# Near-Optimal Goal-Oriented Reinforcement Learning in Non-Stationary Environments

**Liyu Chen**
University of Southern California
liyuc@usc.edu

**Haipeng Luo**
University of Southern California
haipengl@usc.edu

## Abstract

We initiate the study of dynamic regret minimization for goal-oriented reinforcement learning modeled by a non-stationary stochastic shortest path problem with changing cost and transition functions. We start by establishing a lower bound $\Omega((B_\star SAT_\star(\Delta_c + B_\star^2\Delta_P))^{1/3}K^{2/3})$, where $B_\star$ is the maximum expected cost of the optimal policy of any episode starting from any state, $T_\star$ is the maximum hitting time of the optimal policy of any episode starting from the initial state, $SA$ is the number of state-action pairs, $\Delta_c$ and $\Delta_P$ are the amount of changes of the cost and transition functions respectively, and $K$ is the number of episodes. The different roles of $\Delta_c$ and $\Delta_P$ in this lower bound inspire us to design algorithms that estimate costs and transitions separately. Specifically, assuming the knowledge of $\Delta_c$ and $\Delta_P$, we develop a simple but sub-optimal algorithm and another more involved minimax optimal algorithm (up to logarithmic terms). These algorithms combine the ideas of finite-horizon approximation [Chen et al., 2022a], special Bernstein-style bonuses of the MVP algorithm [Zhang et al., 2020], adaptive confidence widening [Wei and Luo, 2021], as well as some new techniques such as properly penalizing long-horizon policies. Finally, when $\Delta_c$ and $\Delta_P$ are unknown, we develop a variant of the MASTER algorithm [Wei and Luo, 2021] and integrate the aforementioned ideas into it to achieve $\tilde{O}(\min\{B_\star S\sqrt{ALK}, (B_\star^2 S^2 AT_\star(\Delta_c + B_\star\Delta_P))^{1/3}K^{2/3}\})$ regret, where $L$ is the unknown number of changes of the environment.

## 1 Introduction

Goal-oriented reinforcement learning studies how to achieve a certain goal with minimal total cost in an unknown environment via sequential interactions. It has often been modeled as online learning in an episodic Stochastic Shortest Path (SSP) model, where in each episode, starting from a fixed initial state, the learner sequentially takes an action, suffers a cost, and transits to the next state, until the goal state is reached. The performance of the learner can be measured by her *regret*, generally defined as the difference between her total cost and that of a sequence of benchmark policies (one for each episode).

Despite the recent surge of studies on this problem, all previous works consider minimizing *static regret*, a special case where the benchmark policy is the same for every episode. This is reasonable only for (near) stationary environments where one single policy performs well over all episodes. In reality, however, the environment is often non-stationary with both the cost function and the transition function changing over episodes, making static regret an unreasonable metric. Instead, the desired objective is to minimize *dynamic regret*, where the benchmark policy for each episode is the optimal policy for that corresponding environment, and the hope is to obtain sublinear dynamic regret whenever the non-stationarity is not too large.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Based on this motivation, we initiate the study of dynamic regret minimization for non-stationary SSP and develop the first set of results. Specifically, our contributions are as follows:

- To get a sense on the difficulty of the problem, we start by establishing a dynamic regret lower bound in Section 3. Specifically, we prove that $\Omega((B_\star SAT_\star(\Delta_c + B_\star^2\Delta_P))^{1/3}K^{2/3})$ regret is unavoidable, where $B_\star$ is the maximum expected cost of the optimal policy of any episode starting from any state, $T_\star$ is the maximum hitting time of the optimal policy of any episode starting from the initial state, $S$ and $A$ are the number of states and actions respectively, $\Delta_c$ and $\Delta_P$ are the amount of changes of the cost and transition functions respectively, and $K$ is the number of episodes. Note the different roles of $\Delta_c$ and $\Delta_P$ here — the latter is multiplied with an extra $B_\star^2$ factor, which we find surprising for a technical reason discussed in Section 3. More importantly, this inspires us to estimate costs and transitions independently in subsequent algorithm design.
- For algorithms, we first present a simple one (Algorithm 2 in Section 5) that achieves sub-optimal regret of $\tilde{\mathcal{O}}((B_\star SAT_{\max}(\Delta_c + B_\star^2\Delta_P))^{1/3}K^{2/3})$, where $T_{\max} \geq T_\star$ is the maximum hitting time of the optimal policy of any episode starting from any state. Except for replacing $T_\star$ with the larger quantity $T_{\max}$, this bound is optimal in all other parameters. Moreover, this also translates to a minimax optimal regret bound in the finite-horizon setting (a special case of SSP), making Algorithm 2 the first model-based algorithm with the optimal $(SA)^{1/3}$ dependency.
- To improve the $T_{\max}$ dependency to $T_\star$, in Section 6, we present a more involved algorithm (Algorithm 4) that achieves a near minimax optimal regret bound matching the earlier lower bound up to logarithmic terms.
- Both algorithms above require the knowledge of $\Delta_c$ and $\Delta_P$. Moreover, for a special kind of non-stationary environments where the cost/transition function only changes $L$ times, they are not able to achieve a more favorable dynamic regret bound of the form $\sqrt{LK}$. To overcome these issues altogether, in Section 7, we develop a variant of the MASTER algorithm [Wei and Luo, 2021] and integrate the earlier algorithmic ideas into it, which finally leads to a (sub-optimal) $\tilde{\mathcal{O}}(\min\{B_\star S\sqrt{ALK}, (B_\star^2 S^2 AT_\star(\Delta_c + B_\star\Delta_P))^{1/3}K^{2/3}\})$ regret bound without knowing the non-stationarity $\Delta_c$, $\Delta_P$, or $L$.

**Techniques** All our algorithms are built on top of a finite-horizon approximation scheme first proposed by Cohen et al. [2021] and later improved by Chen et al. [2022a]; see Section 4. Both the sub-optimal Algorithm 2 and the optimal Algorithm 4 are then developed based on ideas from the MVP algorithm [Zhang et al., 2020] (for the finite-horizon setting), which adopts a UCBVI-style update rule [Azar et al., 2017] with a special Bernstein-style bonus term. The sub-optimal algorithm further integrates the idea of adaptive confidence widening [Wei and Luo, 2021] into the UCBVI-style update by subtracting a bias from the cost function uniformly over all state-action pairs, which helps control the magnitude of the estimated value function. The minimax optimal algorithm, on the other hand, adds a positive correction term to the cost function to penalize long-horizon policies, which helps improve the $T_{\max}$ dependency to $T_\star$. It also incorporates several non-stationarity tests to ensure that the algorithm resets its knowledge of the environment when the amount of non-stationarity is large. Both algorithms maintain (update and reset) cost and transition estimation independently, which is the key to achieve the correct $B_\star$ dependency for both the $\Delta_c$-related and $\Delta_P$-related terms.

To handle unknown non-stationarity, we adopt the idea of the MASTER algorithm from [Wei and Luo, 2021]. Although the nature of MASTER is a blackbox reduction, we cannot apply it directly due to the presence of the correction term that changes continuously and brings extra challenges in tracking the learner's performance. We handle this by redesigning the first non-stationarity test of the MASTER algorithm. Specifically, we maintain multiple running averages of the estimated value function to detect different levels of non-stationarity.

**Related Work** Static regret minimization in SSP has been heavily studied in recent years, for both stochastic costs [Tarbouriech et al., 2020, Cohen et al., 2020, 2021, Tarbouriech et al., 2021, Chen et al., 2021a, Jafarnia-Jahromi et al., 2021, Vial et al., 2021, Min et al., 2021, Chen et al., 2022a] and adversarial costs [Rosenberg and Mansour, 2021, Chen et al., 2021b, Chen and Luo, 2021, Chen et al., 2022b]. To the best of our knowledge, we are the first to study dynamic regret for non-stationary SSP.

There is also a surge of studies on online learning in non-stationary environments, ranging from bandits [Auer et al., 2019, Chen et al., 2019, 2021c, Russac et al., 2020, Faury et al., 2021, Abbasi-Yadkori et al., 2022, Suk and Kpotufe, 2021] to reinforcement learning [Gajane et al., 2018, Ortner et al., 2020, Cheung et al., 2020, Fei et al., 2020, Mao et al., 2021, Zhou et al., 2020, Touati and Vincent, 2020, Domingues et al., 2021, Wei and Luo, 2021, Ding and Lavaei, 2022, Lykouris et al., 2021, Wei et al., 2022]. Compared to previous work, the model we study is quite general and subsumes multi-armed bandit and finite-horizon reinforcement learning. On the other hand, it also introduces extra and unique challenges as we will discuss.

## 2 Preliminaries

A non-stationary SSP instance consists of state space $\mathcal{S}$, action space $\mathcal{A}$, initial state $s_{\text{init}} \in \mathcal{S}$, goal state $g \notin \mathcal{S}$, a set of cost mean functions $\{c_k\}_{k=1}^K$ with $c_k \in [0,1]^{\mathcal{S}\times\mathcal{A}}$, and a set of transition functions $\{P_k\}_{k=1}^K$ with $P_k = \{P_{k,s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ and $P_{k,s,a} \in \Delta_{\mathcal{S}_+}$, where $\mathcal{S}_+ = \mathcal{S} \cup \{g\}$, $\Delta_{\mathcal{S}_+}$ is the simplex over $\mathcal{S}_+$, and $K$ is the number of episodes. The set of cost and transition functions are unknown to the learner and determined by the environment before learning starts.

The learning protocol is as follows: the learner interacts with the environment for $K$ episodes. In episode $k$, starting from the initial state $s_{\text{init}}$, the learner sequentially takes an action, incurs a cost, and transits to the next state until reaching the goal state. We denote by $(s_i^k, a_i^k, c_i^k, s_{i+1}^k)$ the $i$-th state-action-cost-afterstate tuple observed in episode $k$, where $c_i^k$ is sampled from an unknown distribution with support $[0,1]$ and mean $c_k(s_i^k, a_i^k)$, and $s_{i+1}^k$ is sampled from $P_{k,s_i^k,a_i^k}$. We denote by $I_k$ the total number of steps in episode $k$, such that $s_{I_k+1}^k = g$.

