved similarly. By the definition of $q^{m,+}_{h,2}$ and $q^m_{h,3}$, it holds that

$$q^{m,+}_{h,2} - q^m_{h,3} = \alpha^0_t \left(e^{\beta(H-h+1)} - e^{\beta Q^{*,m}_h(s,a)}\right)$$
$$+ \sum_{i\in[t]} \alpha^i_t \left[e^{\beta\left[r^{m^{\mathcal{E}}_i}_h(s,a) + V^{*,m^{\mathcal{E}}_i}_{h+1}\left(s^{m^{\mathcal{E}}_i}_{h+1}\right)\right]} + \Gamma_{h,i} - \mathbb{E}_{s'\sim P^m_h(\cdot|s,a)} e^{\beta\left[r^m_h(s,a) + V^{*,m}_{h+1}(s')\right]} \right].$$

Due to $e^{\beta(H-h+1)} \ge e^{\beta Q^{*,m}_h(s,a)} \ge 1$ and Lemma C.2, we have

$$q^{m,+}_{h,2} - q^m_{h,3} \ge -2g_h(\beta) B_{r,\mathcal{E}} - \left(g_h(\beta) H + \left|e^{\beta(H-h+1)} - 1\right|\right) B_{\mathcal{P},\mathcal{E}}$$

and

$$q^{m,+}_{h,2} - q^m_{h,3} \le \alpha^0_t \left(e^{\beta(H-h+1)} - 1\right) + 2\sum_{i\in[t]} \alpha^i_t \Gamma_{h,i}$$
$$+ 2g_h(\beta) B_{r,\mathcal{E}} + \left(g_h(\beta) H + \left|e^{\beta(H-h+1)} - 1\right|\right) B_{\mathcal{P},\mathcal{E}}.$$

Furthermore, if $q^{m,+}_{h,2} \le e^{\beta(H-h+1)}$, then we have $q^m_{h,2} = q^{m,+}_{h,2}$. On the other hand, if $q^{m,+}_{h,2} \ge e^{\beta(H-h+1)}$, then $q^m_{h,2} = e^{\beta(H-h+1)} \le q^{m,+}_{h,2}$. Thus, it holds that $0 \le q^m_{h,2} - q^m_{h,3} \le q^{m,+}_{h,2} - q^m_{h,3}$. This completes the proof. □

The next two lemmas compare the iterate $e^{\beta\cdot Q^m_h}$ (and $e^{\beta\cdot V^m_h}$) with the optimal exponential value function $e^{\beta\cdot Q^{*,m}_h}$ (and $e^{\beta\cdot V^{*,m}_h}$).

Lemma C.4 *For all (m,h,s,a) and $\delta \in (0,1]$, it holds with probability at least $1-\delta$:*

- *If $\beta > 0$:*

$$(e^{\beta Q^m_h} - e^{\beta Q^{*,m}_h})(s,a)$$
$$\ge -(H-h+1)\left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right).$$

- *If $\beta < 0$:*

$$(e^{\beta Q^m_h} - e^{\beta Q^{*,m}_h})(s,a) \le (H-h+1)\left(-2\beta B_{r,\mathcal{E}} + \left(-\beta H + \left(1 - e^{\beta(H-h+1)}\right)\right) B_{\mathcal{P},\mathcal{E}}\right).$$

Proof. We focus only on the case where $\beta > 0$ since the proof for $\beta < 0$ is similar. For the purpose of the proof, we set $Q^m_{H+1}(s,a) = Q^*_{H+1}(s,a) = 0$ for all $(m,s,a) \in [M]\times\mathcal{S}\times\mathcal{A}$. We fix a pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ and use strong induction on m and h. Without loss of generality, we assume that there exists a pair (m,h) such that $(s,a) = (s^m_h, a^m_h)$ (that is, (s,a) has been visited at some point in Algorithm 2), since otherwise $e^{\beta\cdot Q^m_h(s,a)} = e^{\beta(H-h+1)} \ge e^{\beta\cdot Q^*_h(s,a)}$ for all $(m,h) \in [M]\times[H]$ and we are done.

The base case for $m = 1$ and $h = H+1$ is satisfied since $e^{\beta\cdot Q^{m'}_{H+1}(s,a)} = e^{\beta\cdot Q^{*,m}_{H+1}(s,a)}$ for $m' \in [M]$ by definition. We fix a pair $(m,h) \in [M]\times[H]$ and assume that

$$e^{\beta\cdot Q^{m^{\mathcal{E}}_i}_{h+1}(s,a)} - e^{\beta\cdot Q^{*,m^{\mathcal{E}}_i}_{h+1}(s,a)} \ge -(H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right)$$

for each $m^{\mathcal{E}}_1, \ldots, m^{\mathcal{E}}_t$ (here $t = N^m_h(s,a)$). We have for $i \in [t]$ that

$$e^{\beta\cdot V^{m^{\mathcal{E}}_i}_{h+1}(s)} = \max_{a'\in\mathcal{A}} e^{\beta\cdot Q^{m^{\mathcal{E}}_i}_{h+1}(s,a')} - (H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right)$$
$$\ge \max_{a'\in\mathcal{A}} e^{\beta\cdot Q^{*,m^{\mathcal{E}}_i}_{h+1}(s,a')} - (H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right)$$
$$= e^{\beta\cdot V^{*,m^{\mathcal{E}}_i}_{h+1}(s)} - (H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \quad (33)$$

where the first equality holds by the update procedure in Algorithm 2. Then, it holds that

$$
\begin{aligned}
(q_{h,1}^{m,+} - q_{h,2}^{m})(s,a) &\geq (q_{h,1}^{m,+} - q_{h,2}^{m,+})(s,a) \\
&\geq \sum_{i\in[t]} \alpha_t^i \left[e^{\beta\left[r_h^{m_i^\mathcal{E}}(s,a) + V_{h+1}^{m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} - e^{\beta\left[r_h^{m_i^\mathcal{E}}(s,a) + V_{h+1}^{*,m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} \right] \\
&= \sum_{i\in[t]} \alpha_t^i e^{\beta r_h^{m_i^\mathcal{E}}(s,a)} \left[e^{\beta\left[V_{h+1}^{m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} - e^{\beta\left[V_{h+1}^{*,m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} \right] \\
&\geq -(H-h) \sum_{i\in[t]} \alpha_t^i e^{\beta} \left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \\
&\geq -(H-h) \sum_{i\in[t]} \alpha_t^i \left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \\
&\geq -(H-h) \left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right)
\end{aligned}
$$

where the first inequality follows from the definitions of $q_{h,1}^{m,+}$, $q_{h,2}^{m,+}$, the second inequality holds by the induction hypothesis, the third inequality follows from $e^\beta > 1$ for $\beta > 0$, and the last inequality holds by $\sum_{i\in[t]} \alpha_t^i \leq 1$ from Lemma F.3. Furthermore, when $q_{h,1}^m = e^{\beta(H-h+1)} \leq q_{h,1}^{m,+}$, we have $q_{h,1}^m - q_{h,2}^m \geq 0$ since $q_{h,2}^m \leq e^{\beta(H-h+1)}$ by definition. Thus, we can conclude that

$$
(q_{h,1}^{m} - q_{h,2}^{m})(s,a) \geq -(H-h) \left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \tag{34}
$$

In addition, from Lemma C.3 , we also have

$$
(q_{h,2}^{m} - q_{h,3}^{m})(s,a) \geq -2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} - \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}} \tag{35}
$$

Finally, by combining (34) and (35), we obtain

$$
\begin{aligned}
&(e^{\beta Q_h^m} - e^{\beta Q_h^{*,m}})(s,a) \\
=&(q_{h,1}^{m} - q_{h,2}^{m})(s,a) + (q_{h,2}^{m} - q_{h,3}^{m})(s,a) \\
\geq& -\left(H-h+1\right) \left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right).
\end{aligned}
$$

The induction is completed. □

Lemma C.5 *For all (m,h,s,a) and $\delta \in (0,1]$, it holds with probability at least $1-\delta$:*

- *If $\beta > 0$:*

$$
\begin{aligned}
&(e^{\beta Q_h^m} - e^{\beta Q_h^{*,m}})(s,a) \\
&\leq \sum_{i\in[t]} \alpha_t^i e^{\beta r_h^{m_i^\mathcal{E}}(s,a)} \left[e^{\beta\left[V_{h+1}^{m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} - e^{\beta\left[V_{h+1}^{*,m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} \right] + 3\sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} \\
&\quad + \alpha_t^0 \left(e^{\beta(H-h+1)} - 1\right) + 2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}.
\end{aligned}
$$