**Learning Objective** Intuitively, in each episode the learner aims at finding a policy that minimizes the total cost of reaching the goal state. Formally, a policy $\pi \in \mathcal{A}^{\mathcal{S}}$ assigns an action $\pi(s)$ to each state $s \in \mathcal{S}$, and its expected cost for episode $k$ starting from a state $s$ is denoted as $V_k^\pi(s) = \mathbb{E}\big[\sum_{i=1}^{I_k} c_k(s_i^k, \pi(s_i^k)) | P_k, s_1^k = s\big]$ where the expectation is with respect to the randomness of next states $s_{i+1}^k \sim P_{k,s_i^k,\pi(s_i^k)}$ and the number of steps $I_k$ before reaching $g$. The optimal policy $\pi_k^\star$ for episode $k$ is then the policy that minimizes $V_k^\pi(s)$ for all $s$. Using $V_k^\star$ as a shorthand for $V_k^{\pi_k^\star}$, we formally define the dynamic regret of the learner as

$$R_K = \sum_{k=1}^K \left( \sum_{i=1}^{I_k} c_i^k - V_k^\star(s_{\text{init}}) \right).$$

When $I_k = \infty$ for some $k$, we let $R_K = \infty$.

**Remark 1.** *Note that our learning setting does not fall into the general non-stationary reinforcement learning framework in [Wei and Luo, 2021]. In their framework, they fix a policy to play throughout an episode, and the cost incurs by any policy is bounded. While in our case, the learner may follow several different policies within an episode. This is necessary since under unknown and changing transition, the learner may not be able to identify a proper policy (which reaches the goal state with probability* 1*) at the beginning of an episode, and committing to a single policy within an episode may lead to infinite regret.*

Several parameters play a key role in characterizing the difficulty of this problem: $B_\star = \max_{k,s} V_k^\star(s)$, the maximum cost of the optimal policy of any episode starting from any state; $T_\star = \max_k T_k^{\pi_k^\star}(s_{\text{init}})$ (where $T_k^\pi(s)$ is expected number of steps it takes for policy $\pi$ to reach the goal in episode $k$ starting from state $s$), the maximum hitting time of the optimal policy of any episode starting from the initial state; $T_{\max} = \max_{k,s} T_k^{\pi_k^\star}(s)$, the maximum hitting time of the optimal policy of any episode starting from any state; $\Delta_c = \sum_{k=1}^{K-1} \|c_{k+1} - c_k\|_\infty$, the amount of non-stationarity in the cost functions; and finally $\Delta_P = \sum_{k=1}^{K-1} \max_{s,a} \|P_{k+1,s,a} - P_{k,s,a}\|_1$, the amount of non-stationarity in the transition functions. Throughout the paper we assume the knowledge of $B_\star$, $T_\star$, and $T_{\max}$, and also $B_\star \geq 1$ for simplicity. $\Delta_c$ and $\Delta_P$ are assumed to be known for the first two algorithms we develop, but unknown for the last one.

**Other Notations** For a value function $V \in \mathbb{R}^{\mathcal{S}_+}$ and a distribution $P$ over $\mathcal{S}_+$, define $PV = \mathbb{E}_{s'\sim P}[V(s')]$ (mean) and $\mathbb{V}(P,V) = \mathbb{E}_{s'\sim P}[V(s')^2] - (PV)^2$ (variance). Let $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$ be the number of states and actions respectively. The notation $\tilde{\mathcal{O}}(\cdot)$ hides all logarithmic dependency including $\ln K$ and $\ln \frac{1}{\delta}$ for some failure probability $\delta \in (0,1)$. Also define a value function upper bound $B = 16B_\star$. For integers $s$ and $e$, we define $[s,e] = \{s, s+1, \dots, e\}$ and $[e] = \{1, \dots, e\}$.

**Algorithm 1** Finite-Horizon Approximation of SSP

**Input:** Algorithm $\mathfrak{A}$ for finite-horizon MDP $\mathring{\mathcal{M}}$ with horizon $H = 4T_{\max}\ln(8K)$.
**Initialize:** interval counter $m \leftarrow 1$.
**for** $k = 1, \ldots, K$ **do**
1 Set $s_1^m \leftarrow s_{\text{init}}$.
2 **while** $s_1^m \neq g$ **do**
3 Feed initial state $s_1^m$ to $\mathfrak{A}$, $h \leftarrow 1$.
4 **while** *True* **do**
5 Receive action $a_h^m$ from $\mathfrak{A}$, play it, and observe cost $c_h^m$ and next state $s_{h+1}^m$.
6 Feed $c_h^m$ and $s_{h+1}^m$ to $\mathfrak{A}$.
7 **if** $h = H$ *or* $s_{h+1}^m = g$ *or* $\mathfrak{A}$ *requests to start a new interval* **then**
8 $H_m \leftarrow h$. **break**.
9 **else** $h \leftarrow h + 1$.
10 Set $s_1^{m+1} = s_{H_m+1}^m$ and $m \leftarrow m + 1$.

## 3 Lower Bound

To better understand the difficulty of learning non-stationary SSP, we first establish the following dynamic regret lower bound.

**Theorem 1.** *In the worst case, the learner's regret is at least* $\Omega((B_\star SAT_\star(\Delta_c + B_\star^2\Delta_P))^{1/3}K^{2/3})$.

The lower bound construction is similar to that in [Mao et al., 2021], where the environment is piecewise stationary. In each stationary period, the learner is facing a hard SSP instance with a slightly better hidden state. Details are deferred to Appendix B.2.

In a technical lemma in Appendix B.1, we show that for any two episodes $k_1$ and $k_2$, the change of the optimal value function due to non-stationarity satisfies $V^\star_{k_1}(s_{\text{init}}) - V^\star_{k_2}(s_{\text{init}}) \leq (\Delta_c + B_\star\Delta_P)T_\star$, with only one extra $B_\star$ factor for the $\Delta_P$-related term. We thus find our lower bound somewhat surprising since an extra $B_\star^2$ factor shows up for the $\Delta_P$-related term. This comes from the fact that constructing the hard instance with perturbed costs requires a larger amount of perturbation compared to that with perturbed transitions; see Theorem 7 and Theorem 8 for details.

More importantly, this observation implies that simply treating these two types of non-stationarity as a whole and only consider the non-stationarity in value function as done in [Wei and Luo, 2021] does not give the right $B_\star$ dependency. This further inspires us to consider cost and transition estimation independently in our subsequent algorithm design.

## 4 Basic Framework: Finite-Horizon Approximation

Our algorithms are all built on top of the finite-horizon approximation scheme of [Cohen et al., 2021], whose analysis is greatly simplified and improved by [Chen et al., 2022a], making it applicable to our non-stationary setting as well. This scheme makes use of an algorithm $\mathfrak{A}$ that deals with a special case of SSP where each episode ends within $H = \tilde{\mathcal{O}}(T_{\max})$ steps, and applies it to the original SSP following Algorithm 1. Specifically, call each "mini-episode" $\mathfrak{A}$ is facing an *interval*. At each step $h$ of interval $m$, the learner receives the decision $a_h^m$ from $\mathfrak{A}$, takes this action, observes the cost $c_h^m$, transits to the next state $s_{h+1}^m$, and then feed the observation $c_h^m$ and $s_{h+1}^m$ to $\mathfrak{A}$ (Line 5 and Line 6). The interval $m$ ends whenever one of the following happens (Line 7): the goal state is reached, $H$ steps have passed, or $\mathfrak{A}$ requests to start a new interval.[1] In the first case, the initial state $s_1^{m+1}$ of the next interval $m+1$ will be set to $s_{\text{init}}$, while in the other two cases, it is naturally set to the learner's current state, which is also $s_{H_m+1}^m$ where $H_m$ is the length of interval $m$ (see Line 10). At the end of each interval, we artificially let $\mathfrak{A}$ suffer a *terminal cost* $c_f(s_{H_m+1}^m)$ where $c_f(s) = 2B_\star\mathbb{I}\{s \neq g\}$.

[1]This last condition is not present in prior works. We introduce it since later our instantiation of $\mathfrak{A}$ will change its policy in the middle of an interval, and creating a new interval in this case allows us to make sure that the policy in each interval is always fixed, which simplifies the analysis.

**Algorithm 2** Non-Stationary MVP

**Parameters:** window sizes $W_c$ (for costs) and $W_P$ (for transitions), and failure probability $\delta$.
**Initialize:** for all $(s, a, s')$, $\mathbf{C}(s, a) \leftarrow 0$, $\mathbf{M}(s, a) \leftarrow 0$, $\mathbf{N}(s, a) \leftarrow 0$, $\mathbf{N}(s, a, s') \leftarrow 0$.
**Initialize:** `Update(1)`.
**for** $m = 1, \ldots, M$ **do**
  **for** $h = 1, \ldots, H$ **do**
1    Play action $a_h^m \leftarrow \operatorname{argmin}_a Q_h(s_h^m, a)$, receive cost $c_h^m$ and next state $s_{h+1}^m$.
    $\mathbf{C}(s_h^m, a_h^m) \xleftarrow{+} c_h^m$, $\mathbf{M}(s_h^m, a_h^m) \xleftarrow{+} 1$, $\mathbf{N}(s_h^m, a_h^m) \xleftarrow{+} 1$, $\mathbf{N}(s_h^m, a_h^m, s_{h+1}^m) \xleftarrow{+} 1$.[2]
2    **if** $s_{h+1}^m = g$ *or* $\mathbf{M}(s_h^m, a_h^m) = 2^l$ *or* $\mathbf{N}(s_h^m, a_h^m) = 2^l$ *for some integer* $l \geq 0$ **then**
      **break** (which starts a new interval).
3  **if** $W_c$ *divides* $m$ **then** reset $\mathbf{C}(s, a) \leftarrow 0$ and $\mathbf{M}(s, a) \leftarrow 0$ for all $(s, a)$.
4  **if** $W_P$ *divides* $m$ **then** reset $\mathbf{N}(s, a, s') \leftarrow 0$ and $\mathbf{N}(s, a) \leftarrow 0$ for all $(s, a, s')$.
  `Update(`$m+1$`)`.

**Procedure** `Update(`$m$`)`
  $V_{H+1}(s) \leftarrow 2B_\star \mathbb{I}\{s \neq g\}$, $V_h(g) \leftarrow 0$ for $h \leq H$, $\iota \leftarrow 2^{11} \cdot \ln\left(\frac{2SAHKm}{\delta}\right)$, and $x \leftarrow \frac{1}{mH}$.
  **for** *all* $(s, a)$ **do**
    $\mathbf{N}^+(s, a) \leftarrow \max\{1, \mathbf{N}(s, a)\}$, $\mathbf{M}^+(s, a) \leftarrow \max\{1, \mathbf{M}(s, a)\}$, $\bar{c}(s, a) \leftarrow \frac{\mathbf{C}(s,a)}{\mathbf{M}^+(s,a)}$,
    $\widehat{c}(s, a) \leftarrow \max\left\{0, \bar{c}(s, a) - \sqrt{\frac{\bar{c}(s,a)\iota}{\mathbf{M}^+(s,a)}} - \frac{\iota}{\mathbf{M}^+(s,a)}\right\}$, $\bar{P}_{s,a}(\cdot) \leftarrow \frac{\mathbf{N}(s,a,\cdot)}{\mathbf{N}^+(s,a)}$.
  **while** *True* **do**
    **for** $h = H, \ldots, 1$ **do**
5      $b_h(s, a) \leftarrow \max\left\{7\sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V_{h+1})\iota}{\mathbf{N}^+(s,a)}}, \frac{49B\sqrt{S}\iota}{\mathbf{N}^+(s,a)}\right\}$ for all $(s, a)$.
6      $Q_h(s, a) \leftarrow \max\{0, \widehat{c}(s, a) + \bar{P}_{s,a} V_{h+1} - b_h(s, a) - x\}$ for all $(s, a)$.
      $V_h(s) \leftarrow \min_a Q_h(s, a)$ for all $s$.
7    **if** $\max_{s,a,h} Q_h(s, a) \leq B/4$ **then break**; **else** $x \leftarrow 2x$.

This procedure (adaptively) generates a non-stationary finite-horizon Markov Decision Process (MDP) that $\mathfrak{A}$ faces: $\mathring{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, g, \{c^m\}_{m=1}^M, \{P^m\}_{m=1}^M, c_f, H)$. Here, $c^m = c_{k(m)}$ and $P^m = P_{k(m)}$ where $k(m)$ is the unique episode that interval $m$ belongs to, and $M$ is the total number of intervals over $K$ episodes, a random variable determined by the interactions. Note that $c^m$ and $P^m$ always lie in the oblivious sets $\{c_k\}_{k=1}^K$ and $\{P_k\}_{k=1}^K$ respectively, but $c^m$ and $P^m$ are not oblivious since their values depend on the interaction history. Let $V_1^{\pi,m}(s)$ be the expected cost (including the terminal cost) of following policy $\pi$ starting from state $s$ in interval $m$. Define the regret of $\mathfrak{A}$ over the first $M'$ intervals in $\mathring{\mathcal{M}}$ as $\mathring{R}_{M'} = \sum_{m=1}^{M'}(\sum_{h=1}^{H_m+1} c_h^m - V_1^{\pi^\star_{k(m)},m}(s_1^m))$ where we use $c_{H_m+1}^m$ as a shorthand for the terminal cost $c_f(s_{H_m+1}^m)$. Following similar arguments as in [Cohen et al., 2021, Chen et al., 2022a], the regret in $\mathcal{M}$ and $\mathring{\mathcal{M}}$ are close in the following sense.

**Lemma 1.** *Algorithm 1 ensures* $R_K \leq \mathring{R}_M + B_\star$.

See Appendix C for the proof. Based on this lemma, in following sections we focus on developing the finite-horizon algorithm $\mathfrak{A}$ and analyzing how large $\mathring{R}_M$ is. Note, however, that while this finite-horizon reduction is very useful, it does not mean that our problem is as easy as learning non-stationary finite-horizon MDPs and that we can directly plug in an existing algorithm as $\mathfrak{A}$. Great care is still needed when designing $\mathfrak{A}$ in order to obtain tight regret bounds as we will show.

## 5 A Simple Sub-Optimal Algorithm

In this section, we present a relatively simple finite-horizon algorithm $\mathfrak{A}$ for $\mathring{\mathcal{M}}$ which, in combination with the reduction of Algorithm 1, achieves a regret bound that almost matches our lower bound except that $T_\star$ is replaced by $T_{\max}$. The key steps are shown in Algorithm 2. It follows the ideas

[2] $z \xleftarrow{+} y$ is a shorthand for $z \leftarrow z + y$.

of the MVP algorithm [Zhang et al., 2020] and adopts a UCBVI-style update rule (Line 6) with a Bernstein-type bonus term (Line 5) to maintain a set of $Q_h$ functions, which then determines the action at each step in a greedy manner (Line 1). The two crucial new elements are the following. First, in the update rule Line 6, we subtract a positive value $x$ uniformly over all state-action pairs so that $\|Q_h\|_\infty$ is of order $\mathcal{O}(B_\star)$ (recall $B = 16B_\star$), and we find the (almost) smallest such $x$ via a doubling trick (Line 7). This is similar to the adaptive confidence widening technique of [Wei and Luo, 2021], where they increase the size of the transition confidence set to ensure a bounded magnitude on the estimated value function; our approach is an adaptation of their idea to the UCBVI style update rule.

Second, we periodically restart the algorithm (by resetting some counters and statistics) in Line 3 and Line 4. While periodic restart is a standard idea to deal with non-stationarity, the novelty here is a two-scale restart schedule: we set one window size $W_c$ related to costs and another one $W_P$ related to transitions, and restart after every $W_c$ intervals or every $W_P$ intervals. As mentioned, this two-scale schedule is inspired by the lower bound in Section 3, which indicates that cost estimation and transition estimation play different roles in the final regret and should be treated separately.

Another small modification is that we start a new interval when the visitation to some $(s, a)$ doubles (Line 2), which helps remove $T_{\max}$ dependency in lower-order terms and is important for following sections. With all these elements, we prove the following regret guarantee of Algorithm 2.

**Theorem 2.** *For any $M' \leq M$, with probability at least $1 - 22\delta$ Algorithm 2 ensures* $\mathring{R}_{M'} = \tilde{\mathcal{O}}\left(M'\left(\sqrt{B_\star SA\left(1/W_c + B_\star/W_P\right)} + B_\star SA\left(1/W_c + S/W_P\right)\right) + (\Delta_c W_c + B_\star \Delta_P W_P)T_{\max}\right)$.

Thus, with a proper tunning of $W_c$ and $W_P$ (that is in term of $M'$), Algorithm 2 ensures $\mathring{R}_{M'} = \tilde{\mathcal{O}}((B_\star SAT_{\max}(\Delta_c + B_\star^2\Delta_P))^{1/3}M'^{2/3})$. However, this does not directly imply a bound on $\mathring{R}_M$ since $M$ is a random variable (and the tunning above would depend on $M$). Fortunately, to resolve this it suffices to perform a doubling trick on the number of intervals, that is, first make a guess on $M$, and then double the guess whenever $M$ exceeds it. We summarize this idea in Algorithm 3. Finally, combining it with Algorithm 1, Lemma 1, and the simplified analysis of [Chen et al., 2022a] which is able to bound the total number of intervals $M$ in terms of the total number of episodes $K$ (Lemma 16), we obtain the following result (all proofs are deferred to Appendix D).

**Theorem 3.** *With probability at least $1-22\delta$, applying Algorithm 1 with $\mathfrak{A}$ being Algorithm 3 ensures $R_{K'} = \tilde{\mathcal{O}}((B_\star SAT_{\max}(\Delta_c + B_\star^2\Delta_P))^{1/3}K'^{2/3})$ (ignoring lower order terms) for any $K' \leq K$.*

Note that Theorem 3 actually provides an anytime regret guarantee (that is, holds for any $K' \leq K$), which is important in following sections. Compared to our lower bound in Theorem 1, the only sub-optimality is in replacing $T_\star$ with the larger quantity $T_{\max}$. Despite its sub-optimality for SSP, however, as a side result our algorithm in fact implies the first model-based finite-horizon algorithm that achieves the optimal dependency on $SA$ and matches the minimax lower bound of [Mao et al., 2021]. Specifically, in previous works, the optimal $SA$ dependency is only achievable by model-free algorithms, which unfortunately have sub-optimal dependency on the horizon by the current analysis (see [Mao et al., 2021, Lemma 10]). On the other hand, existing model-based algorithms for finite state-action space all follow the idea of extended value iteration, which gives sub-optimal dependency on $S$ and also brings difficulty in incorporating entry-wise Bernstein confidence sets.[3] Our approach, however, resolves all these issues. See Appendix D.4 for more discussions.

**Technical Highlights** The key step of our proof for Theorem 2 is to bound the term $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P^m_{s^m_h,a^m_h}, V^{\star,m}_{h+1} - V^m_{h+1})$, where $V^m_{h+1}$ is the value of $V_{h+1}$ at the beginning of interval $m$, and $V^{\star,m}_{h+1}$ is the optimal value function of $\mathring{\mathcal{M}}$ in interval $m$ (formally defined in Appendix A). The standard analysis on bounding this term requires $V^{\star,m}_{h+1}(s) - V^m_{h+1}(s) \geq 0$, which is only true in a stationary environment due to optimism. To handle this in non-stationarity environments, we carefully choose a set of constants $\{z^m_h\}$ so that $V^{\star,m}_{h+1}(s) + z^m_h - V^m_{h+1}(s) \geq 0$ (Lemma 18), and then apply similar analysis on $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P^m_{s^m_h,a^m_h}, V^{\star,m}_{h+1} - V^m_{h+1}) = \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P^m_{s^m_h,a^m_h}, V^{\star,m}_{h+1} + z^m_h - V^m_{h+1})$. See Lemma 20 for more details.

[3]Note that the transition non-stationarity $\Delta_P$ is defined via $L_1$ norm. Thus, naively applying entry-wise confidence widening to Bernstein confidence sets introduces extra dependency on $S$.

**Algorithm 3** Non-Stationary MVP with a Doubling Trick

**for** $n = 1, 2, \ldots$ **do**

  Initialize an instance of Algorithm 2 with $W_c = \lceil (B_\star SA)^{1/3}(2^{n-1}/(\Delta_c T_{\max}))^{2/3} \rceil$ and $W_P = \lceil (SA)^{1/3}(2^{n-1}/(\Delta_P T_{\max}))^{2/3} \rceil$, and execute it in intervals $m = 2^{n-1}, \ldots, 2^n - 1$.