- *If $\beta < 0$:*

$$
\begin{aligned}
&(e^{\beta Q_h^m} - e^{\beta Q_h^{*,m}})(s,a) \\
&\geq \sum_{i\in[t]} \alpha_t^i e^{\beta r_h^{m_i^\mathcal{E}}(s,a)} \left[e^{\beta\left[V_{h+1}^{*,m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} - e^{\beta\left[V_{h+1}^{m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} \right] - 3\sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} \\
&\quad - \alpha_t^0 \left(1 - e^{\beta(H-h+1)}\right) + 2\beta B_{r,\mathcal{E}} - \left(-\beta H + \left(1 - e^{\beta(H-h+1)}\right)\right) B_{\mathcal{P},\mathcal{E}}.
\end{aligned}
$$

Proof. We focus on the case where $\beta > 0$ since the case for $\beta < 0$ can be proved similarly. By the definition of $q^m_{h,1}$ and $q^m_{h,2}$, we have

$$\begin{aligned}\left(q^m_{h,1} - q^m_{h,2}\right)(s,a) &\le \left(q^{m,+}_{h,1} - q^{m,\circ}_{h,2}\right)(s,a) \\ &\le \sum_{i\in[t]} \alpha^i_t \left[e^{\beta\left[r_h^{m^\mathcal{E}_i}(s,a) + V^{m^\mathcal{E}_i}_{h+1}\left(s^{m^\mathcal{E}_i}_{h+1}\right)\right]} - e^{\beta\left[r_h^{m^\mathcal{E}_i}(s,a) + V^{*,m^\mathcal{E}_i}_{h+1}\left(s^{m^\mathcal{E}_i}_{h+1}\right)\right]} \right] + \sum_{i\in[t]} \alpha^i_t \Gamma_{h,i} \\ &= \sum_{i\in[t]} \alpha^i_t e^{\beta r_h^{m^\mathcal{E}_i}(s,a)} \left[e^{\beta\left[V^{m^\mathcal{E}_i}_{h+1}\left(s^{m^\mathcal{E}_i}_{h+1}\right)\right]} - e^{\beta\left[V^{*,m^\mathcal{E}_i}_{h+1}\left(s^{m^\mathcal{E}_i}_{h+1}\right)\right]} \right] + \sum_{i\in[t]} \alpha^i_t \Gamma_{h,i}\end{aligned}$$

where the first inequality follows from $q^m_{h,1} \le q^{m,+}_{h,1}$ and $q^m_{h,2} \ge q^{m,\circ}_{h,2}$, and the second inequality holds by the definition of $q^{m,+}_{h,1}$ and $q^{m,\circ}_{h,2}$. Then, by Lemma C.3, we obtain

$$\begin{aligned}&(e^{\beta Q^m_h} - e^{\beta Q^{*,m}_h})(s,a) \\ =&(q^m_{h,1} - q^m_{h,2})(s,a) + (q^m_{h,2} - q^m_{h,3})(s,a) \\ \le& \sum_{i\in[t]} \alpha^i_t e^{\beta r_h^{m^\mathcal{E}_i}(s,a)} \left[e^{\beta\left[V^{m^\mathcal{E}_i}_{h+1}\left(s^{m^\mathcal{E}_i}_{h+1}\right)\right]} - e^{\beta\left[V^{*,m^\mathcal{E}_i}_{h+1}\left(s^{m^\mathcal{E}_i}_{h+1}\right)\right]} \right] + 3\sum_{i\in[t]} \alpha^i_t \Gamma_{h,i} \\ &+ \alpha^0_t \left(e^{\beta(H-h+1)} - 1\right) + 2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}.\end{aligned}$$

This completes the proof. □

C.3 Proof of Theorem 3.2

For now, we consider the case for $\beta > 0$. We define the following quantities to ease the notations for the proof:

$$\begin{aligned}\delta^m_h &:= e^{\beta\cdot V^m_h(s^m_h)} - e^{\beta\cdot V^{\pi^m}_h(s^m_h)}, \\ \phi^m_h &:= e^{\beta\cdot V^m_h(s^m_h)} - e^{\beta\cdot V^{*,m}_h(s^m_h)}, \\ \xi^m_{h+1} &:= \left[\left(\mathcal{P}^m_h - \widehat{\mathcal{P}}^m_h\right)\left(e^{\beta\cdot V^{*,m}_{h+1}} - e^{\beta\cdot V^{\pi^m}_{h+1}}\right)\right](s^m_h, a^m_h)\end{aligned}$$

For each fixed $(m,h) \in [M]\times[H]$, we let $t = N^m_h(s^m_h, a^m_h)$. Then, it holds that

$$\begin{aligned}\delta^m_h &\overset{(i)}{=} e^{\beta\cdot Q^m_h(s^m_h,a^m_h)} - e^{\beta\cdot Q^{\pi^m,m}_h(s^m_h,a^m_h)} \\ &= \left[e^{\beta\cdot Q^m_h(s^m_h,a^m_h)} - e^{\beta\cdot Q^{*,m}_h(s^m_h,a^m_h)}\right] + \left[e^{\beta\cdot Q^{*,m}_h(s^m_h,a^m_h)} - e^{\beta\cdot Q^{\pi^m}_h(s^m_h,a^m_h)}\right] \\ &\overset{(ii)}{=} \left[e^{\beta\cdot Q^m_h(s^m_h,a^m_h)} - e^{\beta\cdot Q^{*,m}_h(s^m_h,a^m_h)}\right] + e^{\beta\cdot r^m_h(s^m_h,a^m_h)}\left[\mathcal{P}^m_h\left(e^{\beta\cdot V^{*,m}_{h+1}} - e^{\beta\cdot V^{\pi^m,m}_{h+1}}\right)\right](s^m_h,a^m_h) \\ &\overset{(iii)}{\le} \left[e^{\beta\cdot Q^m_h(s^m_h,a^m_h)} - e^{\beta\cdot Q^{*,m}_h(s^m_h,a^m_h)}\right] + e^{\beta}\left[\mathcal{P}^m_h\left(e^{\beta\cdot V^{*,m}_{h+1}} - e^{\beta\cdot V^{\pi^m,m}_{h+1}}\right)\right](s^m_h,a^m_h) \\ &= \left[e^{\beta\cdot Q^m_h(s^m_h,a^m_h)} - e^{\beta\cdot Q^{*,m}_h(s^m_h,a^m_h)}\right] + e^{\beta}\left(\delta^m_{h+1} - \phi^m_{h+1} + \xi^m_{h+1}\right) \\ &\overset{(iv)}{\le} \alpha^0_t\left(e^{\beta(H-h+1)} - 1\right) + 3\sum_{i\in[t]}\alpha^i_t\Gamma_{h,i} + \sum_{i\in[t]}\alpha^i_t \cdot e^{\beta\cdot r_h^{m^\mathcal{E}_i}(s^m_h,a^m_h)}\left[e^{\beta\cdot V^{m^\mathcal{E}_i}_{h+1}\left(s^{m^\mathcal{E}_i}_{h+1}\right)} - e^{\beta\cdot V^{*}_{h+1}\left(s^{m^\mathcal{E}_i}_{h+1}\right)}\right] \\ &\quad + 2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}} \\ &\quad + e^{\beta}\left(\delta^m_{h+1} - \phi^m_{h+1} + \xi^m_{h+1}\right) \\ &= \alpha^0_t\left(e^{\beta(H-h+1)} - 1\right) + \sum_{i\in[t]}\alpha^i_t \cdot e^{\beta\cdot r_h^{m^\mathcal{E}_i}(s^m_h,a^m_h)}\phi^{m^\mathcal{E}_i}_{h+1} + e^{\beta}\left(\delta^m_{h+1} - \phi^m_{h+1} + \xi^m_{h+1}\right) \qquad (36) \\ &\quad + 3\sum_{i\in[t]}\alpha^i_t\Gamma_{h,i} + 2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}} \qquad (37)\end{aligned}$$

where step (i) holds since $V_h^m\left(s_h^m\right) = \max_{a' \in \mathcal{A}} Q_h^m\left(s_h^m, a'\right) = Q_h^m\left(s_h^m, a_h^m\right)$ and $V_h^{\pi^m, m}\left(s_h^m\right) = Q_h^{\pi^m, m}\left(s_h^m, \pi_h^m\left(s_h^m\right)\right) = Q_h^{\pi^m, m}\left(s_h^m, a_h^m\right)$; step (ii) holds by the exponential Bellman equation (3); step (iii) holds since $V_{h+1}^{*,m} \geq V_{h+1}^{\pi^m, m}$ implies $e^{\beta \cdot V_{h+1}^{*,m}} \geq e^{\beta \cdot V_{h+1}^{\pi^m, m}}$ given that $\beta > 0$; step (iv) holds on the event of Lemma C.5.