**Algorithm 4** MVP with Non-Stationarity Tests

**Parameters:** window sizes $W_c$ and $W_P$, coefficients $c_1$, $c_2$, sample probability $p$, and failure probability $\delta$.

**Initialize:** `ResetC()`, `ResetP()`, `Update(1)`.

**for** $m = 1, \ldots, M$ **do**

  **for** $h = 1, \ldots, H$ **do**

    Play action $a_h^m \leftarrow \operatorname{argmin}_a \check{Q}_h(s_h^m, a)$, receive cost $c_h^m$ and next state $s_{h+1}^m$.

    $\mathbf{C}(s_h^m, a_h^m) \overset{+}{\leftarrow} c_h^m$, $\mathbf{M}(s_h^m, a_h^m) \overset{+}{\leftarrow} 1$, $\mathbf{N}(s_h^m, a_h^m) \overset{+}{\leftarrow} 1$, $\mathbf{N}(s_h^m, a_h^m, s_{h+1}^m) \overset{+}{\leftarrow} 1$.

1    $\widehat{\chi}^c \overset{+}{\leftarrow} c_h^m - \widehat{c}(s_h^m, a_h^m)$, $\widehat{\chi}^P \overset{+}{\leftarrow} \check{V}_{h+1}(s_{h+1}^m) - \bar{P}_{s_h^m, a_h^m} \check{V}_{h+1}$.

    **if** $s_{h+1}^m = g$ *or* $\mathbf{M}(s_h^m, a_h^m) = 2^l$ *or* $\mathbf{N}(s_h^m, a_h^m) = 2^l$ *for some integer* $l \geq 0$ **then**

      **break** (which start a new interval).

2  **if** $\widehat{\chi}^c > \chi_m^c$ *(defined in Lemma 24)* **then** `ResetC()`. **(Test 1)**

3  **if** $\widehat{\chi}^P > \chi_m^P$ *(defined in Lemma 25)* **then** `ResetC()` and `ResetP()`. **(Test 2)**

4  **if** $\nu^c = W_c$ **then** `ResetC()`.

5  **if** $\nu^P = W_P$ **then** `ResetC()` and `ResetP()`.

  $\nu^c \overset{+}{\leftarrow} 1$, $\nu^P \overset{+}{\leftarrow} 1$, `Update`$(m+1)$.

6  **if** $\left\|\check{V}_h\right\|_\infty > B/2$ *for some* $h$ ***(Test 3)*** **then**

    `ResetC()`, with probability $p$ execute `ResetP()`, and `Update`$(m+1)$.

**Procedure** `Update`$(m)$

  $\check{V}_{H+1}(s) \leftarrow 2B_\star \mathbb{I}\{s \neq g\}$, $\check{V}_h(g) \leftarrow 0$ for all $h \leq H$, and $\iota \leftarrow 2^{11} \cdot \ln\left(\frac{2SAHKm}{\delta}\right)$.

7  $\rho^c \leftarrow \min\{\frac{c_1}{\sqrt{\nu^c}}, \frac{1}{2^8 H}\}$, $\rho^P \leftarrow \min\{\frac{c_2}{\sqrt{\nu^P}}, \frac{1}{2^8 H}\}$, $\eta \leftarrow \rho^c + B\rho^P$.

  **for** *all* $(s, a)$ **do**

    $\mathbf{N}^+(s,a) \leftarrow \max\{1, \mathbf{N}(s,a)\}$, $\mathbf{M}^+(s,a) \leftarrow \max\{1, \mathbf{M}(s,a)\}$, $\bar{c}(s,a) \leftarrow \frac{\mathbf{C}(s,a)}{\mathbf{M}^+(s,a)}$,

    $\bar{P}_{s,a}(\cdot) \leftarrow \frac{\mathbf{N}(s,a,\cdot)}{\mathbf{N}^+(s,a)}$, $\widehat{c}(s,a) \leftarrow \max\left\{0, \bar{c}(s,a) - \sqrt{\frac{\bar{c}(s,a)\iota}{\mathbf{M}^+(s,a)}} - \frac{\iota}{\mathbf{M}^+(s,a)}\right\}$,

8    $\check{c}(s,a) \leftarrow \widehat{c}(s,a) + 8\eta$.

  **for** $h = H, \ldots, 1$ **do**

    $b_h(s,a) \leftarrow \max\left\{7\sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, \check{V}_{h+1})\iota}{\mathbf{N}^+(s,a)}}, \frac{49B\sqrt{S}\iota}{\mathbf{N}^+(s,a)}\right\}$ for all $(s,a)$.

    $\check{Q}_h(s,a) = \max\{0, \check{c}(s,a) + \bar{P}_{s,a}\check{V}_{h+1} - b_h(s,a)\}$ all $(s,a)$.

    $\check{V}_h(s) = \operatorname{argmin}_a \check{Q}_h(s,a)$ for all $s$.

**Procedure** `ResetC()`

  $\nu^c \leftarrow 1$, $\widehat{\chi}^c \leftarrow 0$, $\mathbf{C}(s,a) \leftarrow 0$, $\mathbf{M}(s,a) \leftarrow 0$ for all $(s,a)$.

**Procedure** `ResetP()`

  $\nu^P \leftarrow 1$, $\widehat{\chi}^P \leftarrow 0$, $\mathbf{N}(s,a,s') \leftarrow 0$, $\mathbf{N}(s,a) \leftarrow 0$ for all $(s,a,s')$.

## 6 A Minimax Optimal Algorithm

In this section, we present an improved algorithm that achieves the minimax optimal regret bound up to logarithmic terms, starting with a refined version of Algorithm 2 shown in Algorithm 4. Below, we focus on describing the new elements introduced in Algorithm 4 (that is, Lines 1-3 and 6-4).[4]

The main challenge in replacing $T_{\max}$ with $T_\star$ is that the regret due to non-stationarity accumulates along the learner's trajectory, which can be as large as $\mathcal{O}((\Delta_c + B_\star\Delta_P)H)$ since the horizon is $H$

[4] Line 4 and Line 5, althogh written in a different form, are similar to Line 3 and Line 4 of Algorithm 2.

**Algorithm 5** A Two-Phase Variant of Algorithm 1

**Initialize:** Phase 1 algorithm instance $\mathfrak{A}_1$ and Phase 2 algorithm instance $\mathfrak{A}_2$.
Execute Algorithm 1 with $\mathfrak{A} = \mathfrak{A}_1$ for every first interval of an episode, and $\mathfrak{A} = \mathfrak{A}_2$ otherwise.

(recall $H = \tilde{\mathcal{O}}(T_{\max})$). Moreover, bounding the number of steps needed for the learner's policy to reach the goal is highly non-trivial due to the changing transitions. Our main idea to address these issues is to incorporate a correction term $\eta$ (computed in Line 7) into the estimated cost (Line 4) to penalize policies that take too long to reach the goal. This correction term is set to be an upper bound of the learner's average regret per interval (defined through $\rho^c$ and $\rho^P$ in Line 7). It introduces the effect of canceling the non-stationarity along the learner's trajectory when it is not too large. When the non-stationarity is large, on the other hand, we detect it through two non-stationary tests (Line 2 and Line 3), and reset the knowledge of the environment (more details to follow).

However, this correction leads to one issue: we cannot perform adaptive confidence widening (that is, the $-x$ bias) anymore as it would cancel out the correction term. To address this, we introduce another test (Line 6, **Test 3**) to directly check whether the magnitude of the estimated value function is bounded as desired. If not, we reset again since that is also an indication of large non-stationarity.

We now provide some intuitions on the design of **Test 1** and **Test 2**. First, one can show that the two quantities $\widehat{\chi}^c$ and $\widehat{\chi}^P$ we maintain in Line 1 are such that their sum is roughly an upper bound on the estimated accumulated regret. So directly checking whether $\widehat{\chi}^c + \widehat{\chi}^P$ is too large would be similar to the second test of the MASTER algorithm [Wei and Luo, 2021]. Here, however, we again break it into two tests where **Test 1** only guards the non-stationarity in cost, and **Test 2** mainly guards the non-stationarity in transition. Note that **Test 2** also involves cost information through $\check{V}$, but our observation is that we can still achieve the desired regret bound as long as the ratio of the number of resets caused by procedures ResetC() and ResetP() is of order $\tilde{\mathcal{O}}(B_\star)$. This inspires us to reset both the cost and the transition estimation when **Test 2** fails, but reset the transition estimation only with some probability $p$ (eventually set to $1/B_\star$) when **Test 3** fails.

For analysis, we first establish a regret guarantee of Algorithm 4 in an ideal situation where the first state of each interval is always $s_{\text{init}}$. (Proofs of this section are deferred to Appendix E.)

**Theorem 4.** *Let $c_1 = \sqrt{B_\star SA}/T_\star$, $c_2 = \sqrt{SA}/T_\star$, $W_c = \lceil (B_\star SA)^{1/3}(K/(\Delta_c T_\star))^{2/3} \rceil$, $W_P = \lceil (SA)^{1/3}(K/(\Delta_P T_\star))^{2/3} \rceil$, and $p = 1/B_\star$. Suppose $s_1^m = s_{\text{init}}$ for all $m \leq K$, then Algorithm 4 ensures $\mathring{R}_K = \tilde{\mathcal{O}}((B_\star SAT_\star(\Delta_c + B_\star^2 \Delta_P))^{1/3} K^{2/3})$ (ignoring lower order terms) with probability at least $1 - 40\delta$.*

The reason that we only analyze this ideal case is that, if the initial state is not $s_{\text{init}}$, then even the optimal policy does not guarantee $T_\star$ hitting time by definition. This also inspires us to eventually deploy a two-phase algorithm slightly modifying Algorithm 1: feed the first interval of each episode into an instance of Algorithm 4, and the rest of intervals into an instance of Algorithm 3 (see Algorithm 5). Thanks to the large terminal cost, we are able to show that the regret in the second phase is upper bounded by a constant, leading to the following final result.