We bound each term in (36) and (37) one by one. First, we have

$$\begin{aligned}\sum_{m \in [M]} \alpha_{n_h^m}^0 \left(e^{\beta(H-h+1)} - 1\right) &= \left(e^{\beta(H-h+1)} - 1\right) \sum_{m \in [M]} \mathbb{1}\left\{n_h^m = 0\right\} \\ &\leq \left(e^{\beta(H-h+1)} - 1\right) |\mathcal{S}||\mathcal{A}|.\end{aligned}$$

To bound the second term in (36), we first define

$$\hat{\phi}_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m, a_h^m\right)} \coloneqq \phi_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m, a_h^m\right)} + (H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right)$$

which is non-negative from Lemma C.4 and (33):

$$\begin{aligned}\sum_{m \in [M]} \left(\sum_{i \in [t]} \alpha_t^i \cdot e^{\beta \cdot r_h^{m_i^{\mathcal{E}}\left(s_h^m, a_h^m\right)}} \phi_{h+1}^{m_i^{\mathcal{E}}}\right) &= \sum_{m \in [M]} \left(\sum_{i \in \left[n_h^m\right]} \alpha_{n_h^m}^i \cdot e^{\beta \cdot r_h^{m_i^{\mathcal{E}}\left(s_h^m, a_h^m\right)}} \phi_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m, a_h^m\right)}\right) \\ &\leq e^{\beta} \sum_{m \in [M]} \left(\sum_{i \in \left[n_h^m\right]} \alpha_{n_h^m}^i \hat{\phi}_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m, a_h^m\right)}\right)\end{aligned}$$

where $m_i^{\mathcal{E}}\left(s_h^m, a_h^m\right)$ denotes the episode in which $\left(s_h^m, a_h^m\right)$ was taken at step h for the i-th time in the epoch $\mathcal{E}$. We re-group the above summation by changing the order of the summation. For every $\hat{m}^{\mathcal{E}}$ in the epoch $\mathcal{E}$, the term $\phi_{h+1}^{\hat{m}^{\mathcal{E}}}$ appears in the summand with $m > \hat{m}^{\mathcal{E}}$ if and only if $\left(s_h^m, a_h^m\right) = \left(s_h^{m'}, a_h^{m'}\right)$ and the episode m is in the epoch $\mathcal{E}$. Since the inverse of the mapping $i \to m_i^{\mathcal{E}}\left(s_h^m, a_h^m\right)$ is $\hat{m}^{\mathcal{E}} \to n_h^{\hat{m}^{\mathcal{E}}}$, we can continue the above display as

$$\begin{aligned}e^{\beta} \sum_{m \in [M]} \left(\sum_{i \in \left[n_h^m\right]} \alpha_{n_h^m}^i \hat{\phi}_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m a_h^m\right)}\right) &\leq e^{\beta} \sum_{\mathcal{E}=1}^{\left\lceil \frac{M}{W} \right\rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \left(\sum_{i \in \left[n_h^m\right]} \alpha_{n_h^m}^i \hat{\phi}_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m a_h^m\right)}\right) \\ &\leq e^{\beta} \sum_{\mathcal{E}=1}^{\left\lceil \frac{M}{W} \right\rceil} \sum_{m'=(\mathcal{E}-1)W}^{\mathcal{E}W} \hat{\phi}_{h+1}^{m'} \left(\sum_{t \geq n_h^{m'}+1} \alpha_t^{n_h^{m'}}\right) \\ &\leq e^{\beta} \left(1 + \frac{1}{H}\right) \sum_{\mathcal{E}=1}^{\left\lceil \frac{M}{W} \right\rceil} \sum_{m'=(\mathcal{E}-1)W}^{\mathcal{E}W} \hat{\phi}_{h+1}^{m'}\end{aligned}$$

where the last step follows the third property in Lemma F.3. Collecting the above results and substituting them into (36)-(37), we have

$$\begin{aligned}
\sum_{m\in[M]}\delta_h^m \le& \left(e^{\beta(H-h+1)}-1\right)|\mathcal{S}||\mathcal{A}| + \left(1+\frac{1}{H}\right)e^{\beta}\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\hat{\phi}_{h+1}^m \\
&+\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} e^{\beta}\left(\delta_{h+1}^m-\phi_{h+1}^m+\xi_{h+1}^m\right)+3\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i} \\
&+\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}}+\left(e^{\beta(H-h+1)}\beta H+\left(e^{\beta(H-h+1)}-1\right)\right)B_{\mathcal{P},\mathcal{E}}\right) \\
\le& \left(e^{\beta(H-h+1)}-1\right)|\mathcal{S}||\mathcal{A}| + \left(1+\frac{1}{H}\right)\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}e^{\beta}\delta_{h+1}^m \\
&+\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\left(3\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}+e^{\beta}\xi_{h+1}^m\right) \\
&+3(H-h)\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}}+\left(e^{\beta(H-h+1)}\beta H+\left(e^{\beta(H-h+1)}-1\right)\right)B_{\mathcal{P},\mathcal{E}}\right) \\
\le& \left(e^{\beta(H-h+1)}-1\right)|\mathcal{S}||\mathcal{A}| + \left(1+\frac{1}{H}\right)\sum_{m\in[M]}e^{\beta}\delta_{h+1}^m \\
&+3\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}+\sum_{m\in[M]}e^{\beta}\xi_{h+1}^m \\
&+3(H-h)\left(2e^{\beta(H-h+1)}\beta W B_r+\left(e^{\beta(H-h+1)}\beta H+\left(e^{\beta(H-h+1)}-1\right)\right)W B_{\mathcal{P}}\right)
\end{aligned}$$

where the second step holds since $\delta_{h+1}^m \ge \phi_{h+1}^m$ (due to the fact that $\beta > 0$ and $V_{h+1}^{*,m} \ge V_{h+1}^{\pi^m,m}$) and the definition of $\hat{\phi}_{h+1}^m$; the last step follows from the definition of B_r and $B_{\mathcal{P}}$. Now, we unroll the quantity $\sum_{m\in[M]}\delta_h^m$ recursively in the form of Equation (36), and get

$$\begin{aligned}
&\sum_{m\in[M]}\delta_1^m \qquad (38)\\
\le& \sum_{h\in[H]}\left[\left(1+\frac{1}{H}\right)e^{\beta}\right]^{h-1}\Bigg[\left(e^{\beta(H-h+1)}-1\right)|\mathcal{S}||\mathcal{A}|+3\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}+\sum_{m\in[M]}\left(e^{\beta}\xi_{h+1}^m\right) \\
&+3(H-h)\left(2e^{\beta(H-h+1)}\beta W B_r+\left(e^{\beta(H-h+1)}\beta H+\left(e^{\beta(H-h+1)}-1\right)\right)W B_{\mathcal{P}}\right)\Bigg] \\
\le& \sum_{h\in[H]}\left(1+\frac{1}{H}\right)^{h-1}\Bigg[\left(e^{\beta H}-1\right)|\mathcal{S}||\mathcal{A}|+3\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}e^{\beta(h-1)}\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}+\sum_{m\in[M]}e^{\beta h}\xi_{h+1}^m \\
&+3(H-h)\left(2e^{\beta H}\beta W B_r+\left(e^{\beta H}\beta H+\left(e^{\beta H}-1\right)\right)W B_{\mathcal{P}}\right)\Bigg] \\
\le& e\Bigg[\left(e^{\beta H}-1\right)H|\mathcal{S}||\mathcal{A}|+3e\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{h\in[H]}e^{\beta(h-1)}\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}\Bigg]+\sum_{h\in[H]}\sum_{m\in[M]}\left(1+\frac{1}{H}\right)^{h-1}e^{\beta h}\xi_{h+1}^m \\
&+3eH^2\left(2e^{\beta H}\beta W B_r+\left(e^{\beta H}\beta H+\left(e^{\beta H}-1\right)\right)W B_{\mathcal{P}}\right)
\end{aligned}$$

where the first step uses the fact that $\delta_{H+1}^m = 0$ for $m \in [M]$; the last step holds since $(1+1/H)^h \le (1+1/H)^H \le e$ for all $h \in [H]$. Furthermore, the definition of $\Gamma_{h,i}$ and Lemma F.3 imply that

$$\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i} \le C_2\left(e^{\beta(H-h+1)-1}\right)\sqrt{\frac{H\iota}{t}}.$$

for some constant $C_2 > 0$. By the pigeonhole principle, for any $h \in [H]$ we have

$$\begin{aligned}\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{h\in[H]} e^{\beta(h-1)} \sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} &\le C_2 \left(e^{\beta H}-1\right) \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sqrt{\frac{H\iota}{n_h^m}} \\ &\le C_2 \left(e^{\beta H}-1\right) \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sqrt{W} \sqrt{\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \frac{H\iota}{n_h^m}} \\ &\le C_2 \left(e^{\beta H}-1\right) M\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/W} \end{aligned} \tag{39}$$

where the second step follows from the Cauchy-Schwarz inequality, the third step holds since $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_h^m(s,a) = W$ and the right-hand side of the second step is maximized when $N_h^m(s,a) = W/(|\mathcal{S}||\mathcal{A}|)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Finally, the Azuma-Hoeffding inequality and the fact that $\left|\left(1+\frac{1}{H}\right)^{h-1} e^{\beta h}\xi_{h+1}^m\right| \le e\left(e^{\beta H}-1\right)$ for $h \in [H]$ together imply that with probability at least $1-\delta$, we have

$$\left|\sum_{h\in[H]} \sum_{m\in[M]} \left(1+\frac{1}{H}\right)^{h-1} e^{\beta h}\xi_{h+1}^m\right| \le C_3\left(e^{\beta H}-1\right)\sqrt{HM\iota} \tag{40}$$

for some constant $C_3 > 0$. Plugging Equations (39) and (40) into (38), we have

$$\begin{aligned}\sum_{m\in[M]} \delta_1^m \le& \mathcal{O}\Big(\left(e^{\beta H}-1\right) M\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/W} + \left(e^{\beta H}-1\right)\sqrt{HM\iota} \\ &+H^2\left(2e^{\beta H}\beta W B_r + \left(e^{\beta H}\beta H + \left(e^{\beta H}-1\right)\right) W B_{\mathcal{P}}\right)\Big)\end{aligned} \tag{41}$$

when M is large enough. Invoking Lemma F.1 yields that

$$\begin{aligned}&\text{D-Regret}(M) \\ \le& \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^{*,m}(s_1^m)} - e^{\beta\cdot V_1^m(s_1^m)}\right] + \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^m(s_1^m)} - e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right] \\ \le& \frac{1}{\beta}\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} H\left(2e^{\beta H}\beta B_{r,\mathcal{E}} + \left(e^{\beta H}\beta H + \left(e^{\beta H}-1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \\ &+ \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^m(s_1^m)} - e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right] \\ \le& \frac{1}{\beta} WH\left(2e^{\beta H}\beta B_r + \left(e^{\beta H}\beta H + \left(e^{\beta H}-1\right)\right) B_{\mathcal{P}}\right) + \frac{1}{\beta}\sum_{m\in[M]} \delta_1^m \\ \le& \frac{1}{\beta} WH\left(2e^{\beta H}\beta B_r + \left(e^{\beta H}\beta H + \left(e^{\beta H}-1\right)\right) B_{\mathcal{P}}\right) \\ &+ \frac{1}{\beta}\mathcal{O}\Big(\left(e^{\beta H}-1\right) M\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/W} + \left(e^{\beta H}-1\right)\sqrt{HM\iota} \\ &+ \quad H^2\left(2e^{\beta H}\beta W B_r + \left(e^{\beta H}\beta H + \left(e^{\beta H}-1\right)\right) W B_{\mathcal{P}}\right)\Big) \\ \le& \mathcal{O}\left(e^{\beta H} HM\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/W} + e^{\beta H} H\sqrt{HM\iota} + H^2 e^{\beta H} W\left(B_r + HB_{\mathcal{P}}\right)\right) \qquad (42) \\ \le& \widetilde{\mathcal{O}}\left(e^{\beta H} M\sqrt{H^3|\mathcal{S}||\mathcal{A}|/W} + e^{\beta H}\sqrt{H^3 M} + H^3 e^{\beta H} W\left(B_r + B_{\mathcal{P}}\right)\right) \qquad (43)\end{aligned}$$

where the second step holds by (33), the third inequality holds because of the definition of $B_{\mathcal{P}}$, B_r and δ_1^m, the forth inequality is due to (41), and the fifth inequality follows from $e^{\beta H} - 1 \le \beta H e^{\beta H}$ for $\beta > 0$. Finally, by setting $W = M^{\frac{2}{3}} H^{-\frac{3}{4}} \left(B_{\mathcal{P}} + B_r\right)^{-\frac{2}{3}} |S|^{\frac{1}{3}} |A|^{\frac{1}{3}}$, we conclude that

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

The proof is similar for the case of $\beta < 0$, and one only needs to exchange the role of V_h^m, $V_h^{\pi^m,m}$ and $V_h^{*,m}$ in the definitions of $\delta_h^m, \phi_h^m, \xi_h^m$:

$$\begin{aligned}
\delta_h^m &:= e^{\beta \cdot V_h^{\pi^m}(s_h^m)} - e^{\beta \cdot V_h^m(s_h^m)}, \\
\phi_h^m &:= e^{\beta \cdot V_h^{*,m}(s_h^m)} - e^{\beta \cdot V_h^m(s_h^m)}, \\
\xi_{h+1}^m &:= \left[\left(\mathcal{P}_h^m - \widehat{\mathcal{P}}_h^m \right) \left(e^{\beta \cdot V_{h+1}^{\pi^m}} - e^{\beta \cdot V_{h+1}^{*,m}} \right) \right] (s_h^m, a_h^m)
\end{aligned}$$

to derive the counterparts of (36) and (37), and complete the remaining analysis.

D Proof of Theorem 4.1

D.1 Multi-scale ALG Initialization

Algorithm 4 Multi-scale ALG Initialization (MALG-initialization)

Inputs: ALG and its associated $\rho(\cdot)$, n;
for $\tau = 0, \ldots, 2^n - 1$ **do**
  for k=n,n-1,...,0 **do**
    If τ is a multiple of 2^k, with probability $\frac{\rho(2^n)}{\rho(2^k)}$, schedule a new instance alg of ALG that starts at $alg.s = \tau + 1$ and ends at $alg.e = \tau + 2^k$
  end for
end for

D.2 An illustrative example

For better illustration, we give an example with $n = 4$. This example has also been shown in [40] and we present here for completeness. By Algorithm 4, one possible realization of the MALG initialization is shown in Figure 2 with one order-4 instance (red), zero order-3 instance, two order-2 instances (green), two order-1 instances (purple) and five order-0 instances (blue). The bolder part of the segment indicates the period of time when the instances are active, while the thinner part indicates the inactive period. At any point of time, the active instance is always the one with the shortest length. The dashed arrow marked with ① indicates that ALG is executed as of the two sides of the arrow are concatenated. On the other hand, the two blue instances on the two sides of the dashed line marked with ② are two different order-0 instances, so the second one should start from scratch even though they are consecutive.

D.3 Preliminaries

Similar to [40], our approach takes a base algorithm that tackles the risk-sensitive RL problem when the environment is (near-)stationary, and turns it into another algorithm that can deal with non-stationary environments. The base algorithm is assumed to satisfy the following requirement:

Assumption D.1 *ALG outputs an auxiliary quantity $e^{\beta V_1^m(s_1)} \in [0, e^{\beta H}]$ at the beginning of each round m. There exist a non-stationarity measure Δ and a non-increasing function $\rho : [M] \to \mathbb{R}$ such that running ALG satisfies the following: for all $m \in [M]$, as long as $\Delta_{[1,m]} \le \rho(m)$, without knowing $\Delta_{[1,m]}$ ALG ensures with probability at least $1 - \frac{\delta}{M}$: if $\beta > 0$, it holds that*

$$e^{\beta V_1^m(s_1)} \ge \min_{\tau \in [1,m]} e^{\beta V_1^{*,\tau}(s_1)} - \Delta_{[1,m]} \quad \text{and} \quad \frac{1}{m} \sum_{\tau=1}^{m} \left(e^{\beta V_1^\tau(s_1)} - e^{\beta \sum_{h=1}^H r_h^\tau} \right) \le \rho(m) + \Delta_{[1,m]},$$

and if $\beta < 0$, it holds that

$$\max_{\tau \in [1,m]} e^{\beta V_1^{*,\tau}(s_1)} \ge e^{\beta V_1^m(s_1)} - \Delta_{[1,m]} \quad \text{and} \quad \frac{1}{m} \sum_{\tau=1}^{m} \left(e^{\beta \sum_{h=1}^H r_h^\tau} - e^{\beta V_1^\tau(s_1)} \right) \le \rho(m) + \Delta_{[1,m]},$$

Furthermore, we assume that $\rho(m) \ge \frac{1}{\sqrt{m}}$ and $C(m) = m\rho(m)$ is a non-decreasing function.

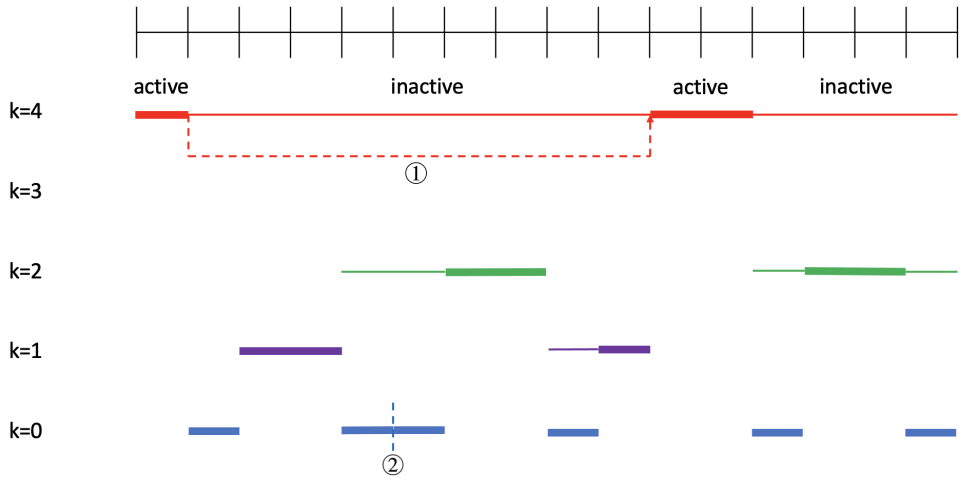


Figure 2: An illustrate example of MALG with $n = 4$.