**Theorem 5.** *Algorithm 5 with $\mathfrak{A}_1$ being Algorithm 4 and $\mathfrak{A}_2$ being Algorithm 3 ensures $R_K = \tilde{\mathcal{O}}((B_\star SAT_\star(\Delta_c + B_\star^2 \Delta_P))^{1/3} K^{2/3})$ (ignoring lower order terms) with probability at least $1 - 64\delta$.*

Ignoring logarithmic and lower-order terms, our bound is minimax optimal. Also note that the bound is sub-linear (in $K$) as long as $\Delta_c$ and $\Delta_P$ are sub-linear (that is, not the worst case).

## 7 Learning without Knowing $\Delta_c$ and $\Delta_P$

To handle unknown non-stationarity, we combine our algorithmic ideas in previous sections with a new variant of the MASTER algorithm [Wei and Luo, 2021]. The original MASTER algorithm is a blackbox reduction that takes a base algorithm for (near) stationary environments as input, and turns it into another algorithm for non-stationarity environments. For many problems (including multi-armed bandits, contextual bandits, linear bandits, finite-horizon or infinite-horizon MDPs), Wei and Luo [2021] show that the final algorithm achieves optimal regret without knowing the non-stationarity. While powerful, MASTER can not be directly used in our problem to achieve the same strong result. As we will discuss, some modification is needed, and even with this modification, some extra difficulty unique to SSP still prevents us from eventually obtaining the optimal regret.

Specifically, in order to obtain $T_\star$ dependency, we again follow the two-phase procedure Algorithm 5 and instantiate a MASTER algorithm with a different base algorithm in each phase. In Phase 1, since it is unclear how to update cost and transition estimation independently under the framework of MASTER, we adopt a simpler version of Algorithm 4 as the base algorithm, which performs synchronized cost and transition estimation and a simpler non-stationarity test; see Algorithm 6 (all algorithms/proofs in this section are deferred to Appendix F due to space limit). In Phase 2, we use Algorithm 2 as the base algorithm.

Our version of the MASTER algorithm (Algorithm 8) requires a different **Test 1** compared to that in [Wei and Luo, 2021], which is essential due to the presence of the correction terms in Algorithm 6. Specifically, it no longer makes sense to simply maintain the maximum of estimated value functions over the past intervals, since the cost function combined with the correction term is changing adaptively, and a large correction term will interfere with the detection of a small amount of non-stationarity. Our key observation is that for a base algorithm scheduled on a given range by MASTER, the average of its correction terms within the same range is of the desired order that does not interfere with non-stationarity detection. This inspires us to maintain multiple running averages of the estimated value functions with different scales (see Line 2 of Algorithm 8). Then, to detect a certain level of non-stationarity, we refer to the running average with the matching scale (see Line 3).

We show that the algorithm described above achieves the following regret guarantee without knowledge of the non-stationarity.

**Theorem 6.** *Let $\mathfrak{A}_1$ be an instance of Algorithm 8 with Algorithm 6 as the base algorithm and $\mathfrak{A}_2$ be an instance of Algorithm 8 with Algorithm 2 as the base algorithm. Then Algorithm 5 with $\mathfrak{A}_1$ and $\mathfrak{A}_2$ ensures with high probability (ignoring lower order terms):*

$$R_K = \tilde{\mathcal{O}}\left(\min\left\{B_\star S\sqrt{ALK}, B_\star S\sqrt{AK} + (B_\star^2 S^2 A(\Delta_c + B_\star \Delta_P)T_\star)^{1/3} K^{2/3}\right\}\right),$$

*where $L = 1 + \sum_{k=1}^{K-1} \mathbb{I}\{P_{k+1} \neq P_k \text{ or } c_{k+1} \neq c_k\}$ is the number changes of the environment (plus one). Moreover, this is achieved without the knowledge of $\Delta_c$, $\Delta_P$, or $L$.*

The advantage of this result compared to Theorem 5 is two-fold. First, it adapts to different levels of non-stationarity ($\Delta_c$, $\Delta_P$, and $L$) automatically. Second, it additionally achieves a bound of order $\tilde{\mathcal{O}}(B_\star S\sqrt{ALK})$, which could be much better than that in Theorem 5; for example, when $L = \mathcal{O}(1)$, the former is a $\sqrt{K}$-order bound while the latter is of order $K^{2/3}$. As discussed in [Wei and Luo, 2021], this is a unique benefit brought by the MASTER algorithm and is not achieved by any other algorithms even with the knowledge of $L$.

The disadvantage of Theorem 6, on the other hand, is its sub-optimality in the $B_\star$ dependency for the $\Delta_c$-related term and the $S$ dependency for both terms. The extra $B_\star$ dependency is due to the synchronized cost and transition estimation. As mentioned, it is unclear how to update cost and transition estimation independently as we do in Algorithm 4 under the framework of MASTER, which we leave as an important future direction. On the other hand, the extra $S$ dependency comes from the fact that the lower-order term in the regret bound of the base algorithm affects the final regret bound (see the statement of Theorem 13). Specifically, the lower-order term is $B_\star S^2 A$ instead of $B_\star SA$, which eventually leads to extra $S$ dependency. How to remove the extra $S$ factor in the base algorithm, or eliminate the undesirable lower-order term effect brought by the MASTER algorithm, is another important future direction.

## 8 Conclusion

In this work, we develop the first set of results for dynamic regret minimization in non-stationary SSP, including a (near) minimax optimal algorithm and two others that are either simpler or advantageous in some other cases. Besides the immediate next step such as improving our results when the non-stationarity is unknown, our work also opens up many other possible future directions on this topic, such as extension to more general settings with function approximation. It would also be interesting to study more adaptive dynamic regret bounds in this setting. For example, our $B_\star$ and $T_\star$ are defined as the maximum optimal expected cost and hitting time over all episodes, which is undesirable if only a few episodes admit a large optimal expected cost or hitting time. Ideally, some kind of (weighted) average would be a more reasonable measure in these cases.

## Acknowledgments and Disclosure of Funding

The authors thank Aviv Rosenberg and Chen-Yu Wei for many helpful discussions. HL is supported by NSF Award IIS-1943607 and a Google Research Scholar Award.

## References

Yasin Abbasi-Yadkori, Andras Gyorgy, and Nevena Lazic. A new look at dynamic regret for non-stationary stochastic bandits. *arXiv preprint arXiv:2201.06532*, 2022.

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.

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

Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, 2021.

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 2021a.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Conference on Learning Theory*, pages 1180–1215. PMLR, 2021b.

Liyu Chen, Rahul Jain, and Haipeng Luo. Improved no-regret algorithms for stochastic shortest path with linear MDP. In *International Conference on Machine Learning*, 2022a.

Liyu Chen, Haipeng Luo, and Aviv Rosenberg. Policy optimization for stochastic shortest path. *Conference on Learning Theory*, 2022b.

Wei Chen, Liwei Wang, Haoyu Zhao, and Kai Zheng. Combinatorial semi-bandit in the non-stationary environment. In *Uncertainty in Artificial Intelligence*, pages 865–875. PMLR, 2021c.

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.

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.

Alon Cohen, Haim Kaplan, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret bounds for stochastic shortest path. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8210–8219. PMLR, 2020.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 2021.

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.

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.

Louis Faury, Yoan Russac, Marc Abeille, and Clément Calauzènes. Regret bounds for generalized linear bandits under parameter drift. *arXiv preprint arXiv:2103.05750*, 2021.

Yingjie Fei, Zhuoran Yang, Zhaoran Wang, and Qiaomin Xie. Dynamic regret of policy optimization in non-stationary environments. *Advances in Neural Information Processing Systems*, 33:6743–6754, 2020.

Pratik Gajane, Ronald Ortner, and Peter Auer. A sliding-window algorithm for Markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*, 2018.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. *Advances in Neural Information Processing Systems*, 29, 2016.

Mehdi Jafarnia-Jahromi, Liyu Chen, Rahul Jain, and Haipeng Luo. Online learning for stochastic shortest path model via posterior sampling. *arXiv preprint arXiv:2106.05335*, 2021.

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

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245. PMLR, 2021.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In *International Conference on Machine Learning*, pages 7447–7458. PMLR, 2021.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Learning stochastic shortest path with linear function approximation. *arXiv preprint arXiv:2110.12727*, 2021.

Ronald Ortner, Pratik Gajane, and Peter Auer. Variational regret bounds for reinforcement learning. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 81–90. PMLR, 22–25 Jul 2020.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 2021.

Yoan Russac, Olivier Cappé, and Aurélien Garivier. Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113*, 2020.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8604–8613, 2020.

Joe Suk and Samory Kpotufe. Tracking most severe arm changes in bandits. *arXiv preprint arXiv:2112.13838*, 2021.

Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pages 9428–9437. PMLR, 2020.

Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 2021.

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

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Regret bounds for stochastic shortest path problems with linear function approximation. *arXiv preprint arXiv:2105.01593*, 2021.

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.

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR, 2022.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference On Learning Theory*, 2020.

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

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] After each main theorem.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Pure theoretical work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

## Contents of Appendix



# A Preliminaries

**Extra Notations** We first define (or restate) some notations used throughout the whole Appendix.

- Let $\Delta_{c,[i,j]} = \sum_{\tau=i}^{j-1} \left\|c^{\tau+1} - c^{\tau}\right\|_\infty$, $\Delta_{P,[i,j]} = \sum_{\tau=i}^{j-1} \max_{s,a} \left\|P^{\tau+1}_{s,a} - P^{\tau}_{s,a}\right\|_1$. It is straightforward to verify that $\Delta_{c,[1,M]} = \Delta_c$ and $\Delta_{P,[1,M]} = \Delta_P$.
- Define $\Delta_{c,m} = \Delta_{c,[i^c_m,m]}$ and $\Delta_{P,m} = \Delta_{P,[i^P_m,m]}$, where $i^c_m$ and $i^P_m$ are the first intervals after the last resets of $\mathbf{M}$ and $\mathbf{N}$ before interval $m$ respectively.
- For all algorithms, denote by $\widehat{c}^m$, $\bar{c}^m$, $\bar{P}^m_{s,a}$, $b^m_h$, $\mathbf{N}^+_m$, $\mathbf{M}^+_m$, $\iota_m$ the value of $\widehat{c}$, $\bar{c}$, $\bar{P}_{s,a}$, $b_h$, $\mathbf{N}^+$, $\mathbf{M}^+$, $\iota$ at the beginning of interval $m$, and define $\widehat{c}^m_h = \widehat{c}^m(s^m_h, a^m_h)$, $\bar{c}^m_h = \bar{c}(s^m_h, a^m_h)$, $\mathbf{N}^m_h = \mathbf{N}^+(s^m_h, a^m_h)$, and $\mathbf{M}^m_h = \mathbf{M}^+(s^m_h, a^m_h)$. We also slightly abuse the notation and write $b^m(s^m_h, a^m_h)$ as $b^m_h$ when there is no confusion.