Under Assumption D.1, the multi-scale nature of MALG allows the learner's regret to also enjoy a multi-scale structure, as shown in the next lemma:

Lemma D.2 *Let $\widehat{n} = \log_2 M + 1$ and $\widehat{\rho}(m) = 6\widehat{n}\log(M/\delta)\rho(m)$. MALG with input $n \le \log_2 M$ guarantees the following: for every instance alg that MALG maintains and every $m \in [alg.s, alg.e]$, as long as $\Delta_{[alg.s\,,t]} \le \rho(m')$ where $m' = m - alg.s + 1$, we have with probability at least $1 - \frac{\delta}{M}$: if $\beta > 0$, it holds that*

$$g_m \ge \min_{\tau\in[alg.s,m]} e^{\beta V_1^{*,\tau}(s_1)} - \Delta_{[alg.s\,,t]}, \quad \frac{1}{m'}\sum_{\tau=alg.s}^{m}\left(g_\tau - e^{\beta\sum_{h=1}^{H} r_h^\tau}\right) \le \widehat{\rho}(m') + \widehat{n}\Delta_{[alg.s,m]},$$

and if $\beta < 0$, it holds that

$$\max_{\tau\in[alg.s,m]} e^{\beta V_1^{*,\tau}(s_1)} \ge g_m - \Delta_{[alg.s\,,t]}, \quad \frac{1}{m'}\sum_{\tau=alg.s}^{m}\left(e^{\beta\sum_{h=1}^{H} r_h^\tau} - g_\tau\right) \le \widehat{\rho}(m') + \widehat{n}\Delta_{[alg.s,m]},$$

where g_m is the UCB-based optimistic estimator $e^{\beta V_1^m(s_1)}$ for the unique active instance alg at the episode m, and the number of instances started within $[alg.s, m]$ is upper bounded by $6\widehat{n}\log(M/\delta)\frac{C(m')}{C(1)}$.

Proof. The proof is similar to that of Lemma 3 in [40] with the standard value functions replaced by the exponential value functions and is thus omitted. □

Lemma D.2 states that even if there are multiple instances interleaving in a complicated way, the regret for a specific interval is still almost the same as running ALG alone on this interval, due to the carefully chosen probability $\frac{\rho(2^n)}{\rho(2^k)}$ in Algorithm 4. Built on Lemma D.2, the regret on a single block $[m_n, E_n]$, where E_n is either $m_n + 2^n - 1$ or something smaller in the case where a restart is triggered, is bounded in the following lemma:

Lemma D.3 *For Algorithm 3 with ALG satisfying Assumption D.1 and on every block $\mathcal{J} = [m_n, E_n]$ where $E_n \le m_n + 2^n - 1$, it holds with high probability that:*

$$\begin{cases}\sum_{\tau\in\mathcal{J}}\left(e^{\beta V_1^{*,\tau}(s_1)} - R_\tau\right) \le \widetilde{\mathcal{O}}\left(\sum_{i=1}^{\ell} C\left(|\mathcal{I}_i'|\right) + \sum_{m=0}^{n}\frac{\rho(2^m)}{\rho(2^n)}C\left(2^m\right)\right), & \text{if } \beta > 0,\\ \sum_{\tau\in\mathcal{J}}\left(R_\tau - e^{\beta V_1^{*,\tau}(s_1)}\right) \le \widetilde{\mathcal{O}}\left(\sum_{i=1}^{\ell} C\left(|\mathcal{I}_i'|\right) + \sum_{m=0}^{n}\frac{\rho(2^m)}{\rho(2^n)}C\left(2^m\right)\right), & \text{if } \beta < 0,\end{cases}$$

where $\{\mathcal{I}_1', \ldots, \mathcal{I}_\ell'\}$ is any partition of $\mathcal{J}$ such that $\Delta_{\mathcal{I}_i'} \le \rho(|\mathcal{I}_i'|)$ for all i.

Proof. The proof is similar to that of Lemma 4 in [40] with the standard value functions replaced by the exponential value functions and is thus omitted. □

Built on the dynamic regret over a block, we can further bound the dynamic regret over a single-epoch. The epoch is defined as an interval that starts at the first episode after a restart and ends at the first time when the restart is triggered.

Lemma D.4 *Assume that $C(m)$ takes the form of $C(m) = c_1 m^{\frac{1}{2}}$ for some constant c_1. Then, for Algorithm 3 with ALG satisfying Assumption D.1 and on every epoch $\mathcal{E}$, it holds with high probability that:*

$$\begin{cases} \sum_{\tau\in\mathcal{E}} \left(e^{\beta V_1^{*,\tau}(s_1)} - R_\tau\right) \le \widetilde{\mathcal{O}}\left(c_1^{\frac{2}{3}} \Delta_{\mathcal{E}}^{\frac{1}{3}} |\mathcal{E}|^{\frac{2}{3}} + c_1 |\mathcal{E}|^{\frac{1}{2}}\right), \text{ if } \beta > 0, \\ \sum_{\tau\in\mathcal{E}} \left(R_\tau - e^{\beta V_1^{*,\tau}(s_1)}\right) \le \widetilde{\mathcal{O}}\left(c_1^{\frac{2}{3}} \Delta_{\mathcal{E}}^{\frac{1}{3}} |\mathcal{E}|^{\frac{2}{3}} + c_1 |\mathcal{E}|^{\frac{1}{2}}\right), \text{ if } \beta < 0, \end{cases}$$

Proof. The proof is similar to that of Lemma 22 in [40] with the standard value functions replaced by the exponential value functions and is thus omitted. □

Finally, we have the following bound on the number of epoch:

Lemma D.5 (Lemma 24 in [40]) *Assume that $C(m)$ takes the form of $C(m) = c_1 m^{\frac{1}{2}}$ for some constant c_1. Then, with high probability, the number of epoch is upper-bounded by $1 + 2(c_1^{-\frac{1}{3}} \Delta^{\frac{2}{3}} M^{\frac{1}{3}})$.*

D.4 Proof of Theorem 4.1

We first focus on the case for $\beta > 0$. Let $\mathcal{E}_1, \ldots, \mathcal{E}_N$ be epochs in $[1, M]$. If Assumption D.1 holds, by Lemma D.4, the dynamic regret of the exponential value functions over M episodes is upper-bounded by

$$\begin{aligned} \sum_{m=1}^{M} \left(e^{\beta V_1^{*,m}(s_1)} - R_m\right) \le & \widetilde{\mathcal{O}}\left(\sum_{i=1}^{N} \left(c_1^{\frac{2}{3}} \Delta_{\mathcal{E}_i}^{\frac{1}{3}} |\mathcal{E}_i|^{\frac{2}{3}} + c_1 |\mathcal{E}_i|^{\frac{1}{2}}\right)\right) \\ \le & \widetilde{\mathcal{O}}\left(c_1^{\frac{2}{3}} \Delta^{\frac{1}{3}} M^{\frac{2}{3}} + c_1 N^{\frac{1}{2}} M^{\frac{1}{2}}\right) \\ \le & \widetilde{\mathcal{O}}\left(c_1^{\frac{2}{3}} \Delta^{\frac{1}{3}} M^{\frac{2}{3}}\right). \end{aligned} \tag{44}$$

where the second inequality follows from Hölder's inequality and the facts that $\sum_{i=1}^{N} \Delta_{\mathcal{E}_i} \le \Delta$ and $\sum_{i=1}^{N} |\mathcal{E}_i| \le M$, the last step holds by the bound on N from Lemma D.5.

Now, it remains to show that the base algorithms RSVI and RSQ satisfy Assumption D.1 and provide the concrete form of $\Delta(m)$, $\rho(m)$, c_1 and c_2.