- Define $\widetilde{c}^m(s,a) = \frac{1}{\mathbf{M}_m^+(s,a)}\sum_{m'=i_m^c}^{m-1}\sum_{h=1}^{H_{m'}} c^{m'}(s,a)\mathbb{I}\{(s_h^{m'},a_h^{m'})=(s,a)\}$, $\widetilde{c}_h^m = \widetilde{c}^m(s_h^m,a_h^m)$, $\widetilde{P}_{s,a}^m = \frac{1}{\mathbf{N}_m^+(s,a)}\sum_{m'=i_m^P}^{m-1}\sum_{h=1}^{H_{m'}} P_{s,a}^{m'}\mathbb{I}\{(s_h^{m'},a_h^{m'})=(s,a)\}$, $\bar{P}_h^m = \bar{P}_{s_h^m,a_h^m}^m$, and $\widetilde{P}_h^m = \widetilde{P}_{s_h^m,a_h^m}^m$.

- Denote by $L_{c,[i,j]}$ and $L_{P,[i,j]}$ one plus the number of resets of $\mathbf{M}$ and $\mathbf{N}$ within intervals $[i,j]$ respectively, and define $L_{c,m}=L_{c,[1,m]}$, $L_{P,m}=L_{P,[1,m]}$, $L_m=L_{c,m}+L_{P,m}$ for any $m\geq 1$.

- Define $f^c(m)$ (or $f^P(m)$) as the earliest interval at or after interval $m$ in which the learner resets $\mathbf{M}$ (or $\mathbf{N}$).

- Define $\mathbf{m}_h^m = \mathbb{I}\{\mathbf{M}^m(s_h^m,a_h^m)=0\}$, $\mathbf{n}_h^m = \mathbb{I}\{\mathbf{N}^m(s_h^m,a_h^m)=0\}$, $C_{M'} = \sum_{m=1}^{M'}\sum_{h=1}^{H_m+1} c_h^m$, and bonus function $b^m(s,a,V)=\max\left\{7\sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}^m,V)\iota_m}{\mathbf{N}_m^+(s,a)}}, \frac{49B\sqrt{S}\iota_m}{\mathbf{N}_m^+(s,a)}\right\}$.

- Define $T_h^{\pi^\star,m}(s)$ (or $T_h^{\pi^\star,m}(s,a)$) as the hitting time (reaching $g$ or layer $H+1$) of $\pi_{k(m)}^\star$ starting from state $s$ (or state-action pair $(s,a)$) in layer $h$ w.r.t transition $P^m$, such that $T_h^{\pi^\star,m}(s,a)=1+P_{s,a}^m T_{h+1}^{\pi^\star,m}$, $T_h^{\pi^\star,m}(s)=T_h^{\pi^\star,m}(s,\pi_{k(m)}^\star(s))$, and $T_{H+1}^{\pi^\star,m}(s) = T_{H+1}^{\pi^\star,m}(s,a) = T_h^{\pi^\star,m}(g) = T_h^{\pi^\star,m}(g,a)=0$.

- For notational convenience, we often write $V_h^{\pi_{k(m)}^\star,m}$ as $V_h^{\pi^\star,m}$.

- Define $(x)_+ = \max\{0,x\}$.

**Optimal Value Functions of $\mathring{\mathcal{M}}$** We denote by $Q_h^{\star,m}$ and $V_h^{\star,m}$ the optimal value functions in interval $m$. It is not hard to see that they can be defined recursively as follows: $V_{H+1}^{\star,m}=c_f$ and for $h\leq H$,

$$Q_h^{\star,m}(s,a) = c^m(s,a) + P_{s,a}^m V_{h+1}^{\star,m}, \qquad V_h^{\star,m}(s) = \min_a Q_h^{\star,m}(s,a).$$

For notational convenience, we also let $Q_{H+1}^{\star,m}(s,a)=V_{H+1}^{\star,m}(s)$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$.

**Lemma 2.** *For any $m\geq 1$ and $h\leq H+1$, $Q_h^{\star,m}(s,a)\leq Q_h^{\pi^\star,m}(s,a)\leq 4B_\star$.*

*Proof.* This is simply by $Q_h^{\pi^\star,m}(s,a)\leq 1+\max_s V_k^\star(s)+2B_\star\leq 4B_\star$. □

**Auxiliary Lemmas** Below we provide auxiliary lemmas used throughout the whole Appendix and for all algorithms.

**Lemma 3.** *With probability at least $1-3\delta$, $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c^m(s_h^m,a_h^m)-\widetilde{c}_h^m) \leq 3\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\widetilde{c}_h^m\iota_m}{\mathbf{M}_h^m}}+\frac{\iota_m}{\mathbf{M}_h^m}\right)+\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta_{c,m} \leq \tilde{\mathcal{O}}\left(\sqrt{SAL_{c,M'}C_{M'}}+SAL_{c,M'}\right) + 2\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta_{c,m}$ and $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\widetilde{c}_h^m\iota_m}{\mathbf{M}_h^m}}+\frac{\iota_m}{\mathbf{M}_h^m}\right)\leq\tilde{\mathcal{O}}(\sqrt{SAL_{c,M'}C_{M'}}+SAL_{c,M'}+\sqrt{SAL_{c,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta_{c,m}})$ for any $M'\leq M$.*

*Proof.* First note that by Lemma 49, with probability at least $1-\delta$, for any $m\geq 1$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\widetilde{c}^m(s,a)-\bar{c}^m(s,a)\leq\sqrt{\frac{\bar{c}^m(s_h^m,a_h^m)}{\mathbf{M}_m^+(s,a)}}+\frac{1}{\mathbf{M}_m^+(s,a)}. \tag{1}$$

For the first inequality in the first statement, note that

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c^m(s_h^m,a_h^m)-\widehat{c}_h^m)\\
&\leq \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\widetilde{c}^m(s_h^m,a_h^m)-\bar{c}^m(s_h^m,a_h^m)+\sqrt{\frac{\bar{c}_h^m\iota_m}{\mathbf{M}_h^m}}+\frac{\iota_m}{\mathbf{M}_h^m}+\mathbf{m}_h^m\right)+\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta_{c,m}\\
&\qquad\qquad\qquad\text{(definition of } \widehat{c}_h^m \text{ and } c^m(s_h^m,a_h^m)\leq\widetilde{c}^m(s_h^m,a_h^m)+\Delta_{c,m}+\mathbf{m}_h^m)\\
&\leq 3\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\bar{c}_h^m\iota_m}{\mathbf{M}_h^m}}+\frac{\iota_m}{\mathbf{M}_h^m}\right)+\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta_{c,m}. \qquad \text{(Eq. (1) and } \mathbf{m}_h^m\leq\tfrac{1}{\mathbf{M}_h^m})
\end{aligned}$$

The second inequality in the first statement simply follows from applying AM-GM inequality on the second statement. To prove the second statement, first note that by Lemma 49, Cauchy-Schwarz inequality, and Lemma 11, with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\bar{c}_h^m &= \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\widetilde{c}_h^m+\sqrt{\frac{\widetilde{c}_h^m}{\mathbf{M}_h^m}}+\frac{1}{\mathbf{M}_h^m}\right)\right)\\
&= \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\widetilde{c}_h^m+\sqrt{SAL_{c,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\widetilde{c}_h^m}+SAL_{c,M'}\right).
\end{aligned}$$

Solving a quadratic inequality w.r.t $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\bar{c}_h^m$ (Lemma 45) gives $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\bar{c}_h^m = \tilde{\mathcal{O}}(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\widetilde{c}_h^m+SAL_{c,M'})$. Therefore, with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\bar{c}_h^m\iota_m}{\mathbf{M}_h^m}}+\frac{\iota_m}{\mathbf{M}_h^m}\right)=\tilde{\mathcal{O}}\left(\sqrt{SAL_{c,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\bar{c}_h^m}+SAL_{c,M'}\right)\\
&\qquad\qquad\qquad\text{(Cauchy-Schwarz inequality and Lemma 11)}\\
&=\tilde{\mathcal{O}}\left(\sqrt{SAL_{c,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\widetilde{c}_h^m}+SAL_{c,M'}\right)\\
&=\tilde{\mathcal{O}}\left(\sqrt{SAL_{c,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta_{c,m}}+\sqrt{SAL_{c,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}c^m(s_h^m,a_h^m)}+SAL_{c,M'}\right)\\
&=\tilde{\mathcal{O}}\left(\sqrt{SAL_{c,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta_{c,m}}+\sqrt{SAL_{c,M'}C_{M'}}+SAL_{c,M'}\right). \qquad \text{(Lemma 50)}
\end{aligned}$$

This completes the proof. □

**Lemma 4.** *With probability at least* $1-\delta$*, for any* $m\geq 1$*,* $(s,a)\in\mathcal{S}\times\mathcal{A}$ *and* $s'\in\mathcal{S}_+$*,* $\left|\widetilde{P}_{s,a}^m(s')-\bar{P}_{s,a}^m(s')\right|\leq\sqrt{\frac{\widetilde{P}_{s,a}^m(s')\iota_m}{2\mathbf{N}_m^+(s,a)}}+\frac{\iota_m}{2\mathbf{N}_m^+(s,a)}\leq\sqrt{\frac{\bar{P}_{s,a}^m(s')\iota_m}{\mathbf{N}_m^+(s,a)}}+\frac{\iota_m}{\mathbf{N}_m^+(s,a)}$.

*Proof.* The first inequality hold with probability at least $1-\delta/2$ by applying Lemma 49 for each $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $s'\in\mathcal{S}_+$. Also by Lemma 50, we have $\widetilde{P}_{s,a}^m(s')\leq 2\bar{P}_{s,a}^m(s')+\frac{\iota_m}{2\mathbf{N}_m^+(s,a)}$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, $s'\in\mathcal{S}_+$ with probability at least $1-\delta/2$. Substituting this back and applying $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ proves the second inequality. □

**Lemma 5.** *With probability at least* $1-\delta$*, for any* $(s,a)\in\mathcal{S}\times\mathcal{A}$ *and* $m\geq 1$*,* $\widehat{c}^m(s,a)\leq c^m(s,a)+\Delta_{c,m}$.

*Proof.* For any $(s,a)$ and $m \ge 1$, when $\mathbf{M}_m(s,a) = 0$, the statement clearly holds since $\bar{c}^m(s,a) = 0$. Otherwise, by Lemma 49 and Lemma 50, with probability at least $1-\delta$, for all $(s,a)$ and $m \ge 1$ simultaneously,

$$
\begin{aligned}
|\bar{c}^m(s,a) - \widetilde{c}^m(s,a)| &\le 3\sqrt{\frac{\widetilde{c}^m(s,a)}{\mathbf{M}_m^+(s,a)}\ln\frac{32SAm^5}{\delta}} + \frac{2\ln\frac{32SAm^5}{\delta}}{\mathbf{M}_m^+(s,a)} \\
&\le 3\sqrt{\frac{\left(2\bar{c}^m(s,a) + \frac{12\ln\frac{4SAm}{\delta}}{\mathbf{M}_m^+(s,a)}\right)}{\mathbf{M}_m^+(s,a)}\ln\frac{32SAm^5}{\delta}} + \frac{2\ln\frac{32SAm^5}{\delta}}{\mathbf{M}_m^+(s,a)} \le \sqrt{\frac{\bar{c}^m(s,a)\iota_m}{\mathbf{M}_m^+(s,a)}} + \frac{\iota_m}{\mathbf{M}_m^+(s,a)}.
\end{aligned} \tag{2}
$$

Therefore, by $\max\{0,a\} - \max\{0,b\} \le \max\{0, a-b\}$,

$$
\begin{aligned}
&\widehat{c}^m(s,a) - c^m(s,a) \le \widehat{c}^m(s,a) - \widetilde{c}^m(s,a) + \Delta_{c,m} \\
&\le \max\left\{0, \bar{c}^m(s,a) - \widetilde{c}^m(s,a) - \sqrt{\frac{\bar{c}^m(s,a)\iota_m}{\mathbf{M}_m^+(s,a)}} - \frac{\iota_m}{\mathbf{M}_m^+(s,a)}\right\} + \Delta_{c,m} \le \Delta_{c,m},
\end{aligned}
$$

where the last step is by Eq. (2). $\square$

**Lemma 6.** *Given function $V \in [-B,B]^{\mathcal{S}_+}$ for some $B > 0$, we have with probability at least $1-\delta$, $|(\widetilde{P}^m_{s,a} - \bar{P}^m_{s,a})V| \le \tilde{\mathcal{O}}\left(\sqrt{\frac{S\mathbb{V}(P^m_{s,a},V)}{\mathbf{N}_m^+(s,a)}} + \frac{SB}{\mathbf{N}_m^+(s,a)}\right) + \frac{B\Delta_{P,m}}{64}$ for any $m \ge 1$.*

*Proof.* Note that with probability at least $1-\delta$,

$$
\begin{aligned}
|(\widetilde{P}^m_{s,a} - \bar{P}^m_{s,a})V| &= |(\widetilde{P}^m_{s,a} - \bar{P}^m_{s,a})(V - P^m_{s,a}V)| \\
&= \tilde{\mathcal{O}}\left(\sum_{s'}\left(\sqrt{\frac{\widetilde{P}^m_{s,a}(s')}{\mathbf{N}_m^+(s,a)}}|V(s') - P^m_{s,a}V| + \frac{B}{\mathbf{N}_m^+(s,a)}\right)\right) && \text{(Lemma 4)} \\
&= \tilde{\mathcal{O}}\left(\sqrt{\frac{S\widetilde{P}^m_h(V - P^m_{s,a}V)^2}{\mathbf{N}_m^+(s,a)}} + \frac{SB}{\mathbf{N}_m^+(s,a)}\right) && \text{(Cauchy-Schwarz inequality)} \\
&= \tilde{\mathcal{O}}\left(\sqrt{\frac{SP^m_h(V - P^m_{s,a}V)^2}{\mathbf{N}_m^+(s,a)}} + \frac{SB}{\mathbf{N}_m^+(s,a)} + B\sqrt{\frac{S\Delta_{P,m}}{\mathbf{N}_m^+(s,a)}}\right).
\end{aligned}
$$

Applying AM-GM inequality completes the proof. $\square$

**Lemma 7.** *With probability at least $1-\delta$, $\mathbb{V}(\bar{P}^m_h, V^m_{h+1}) \le 2\mathbb{V}(P^m_h, V^m_{h+1}) + \tilde{\mathcal{O}}\left(\frac{SB^2}{\mathbf{N}^m_h}\right) + 2B^2\Delta_{P,m}$ for any $m \ge 1$.*

*Proof.* Note that:

$$
\begin{aligned}
\mathbb{V}(\bar{P}^m_h, V^m_{h+1}) &\le \bar{P}^m_h(V^m_{h+1} - P^m_hV^m_{h+1})^2 && \left(\tfrac{\sum_i p_ix_i}{\sum_i p_i} = \operatorname{argmin}_z \textstyle\sum_i p_i(x_i - z)^2\right) \\
&= \mathbb{V}(P^m_h, V^m_{h+1}) + (\bar{P}^m_h - P^m_h)(V^m_{h+1} - P^m_hV^m_{h+1})^2 \\
&\le \mathbb{V}(P^m_h, V^m_{h+1}) + (\bar{P}^m_h - \widetilde{P}^m_h)(V^m_{h+1} - P^m_hV^m_{h+1})^2 + B^2\Delta_{P,m} \\
&\le \mathbb{V}(P^m_h, V^m_{h+1}) + \tilde{\mathcal{O}}\left(B\sqrt{\frac{S\widetilde{P}^m_h(V^m_{h+1} - P^m_hV^m_{h+1})^2}{\mathbf{N}^m_h}} + \frac{SB^2}{\mathbf{N}^m_h}\right) + B^2\Delta_{P,m} \\
& && \text{(Lemma 4 and Cauchy-Schwarz inequality)} \\
&\le \mathbb{V}(P^m_h, V^m_{h+1}) + \tilde{\mathcal{O}}\left(B\sqrt{\frac{S\mathbb{V}(P^m_h, V^m_{h+1})}{\mathbf{N}^m_h}} + B^2\sqrt{\frac{S\Delta_{P,m}}{\mathbf{N}^m_h}} + \frac{SB^2}{\mathbf{N}^m_h}\right) + B^2\Delta_{P,m} \\
&\le 2\mathbb{V}(P^m_h, V^m_{h+1}) + \tilde{\mathcal{O}}\left(\frac{SB^2}{\mathbf{N}^m_h}\right) + 2B^2\Delta_{P,m}. && \text{(AM-GM inequality)}
\end{aligned}
$$

$\square$

**Lemma 8.** *Given an oblivious set of value functions $\mathcal{V}$ with $|\mathcal{V}| \leq (2HK)^6$ and $\|V\|_\infty \leq B$ for any $V \in \mathcal{V}$, we have with probability at least $1-\delta$, for any $V \in \mathcal{V}$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, and $m \geq 1$, $|(\bar{P}^m_{s,a} - \widetilde{P}^m_{s,a})V| \leq \sqrt{\frac{\mathbb{V}(P^m_{s,a},V)\iota_m}{\mathbf{N}^+_m(s,a)}} + \frac{17B\iota_m}{\mathbf{N}^+_m(s,a)} + \frac{B\Delta_{P,m}}{64}$ and $|(\bar{P}^m_{s,a} - \widetilde{P}^m_{s,a})V| \leq \sqrt{\frac{2\mathbb{V}(\bar{P}^m_{s,a},V)\iota_m}{\mathbf{N}^+_m(s,a)}} + \frac{3B\sqrt{S}\iota_m}{\mathbf{N}^+_m(s,a)}$.*

*Proof.* For each $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $V\in\mathcal{V}$, by Lemma 49, with probability at least $1-\frac{\delta}{2SA(2HK)^6}$, for any $m\geq 1$

$$|(\bar{P}^m_{s,a} - \widetilde{P}^m_{s,a})V| \leq \frac{1}{\mathbf{N}^+_m(s,a)}\left(\sqrt{\sum_{i=1}^{\mathbf{N}_m(s,a)}\mathbb{V}(P^{m_i}_{s,a},V)\iota_m} + B\iota_m\right). \tag{3}$$

Denote by $m_i$ the interval where the $i$-th visits to $(s,a)$ lies in among those $\mathbf{N}_m(s,a)$ visits, we have

$$\begin{aligned}
\frac{1}{\mathbf{N}^+_m(s,a)}\sum_{i=1}^{\mathbf{N}_m(s,a)}\mathbb{V}(P^{m_i}_{s,a},V) &= \frac{1}{\mathbf{N}^+_m(s,a)}\sum_{i=1}^{\mathbf{N}_m(s,a)}P^{m_i}_{s,a}(V-P^{m_i}_{s,a}V)^2\\
&\leq \frac{1}{\mathbf{N}^+_m(s,a)}\sum_{i=1}^{\mathbf{N}_m(s,a)}P^{m_i}_{s,a}(V-P^{m}_{s,a}V)^2 \leq \mathbb{V}(P^m_{s,a},V) + B^2\Delta_{P,m},
\end{aligned}$$

where the second last inequality is by $\frac{\sum_i p_ix_i}{\sum_i p_i} = \mathrm{argmin}_z\sum_i p_i(x_i-z)^2$. Thus by Eq. (3),

$$\begin{aligned}
|(\bar{P}^m_{s,a} - \widetilde{P}^m_{s,a})V| &\leq \sqrt{\frac{\mathbb{V}(P^m_{s,a},V)\iota_m}{\mathbf{N}^+_m(s,a)}} + \frac{B\iota_m}{\mathbf{N}^+_m(s,a)} + B\sqrt{\frac{\Delta_{P,m}\iota_m}{\mathbf{N}^+_m(s,a)}}\\
&\leq \sqrt{\frac{\mathbb{V}(P^m_{s,a},V)\iota_m}{\mathbf{N}^+_m(s,a)}} + \frac{17B\iota_m}{\mathbf{N}^+_m(s,a)} + \frac{B\Delta_{P,m}}{64}. \qquad \text{(AM-GM inequality)}
\end{aligned}$$