- RSVI as the base algorithm: it has been shown in Lemma B.6 and (25) in the proof of Theorem 3.1 that RSVI satisfies Assumption D.1 with the following choices:

  $$\begin{aligned} \Delta(m) &= H\left(\left|e^{\beta H} - 1\right| B_{\mathcal{P},m} + g_1(\beta) B_{r,m}\right), \\ \rho(m) &= \mathcal{O}\left(\left(\left|e^{\beta H} - 1\right| + g_1(\beta)\right) \sqrt{H^2 |S|^2 |A| \iota^2 / m}\right), \\ c_1 &= \left(\left|e^{\beta H} - 1\right| + g_1(\beta)\right) \sqrt{H^2 |S|^2 |A| \iota^2}. \end{aligned}$$

  Then, by plugging in the form of Δ and c_1 in (44), and using $e^{\beta H} - 1 \le \beta H e^{\beta H}$ for $\beta > 0$, we have

  $$\sum_{m=1}^{M} \left(e^{\beta V_1^{*,m}(s_1)} - R_m\right) \le \widetilde{\mathcal{O}}\left(\beta e^{\beta H} H^2 |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}} B^{\frac{1}{3}} M^{\frac{2}{3}}\right).$$

  Invoking the above inequality with Lemma F.1 and applying Azuma's inequality to bound $\sum_{m=1}^{M} (R_m - e^{\beta V_1^{\pi^m, m}})$ yield that:

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

- RSQ as the base algorithm: it has also been shown in Lemma C.4 and (41) in the proof of Theorem 3.2 that RSQ satisfies Assumption D.1 with the following choices:

$$\Delta(m) = H\left(2g_1(\beta)B_{r,m} + \left(g_1(\beta)H + \left|e^{\beta H} - 1\right|\right)B_{\mathcal{P},m}\right)$$
$$\rho(m) = \mathcal{O}\left(\left|e^{\beta H} - 1\right|\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/m}\right),$$
$$c_1 = \mathcal{O}\left(\left|e^{\beta H} - 1\right|\sqrt{H|\mathcal{S}||\mathcal{A}|\iota}\right).$$

Then, by plugging in the form of Δ and c_1 in (44), and using $e^{\beta H} - 1 \le \beta H e^{\beta H}$ for $\beta > 0$, we have

$$\sum_{m=1}^{M}\left(e^{\beta V_1^{*,m}(s_1)} - R_m\right) \le \widetilde{\mathcal{O}}\left(\beta e^{\beta H} H^{\frac{5}{3}}|\mathcal{S}|^{\frac{1}{3}}|\mathcal{A}|^{\frac{1}{3}}B^{\frac{1}{3}}M^{\frac{2}{3}}\right).$$

Invoking the above inequality with Lemma F.1 and applying Azuma's inequality to bound $\sum_{m=1}^{M}(R_m - e^{\beta V_1^{\pi^m,m}})$ yield that:

$$\text{D-Regret(M)} \le \widetilde{\mathcal{O}}\left(e^{\beta H} H^{\frac{5}{3}}|\mathcal{S}|^{\frac{1}{3}}|\mathcal{A}|^{\frac{1}{3}}B^{\frac{1}{3}}M^{\frac{2}{3}}\right).$$

For the case of $\beta < 0$, note that from Lemma F.1, the dynamic regret can be bounded and decomposed as follows:

$$\text{D-Regret}(M) \le \frac{e^{-\beta H}}{(-\beta)}\sum_{m\in[M]}\left[e^{\beta\cdot V_1^m(s_1^m)} - e^{\beta\cdot V_1^{*,m}(s_1^m)}\right] + \frac{e^{-\beta H}}{(-\beta)}\sum_{m\in[M]}\left[e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)} - e^{\beta\cdot V_1^m(s_1^m)}\right].$$

Then, following a procedure similar to the one used for the case $\beta > 0$ and noticing that $g_1(\beta)H = -\beta H \ge 1 - e^{\beta H}$ for $\beta < 0$, we obtain the desired result.

E Proof of Theorem 5.1

E.1 Case $\beta > 0$

Consider a stochastic k-arm and M horizons bandit environment ν, where the reward for pulling arm $j \in \{1, 2, \ldots, k\}$ is given by the scaled Bernoulli random variable $Ber(p_j)$

$$X_j = \begin{cases} H, & \text{with probability } p_j, \\ 0, & \text{with probability } 1 - p_j \end{cases}$$

where $H \ge 1$ specifies the range of the reward. We let the arm i be the unique optimal arm and all the other $k-1$ arms have the same p_j, that is, $p_1 = p_2 = \cdots = p_{i-1} = p_{i+1} = \cdots = p_k = p$ and $p_i = p + \Delta$ for some constants $p > 0$ and $\Delta > 0$. Define X_j^m to be the outcome of arm j (if pulled) in round m, and Y^m to be the outcome of arm actually pulled in round m.

Lemma E.1 *For the Bernoulli bandit ν described above, if $p = e^{-\beta H}$, $\Delta \le e^{-\beta H}$ and $H \ge \frac{\log 2}{\beta}$, then for every policy π, the regret with the entropic risk measure in ν satisfies*

$$\begin{aligned}\text{Regret}(M) &\coloneqq \sum_{m=1}^{M}\frac{1}{\beta}\left(\log\left[\mathbb{E}[\exp\left(\beta X_1^m\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta Y^m\right)]\right]\right) \\ &\ge \sum_{j\in[k]/\,i}\mathbb{E}\left[T_j(M)\right]\frac{\Delta(e^{\beta H} - 1)}{4\beta}\end{aligned}$$

Proof. By the definition of $\text{Regret}(M)$, we have

$$\begin{aligned}\text{Regret(M)} &= \sum_{m=1}^{M}\frac{1}{\beta}\left(\log\left[\mathbb{E}[\exp\left(\beta X_1^m\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta Y^m\right)]\right]\right) \\ &= \sum_{j\in[k]/\,i}\frac{T_i(M)}{\beta}\left(\log\left[\mathbb{E}[\exp\left(\beta X_1\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta X_i\right)]\right]\right) \qquad (45)\end{aligned}$$

where the last step holds because of the independence among $\{X_1^m\}_{m=1}^M$ and the independence among $\{Y^m\}_{m=1}^M$. Taking the expectation over M on both sides of (45), we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Regret(M)}\right] &= \sum_{j\in[k]/\,i} \frac{\mathbb{E}\left[T_i(M)\right]}{\beta}\left(\log\left[\mathbb{E}[\exp\left(\beta X_i\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta X_j\right)]\right]\right) \\
&= \sum_{j\in[k]/\,i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta}\log\left(\frac{(p+\Delta)e^{\beta H}+(1-p-\Delta)}{pe^{\beta H}+(1-p)}\right) \\
&= \sum_{j\in[k]/\,i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta}\log\left(1+\frac{\Delta(e^{\beta H}-1)}{pe^{\beta H}+(1-p)}\right) \\
&= \sum_{j\in[k]/\,i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta}\log\left(1+\frac{\Delta(e^{\beta H}-1)}{2-e^{-\beta H}}\right) \\
&\geq \sum_{j\in[k]/\,i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta}\log\left(1+\frac{\Delta(e^{\beta H}-1)}{2}\right) \\
&\geq \sum_{j\in[k]/\,i} \mathbb{E}\left[T_j(M)\right]\frac{\Delta(e^{\beta H}-1)}{4\beta}
\end{aligned}$$

where the forth equality holds since $p = e^{-\beta H}$, the first inequality follows from $e^{\beta H} \geq 2$, and the second inequality holds since $\Delta \leq e^{-\beta H}$ and $\log(1+x) \geq \frac{x}{2}$ for $x \in [0,1]$. □

Lemma E.2 *Let $k > 1$. For every policy π and sufficiently large M and H, there exists a k-arm bandit instance such that*

$$\mathbb{E}_{\vec{p}}\left[\mathrm{Regret(M)}\right] > \frac{e^{\beta H/2}-1}{\beta}\frac{\sqrt{Mk}}{64e}.$$

Proof. Fix a policy π. Let $\Delta \in [0, e^{-\beta H}]$ be some constant to be chosen later. We start with a Bernoulli bandit where the reward of each arm is a scaled Bernoulli random variable $\mathrm{Ber}(p_i)$ with $\vec{p} \coloneqq (p_1, \ldots, p_k) = (\Delta + p, p, \ldots, p)$. This environment and the policy π give rise to the probability measure $\mathbb{P}_{\vec{p}}$ on the canonical bandit model (Section 4.6 in [29]) induced by the M-round interconnection of π and ν. Expectation under $\mathbb{P}_{\vec{p}}$ will be denoted as $\mathbb{E}_{\vec{p}}$. To choose the second environment, let

$$i = \operatorname*{argmin}_{j>1} \mathbb{E}_{\vec{p}}\left[T_j(M)\right].$$

Since $\sum_{j=1}^k \mathbb{E}_{\vec{p}}\left[T_j(M)\right] = M$, it holds that

$$\mathbb{E}_{\vec{p}}\left[T_i(M)\right] \leq \frac{M}{k-1} \tag{46}$$

The second bandit is also a Bernoulli bandit where the reward of each arm is a scaled Bernoulli random variable $\mathrm{Ber}(p_i')$ with $\vec{p}' \coloneqq (p_1', \ldots, p_k') = (\Delta + p, p, \ldots, 2\Delta + p, p, \ldots, p)$, where specifically $p_i' = 2\Delta + p$. Therefore, $p_j = p_j'$ except at index i and the optimal arm in $\nu_{\vec{p}}$ is the first arm, while in $\nu_{\vec{p}'}$ arm i is optimal. Then, Lemma E.1 and a simple calculation lead to

$$\begin{aligned}
\mathbb{E}_{\vec{p}}\left[\mathrm{Regret(M)}\right] &\geq \mathbb{P}_{\vec{p}}(T_1(M) \leq \frac{M}{2})\frac{M\Delta(e^{\beta H}-1)}{8\beta}, \\
\mathbb{E}_{\vec{p}'}\left[\mathrm{Regret(M)}\right] &> \mathbb{P}_{\vec{p}'}(T_1(M) > \frac{M}{2})\frac{M\Delta(e^{\beta H}-1)}{8\beta}.
\end{aligned}$$

Then, applying the Bretagnolle-Huber inequality in Lemma F.4 leads to

$$\begin{aligned}
&\mathbb{E}_{\vec{p}}\left[\mathrm{Regret(M)}\right] + \mathbb{E}_{\vec{p}'}\left[\mathrm{Regret(M)}\right] \\
>&\frac{M\Delta(e^{\beta H}-1)}{8\beta}\left(\mathbb{P}_{\vec{p}}(T_1(M) \leq \frac{M}{2}) + \mathbb{P}_{\vec{p}'}(T_1(M) > \frac{M}{2})\right) \\
\geq&\frac{M\Delta(e^{\beta H}-1)}{8\beta}\exp\left(-D_{\mathrm{KL}}(\mathbb{P}_{\vec{p}} \mid \mathbb{P}_{\vec{p}'})\right)
\end{aligned}$$

It remains to upper-bound $D_{\mathrm{KL}}(\mathbb{P}_{\vec{p}} \mid \mathbb{P}_{\vec{p}'})$. For this, we use Lemma F.6:

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathbb{P}_{\nu} \mid \mathbb{P}_{\nu'}\right) =& \mathbb{E}_{\mathbb{P}_{\vec{p}}}\left[T_i(M)\right] \mathrm{D}_{\mathrm{KL}}\left(\mathrm{Ber}(p_i) \mid \mathrm{Ber}(p_i')\right) \\ =& \mathbb{E}_{\mathbb{P}_{\vec{p}}}\left[T_i(M)\right] \mathrm{D}_{\mathrm{KL}}\left(p \mid 2\Delta + p\right)) \\ \leq& \mathbb{E}_{\mathbb{P}_{\vec{p}}}\left[T_i(M)\right] \cdot \frac{4\Delta^2}{(2\Delta+p)(1-2\Delta-p)} \\ \leq& \frac{M}{k-1} \cdot \frac{4\Delta^2}{(2\Delta+p)(1-2\Delta-p)} \\ \leq& \frac{16M\Delta^2}{kp} \\ \leq& \frac{16e^{\beta H}M\Delta^2}{k} \end{aligned} \tag{47}$$

where the first inequality follows from Lemma F.5, the second inequality holds by (46), the third step follows from $1-2\Delta-p \geq \frac{1}{2}$ and $k \geq 3$, and the last step holds by $p = e^{-\beta H}$.

Substituting this into the previous expression, we find that

$$\begin{aligned} \mathbb{E}_{\vec{p}}\left[\mathrm{Regret(M)}\right] + \mathbb{E}_{\vec{p}'}\left[\mathrm{Regret(M)}\right] >& \frac{M\Delta(e^{\beta H}-1)}{8\beta} \exp\left(-\frac{16e^{\beta H}M\Delta^2}{k}\right) \\ >& \frac{e^{\beta H/2}-1}{\beta}\frac{\sqrt{Mk}}{32e} \end{aligned}$$

where the second inequality holds by choosing $\Delta = \sqrt{k/(16Me^{\beta H})} \leq e^{-\beta H}$ with M sufficiently large. This result is completed by using $2\max(a,b) \geq a+b$. □

Lemma E.3 *For every policy π and sufficiently large M and H, there exists a MDP instance with horizon H, $S \geq 3$ states and A actions such that*

$$\mathbb{E}\left[\mathrm{Regret(M)}\right] > \frac{e^{\beta H/2}-1}{\beta}\frac{\sqrt{MSA}}{64e}.$$

Proof. Note that the M-round k-arm bandit model described in Lemma E.2 is a special case of an M-episode $(H+2)$-horizon MDP with S states and $\frac{S-1}{2}$ actions where $S \geq 3$ is odd. Let s_1 be the initial state, and all other states be absorbing regardless of actions taken. At the initial state s_1, we may choose to take action $a_1, a_2, \ldots, a_{\frac{S-1}{2}}$. If a_j is taken at state s_1, then we transition to state $s_{1+2(j-1)+1}$ with probability p_j and to state $s_{1+2(j-1)+2}$ with probability $1-p_j$. The reward function satisfies $r_h(s_{1+2(j-1)+1}, a) = 1$, $r_h(s_{1+2(j-1)+2}, a) = 0$ and $r_h(s_1, a) = 0$ for all $h \in [H+2]$, $a \in \mathcal{A}$ and $j = 1, \ldots, \frac{S-1}{2}$. □

Based on Lemma E.3, let us now incorporate the non-stationarity of the MDP and derive a lower bound for the dynamic regret $\mathrm{D\text{-}Regret(M)}$. We will construct the non-stationary environment as a switching-MDP. For each segment of length M_0, the environment is held constant, and the regret lower bound for each segment is $\mathcal{O}\left(\frac{e^{\beta H/2}-1}{\beta}\sqrt{SAM_0}\right)$. At the beginning of each new segment, we uniformly sample a new action at random at the state s_1 from the action space $\mathcal{A}$ to be the optimal action at the state s_1 for the new segment. In this case, the learning algorithm cannot use the information it learned during its previous interactions with the environment, even if it knows the switching structure of the environment. Therefore, the algorithm needs to learn a new (static) MDP in each segment, which leads to a dynamic regret lower bound of

$$\mathcal{O}\left(\frac{e^{\beta H/2}-1}{\beta}L\sqrt{SAM_0}\right) = \mathcal{O}\left(\frac{e^{\beta H/2}-1}{\beta}\sqrt{SAML}\right),$$

where L is the number of segments. Every time that the optimal action at the state s_1 varies, it will cause a variation of magnitude $2\Delta = \sqrt{SA/(4M_0e^{\beta H})}$ in the transition kernel. The constraint of the overall variation budget requires that

$$2\Delta L = \sqrt{\frac{SA}{4M_0e^{\beta H}}}L = \sqrt{\frac{SAL^3}{4Me^{\beta H}}} \leq B,$$

which in turn requires $L \le 4^{\frac{1}{3}} B^{\frac{2}{3}} M^{\frac{1}{3}} e^{\frac{\beta H}{3}} S^{-\frac{1}{3}} A^{-\frac{1}{3}}$. Finally, by assigning the largest possible value to L subject to the variation budget, we obtain a dynamic regret lower bound of

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

This completes the proof of Theorem 5.1 for the case $\beta > 0$.

E.2 Case $\beta < 0$

The proof of the base $\beta < 0$ is similar to that of the case $\beta > 0$. For $\beta < 0$, consider a stochastic k-arm and M horizons bandit environment ν, where the reward for pulling arm $j \in \{1, 2, \ldots, k\}$ is given by the scaled Bernoulli random variable $Ber(1 - p_j)$

$$X_j = \begin{cases} 0, & \text{with probability } p_j, \\ H, & \text{with probability } 1 - p_j \end{cases}$$

where $H \ge 1$ specifies the range of the reward. We let the arm i be the unique optimal arm and all the other $k-1$ arms have the same p_j, that is, $p_1 = p_2 = \cdots = p_{i-1} = p_{i+1} = \cdots = p_k = p$ and $p_i = p + \Delta$ for some constants $p > 0$ and $\Delta < 0$. Define X_j^m to be the outcome of arm j (if pulled) in round m, and Y^m to be the outcome of arm actually pulled in round m.

Lemma E.4 *For the Bernoulli bandit ν described above, if $p = e^{\beta H}$ and $\Delta \ge -e^{\beta H}$, then for every policy π, the regret with the entropic risk measure in ν satisfies*

$$\begin{aligned} \mathrm{Regret}(M) &\coloneqq \sum_{m=1}^{M} \frac{1}{\beta} \left(\log\left[\mathbb{E}[\exp\left(\beta X_1^m\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta Y^m\right)]\right]\right) \\ &\ge \sum_{j \in [k]/\, i} \mathbb{E}\left[T_j(M)\right] \frac{\Delta(e^{-\beta H}-1)}{2\beta} \end{aligned}$$

Proof. Taking the expectation over M on both sides of (45), we have

$$\begin{aligned} \mathbb{E}\left[\mathrm{Regret(M)}\right] &= \sum_{j \in [k]/\, i} \frac{\mathbb{E}\left[T_i(M)\right]}{\beta} \left(\log\left[\mathbb{E}[\exp\left(\beta X_i\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta X_j\right)]\right]\right) \\ &= \sum_{j \in [k]/\, i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(\frac{(1-p-\Delta)e^{\beta H} + (p+\Delta)}{(1-p)e^{\beta H} + p}\right) \\ &= \sum_{j \in [k]/\, i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(1 + \frac{\Delta(1-e^{\beta H})}{(1-p)e^{\beta H} + p}\right) \\ &\ge \sum_{j \in [k]/\, i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(1 + \frac{\Delta(1-e^{\beta H})}{2e^{\beta H}}\right) \\ &\ge \sum_{j \in [k]/\, i} \mathbb{E}\left[T_j(M)\right] \frac{\Delta(e^{-\beta H}-1)}{2\beta} \end{aligned}$$

where the first inequality holds since $p = e^{\beta H}$, the second inequality holds since $\Delta \le e^{-\beta H}$ and $\log(1+x) \le x$ for $x > -1$. □

Lemma E.5 *Let $k > 1$. For every policy π and sufficiently large M and H, there exists a k-arm bandit instance such that*

$$\mathbb{E}_{\vec{p}}\left[\mathrm{Regret(M)}\right] > \frac{e^{-\beta H/2}-1}{-\beta} \frac{\sqrt{Mk}}{64e}.$$

Proof. The proof is similar to that of Lemma E.2 by replacing Lemma E.1 with Lemma E.4, replacing (47) by

$$D_{\mathrm{KL}}\left(\mathbb{P}_\nu \mid \mathbb{P}_{\nu'}\right) = \mathbb{E}_{\mathbb{P}_{\vec{p}}}\left[T_i(M)\right] \mathrm{D}_{\mathrm{KL}}\left(\mathrm{Ber}(1-p_i) \mid \mathrm{Ber}(1-p_i')\right)$$

and by choosing $\Delta = -\sqrt{k/(16Me^{-\beta H})} \ge -e^{\beta H}$. □

The rest of the proof is similar to that for the case $\beta > 0$ and is thus omitted.

F Auxiliary lemmas

Lemma F.1 *For $\beta > 0$, the dynamic regret is bounded by*

$$\text{D-Regret}(M) \le \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^{*}(s_1^m)} - e^{\beta\cdot V_1^{m}(s_1^m)}\right] + \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^{m}(s_1^m)} - e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right],$$

and for $\beta < 0$, the dynamic regret is bounded by

$$\text{D-Regret}(M) \le \sum_{m\in[M]} \frac{e^{-\beta H}}{(-\beta)}\left[e^{\beta\cdot V_1^{m}(s_1^m)} - e^{\beta\cdot V_1^{*,m}(s_1^m)}\right] + \sum_{m\in[M]} \frac{e^{-\beta H}}{(-\beta)}\left[e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)} - e^{\beta\cdot V_1^{m}(s_1^m)}\right].$$

Proof. For $\beta > 0$, we have

$$\begin{aligned}
&\text{D-Regret}(M)\\
&= \sum_{m\in[M]} \left(V_1^{*,m} - V_1^{\pi^m,m}\right)(s_1^m)\\
&= \sum_{m\in[M]} \left(V_1^{*,m} - V_1^{m}\right)(s_1^m) + \sum_{m\in[M]} \left(V_1^{m} - V_1^{\pi^m}\right)(s_1^m)\\
&= \sum_{m\in[M]} \left[\frac{1}{\beta}\log\left\{e^{\beta\cdot V_1^{*,m}(s_1^m)}\right\} - \frac{1}{\beta}\log\left\{e^{\beta\cdot V_1^{m}(s_1^m)}\right\}\right] + \sum_{m\in[M]} \left[\frac{1}{\beta}\log\left\{e^{\beta\cdot V_1^{m}(s_1^m)}\right\} - \frac{1}{\beta}\log\left\{e^{\beta\cdot V_1^{\pi^m}(s_1^m)}\right\}\right]\\
&\le \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^{*,m}(s_1^m)} - e^{\beta\cdot V_1^{m}(s_1^m)}\right] + \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^{m}(s_1^m)} - e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right]
\end{aligned}$$

where the last step holds by the 1-Lipschitzness of the function $f(x) = \log x$ for $x \ge 1$.

For $\beta < 0$, we similarly have

$$\begin{aligned}
&\text{D-Regret}(M)\\
&= \sum_{m\in[M]} \left(V_1^{*,m} - V_1^{\pi^m,m}\right)(s_1^m)\\
&= \sum_{m\in[M]} \left(V_1^{*,m} - V_1^{m}\right)(s_1^m) + \sum_{m\in[M]} \left(V_1^{m} - V_1^{\pi^m}\right)(s_1^m)\\
&= \sum_{m\in[M]} \left[\frac{1}{-\beta}\log\left\{e^{\beta\cdot V_1^{m}(s_1^m)}\right\} - \frac{1}{-\beta}\log\left\{e^{\beta\cdot V_1^{*,m}(s_1^m)}\right\}\right]\\
&\quad + \sum_{m\in[M]} \left[\frac{1}{-\beta}\log\left\{e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right\} - \frac{1}{-\beta}\log\left\{e^{\beta\cdot V_1^{m}(s_1^m)}\right\}\right]\\
&\le \sum_{m\in[M]} \frac{e^{-\beta H}}{(-\beta)}\left[e^{\beta\cdot V_1^{m}(s_1^m)} - e^{\beta\cdot V_1^{*,m}(s_1^m)}\right] + \sum_{m\in[M]} \frac{e^{-\beta H}}{(-\beta)}\left[e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)} - e^{\beta\cdot V_1^{m}(s_1^m)}\right]
\end{aligned}$$

where the last step holds by the $\left(e^{-\beta H}\right)$-Lipschitzness of the function $f(x) = \log x$ for $x \ge e^{\beta H}$. □

Lemma F.2 (Theorem 1 in [1]) *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration and $\{\eta_t\}_{t=1}^{\infty}$ be a $\mathbb{R}$-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable for every $t \ge 0$. Assume that for every $t \ge 0$, conditioning on $\mathcal{F}_t$, η_t is a zero-mean and σ-subGaussian random variable with the variance proxy $\sigma^2 > 0$, i.e., $\mathbb{E}\left[e^{\lambda\eta_t} \mid \mathcal{F}_t\right] \le e^{\lambda^2\sigma^2/2}$ for every $\lambda \in \mathbb{R}$. Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_t$-measurable for every $t \ge 0$. Let $Y \in \mathbb{R}^{d\times d}$ be a deterministic and positive-definite matrix. For every $t \ge 0$, we define*

$$\bar{Y}_t := Y + \sum_{\tau=1}^{t} X_\tau X_\tau^{\top} \text{ and } S_t = \sum_{\tau=1}^{t} \eta_\tau X_\tau.$$

Then, for every fixed $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that

$$\|S_t\|^2_{(\bar{Y}_t)^{-1}} \le 2\sigma^2 \log\left(\frac{\det\left(\bar{Y}_t\right)^{1/2}\det(Y)^{-1/2}}{\delta}\right)$$

for every $t \ge 0$.

Lemma F.3 (Fact 1 in [20]) *The following properties hold for α_t^i defined in* (27)*:*

1. $\frac{1}{\sqrt{t}} \leq \sum_{i\in[t]} \frac{\alpha_t^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t}}$ *for every integer* $t \geq 1$.
2. $\max_{i\in[t]} \alpha_t^i \leq \frac{2H}{t}$ *and* $\sum_{i\in[t]} \left(\alpha_t^i\right)^2 \leq \frac{2H}{t}$ *for every integer* $t \geq 1$.
3. $\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H}$ *for every integer* $i \geq 1$.
4. $\sum_{i\in[t]} \alpha_t^i = 1$ *and* $\alpha_t^0 = 0$ *for every integer* $t \geq 1$*, and* $\sum_{i\in[t]} \alpha_t^i = 0$ *and* $\alpha_t^0 = 1$ *for* $t = 0$.

Lemma F.4 (Lemma 14.2 in [29]) *Let P, Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-D_{KL}(P \mid Q)\right),$$

where D_{KL} denotes the KL divergence and $A^c = \Omega / A$ is the complement of A.

Lemma F.5 (Lemma 14 in [19]) *Let $p, p' \in (0,1)$ be such that $p > p'$. We have* $D_{\mathrm{KL}}\left(\mathrm{Ber}\left(p'\right) \| \mathrm{Ber}(p)\right) \leq \frac{\left(p-p'\right)^2}{p(1-p)}$

Lemma F.6 (Divergence decomposition, Lemma 15.1 in [29]) *Let $\nu = (P_1, \ldots, P_k)$ be the reward distributions associated with one k-armed bandit, and let $\nu' = (P_1', \ldots, P_k')$ be the reward distributions associated with another k-armed bandit. Fix some policy π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ be the probability measures on the canonical bandit model (Section 4.6 in [29]) induced by the M-round interconnection of π and ν (respectively, π and ν'). Then,*

$$\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right) = \sum_{i=1}^{k} \mathbb{E}_\nu\left[T_i(M)\right] \mathrm{D}_{\mathrm{KL}}\left(P_i, P_i'\right)$$