Moreover, again by $\frac{\sum_i p_ix_i}{\sum_i p_i} = \mathrm{argmin}_z\sum_i p_i(x_i-z)^2$,

$$\begin{aligned}
&\frac{1}{\mathbf{N}^+_m(s,a)}\sum_{i=1}^{\mathbf{N}_m(s,a)}\mathbb{V}(P^{m_i}_{s,a},V) \leq \frac{1}{\mathbf{N}^+_m(s,a)}\sum_{i=1}^{\mathbf{N}_m(s,a)}P^{m_i}_{s,a}(V-\bar{P}^m_{s,a}V)^2\\
&\leq \mathbb{V}(\bar{P}^m_{s,a},V) + (\widetilde{P}^m_{s,a}-\bar{P}^m_{s,a})(V-\bar{P}^m_{s,a}V)^2 \leq \mathbb{V}(\bar{P}^m_{s,a},V) + B\sqrt{\frac{S\mathbb{V}(\bar{P}^m_{s,a},V)\iota_m}{\mathbf{N}^+_m(s,a)}} + \frac{SB^2\iota_m}{\mathbf{N}^+_m(s,a)}\\
&\qquad\qquad \text{(Lemma 4 and Cauchy-Schwarz inequality)}\\
&\leq 2\mathbb{V}(\bar{P}^m_{s,a},V) + \frac{2SB^2\iota_m}{\mathbf{N}^+_m(s,a)}. \qquad \text{(AM-GM inequality)}
\end{aligned}$$

Thus by Eq. (3), $|(\bar{P}^m_{s,a} - \widetilde{P}^m_{s,a})V| \leq \sqrt{\frac{2\mathbb{V}(\bar{P}^m_{s,a},V)\iota_m}{\mathbf{N}^+_m(s,a)}} + \frac{3B\sqrt{S}\iota_m}{\mathbf{N}^+_m(s,a)}$. □

**Lemma 9.** *For any sequence of value functions $\{V^m_h\}_{m,h}$ with $\|V^m_h\|_\infty \in [0,B]$, we have with probability at least $1-\delta$, for all $M'\geq 1$, $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P^m_h,V^m_{h+1}) = \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}V^m_{H_m+1}(s^m_{H_m+1})^2 + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}B(V^m_h(s^m_h)-P^m_hV^m_{h+1})_+ + B^2\right)$.*

*Proof.* We decompose the sum of variance as follows:

$$\begin{aligned}
\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P^m_h,V^m_{h+1}) &= \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(P^m_h(V^m_{h+1})^2 - V^m_{h+1}(s^m_{h+1})^2\right)\\
&+ \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(V^m_{h+1}(s^m_{h+1})^2 - V^m_h(s^m_h)^2\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(V^m_h(s^m_h)^2 - (P^m_hV^m_{h+1})^2\right).
\end{aligned}$$

For the first term, by Lemma 49 and Lemma 47, with probability at least $1-\delta$,

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(P_h^m(V_{h+1}^m)^2 - V_{h+1}^m(s_{h+1}^m)^2\right) = \tilde{\mathcal{O}}\left(\sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,(V_{h+1}^m)^2)} + B^2\right)$$
$$= \tilde{\mathcal{O}}\left(B\sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)} + B^2\right).$$

The second term is clearly upper bounded by $\sum_{m=1}^{M'} V_{H_m+1}^m(s_{H_m+1}^m)^2$, and the third term is upper bounded by $2B\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(V_h^m(s_h^m) - P_h^m V_{h+1}^m)_+$ by $a^2-b^2 \leq (a+b)(a-b)_+$. Putting everything together and solving a quadratic inequality (Lemma 45) w.r.t $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^m)$ completes the proof. $\square$

**Lemma 10.** *For any value functions $\{V_h^m\}_{m,h}$ such that $\|V_h^m\|_\infty \leq B$, with probability at least $1-\delta$, for any $M' \geq 1$,*

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m} b^m(s_h^m, a_h^m, V_{h+1}^m)$$
$$= \tilde{\mathcal{O}}\left(\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)} + BS^{1.5}AL_{P,M'} + B\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta_{P,m}}\right).$$

*Proof.* Note that:

$$\sum_{m=1}^{M'}\sum_{h=1}^{H} b^m(s_h^m, a_h^m, V_{h+1}^m) = \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\mathbb{V}(\bar{P}_h^m, V_{h+1}^m)}{\mathbf{N}_h^m}} + \frac{B\sqrt{S}}{\mathbf{N}_h^m}\right)\right)$$
$$= \tilde{\mathcal{O}}\left(\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(\bar{P}_h^m,V_{h+1}^m)} + BS^{1.5}AL_{P,M'}\right)$$

(Cauchy-Schwarz inequality and Lemma 11)

$$= \tilde{\mathcal{O}}\left(\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)} + BS^{1.5}AL_{P,M'} + B\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta_{P,m}}\right).$$

(Lemma 7, Lemma 11, and $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$)

$\square$

**Lemma 11.** *For any $M' \geq 1$, $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{1}{\mathbf{M}_h^m} = \tilde{\mathcal{O}}(SAL_{c,M'})$ and $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{1}{\mathbf{N}_h^m} = \tilde{\mathcal{O}}(SAL_{P,M'})$.*

*Proof.* This simply follows from the fact that the sum of $\frac{1}{\mathbf{M}_h^m}$ (or $\frac{1}{\mathbf{N}_h^m}$) between consecutive resets of $\mathbf{M}_h^m$ (or $\mathbf{N}_h^m$) is of order $\tilde{\mathcal{O}}(SA)$. $\square$

**Lemma 12.** $\sum_{m=1}^{M'}\mathbb{I}\{H_m < H, s_{H_m+1}^m \neq g\} = \tilde{\mathcal{O}}(SAL_{M'})$ *for any $M' \leq M$.*

*Proof.* This simply follows from the fact that between consecutive resets of $\mathbf{M}$ or $\mathbf{N}$, the number of times that the number of visits to some $(s,a)$ is doubled is $\tilde{\mathcal{O}}(SA)$. $\square$

**Lemma 13.** *Suppose $r(m) = \min\{\frac{c_1}{\sqrt{m}} + c_2, c_3\}$, $\Delta \in \mathbb{R}_+^{\mathbb{N}_+}$ is a non-stationarity measure, and define $\Delta_{[i,j]} = \sum_{i=1}^{j-1}\Delta(i)$. If for a given interval $\mathcal{J}$, there is a way to partition $\mathcal{J}$ into $\ell$ intervals $\{\mathcal{I}_i\}_{i=1}^{\ell}$ with $\mathcal{I}_i = [s_i, e_i]$ such that $\Delta_{[s_i,e_i+1]} > r(|\mathcal{I}_i|+1)$ for $i \leq \ell-1$ (note that $|\mathcal{I}_i| = e_i - s_i + 1$), then $\ell \leq 1 + (2c_1^{-1}\Delta_{\mathcal{J}})^{2/3}|\mathcal{J}|^{1/3} + c_3^{-1}\Delta_{\mathcal{J}}$.*

*Proof.* Note that

$$\Delta_{\mathcal{J}} \geq \sum_{i=1}^{\ell-1} \Delta_{[s_i,e_i+1]} > \sum_{i=1}^{\ell-1} r(|\mathcal{I}_i|+1) \geq \sum_{i=1}^{\ell-1} \min\left\{c_1(|\mathcal{I}_i|+1)^{-1/2}, c_3\right\}$$
$$\geq \sum_{i=1}^{\ell-1} \min\left\{\frac{c_1}{2}|\mathcal{I}_i|^{-1/2}, c_3\right\} = \sum_{i=1}^{\ell_1} \frac{c_1}{2}|\mathcal{I}_i|^{-1/2} + \ell_2 c_3,$$

where in the last step we assume $|\mathcal{I}_i|$ is decreasing in $i$ without loss of generality and $\ell_1 + \ell_2 = \ell - 1$. The inequality above implies $\ell_2 \leq c_3^{-1}\Delta_{\mathcal{J}}$ and

$$\ell_1 = \sum_{i=1}^{\ell_1} |\mathcal{I}_i|^{-\frac{1}{3}}|\mathcal{I}_i|^{\frac{1}{3}} \leq \left(\sum_{i=1}^{\ell_1}|\mathcal{I}_i|^{-1/2}\right)^{\frac{2}{3}}\left(\sum_{i=1}^{\ell_1}|\mathcal{I}_i|\right)^{\frac{1}{3}} \leq \left(\frac{2\Delta_{\mathcal{J}}}{c_1}\right)^{\frac{2}{3}}|\mathcal{J}|^{\frac{1}{3}}$$

(Hölder's inequality with $p = \frac{3}{2}$ and $q = 3$)

Combining them completes the proof. □

# B Omitted Details in Section 3

In this section we provide omitted proofs and discussions in Section 3.

## B.1 Optimal Value Change w.r.t Non-stationarity

Below we provide a bound on the change of optimal value functions w.r.t cost and transition non-stationarity.

**Lemma 14.** *For any $k_1, k_2 \in [K]$, $V^\star_{k_1}(s_{\text{init}}) - V^\star_{k_2}(s_{\text{init}}) \leq (\Delta_c + B_\star\Delta_P)T_\star$.*

*Proof.* Denote by $q^\star_{k_2}(s,a)$ (or $q^\star_{k_2}(s)$) the number of visits to $(s,a)$ (or $s$) before reaching $g$ following $\pi^\star_{k_2}$. By the extended value difference lemma [Shani et al., 2020, Lemma 1] (note that their result is for finite-horizon MDP, but the nature generalization to SSP holds), we have

$$\begin{aligned}
&V^\star_{k_1}(s_{\text{init}}) - V^\star_{k_2}(s_{\text{init}}) \\
&= \sum_s q^\star_{k_2}(s)(V^\star_{k_1}(s) - Q^\star_{k_1}(s,\pi^\star_{k_2}(s))) + \sum_{s,a} q^\star_{k_2}(s,a)(Q^\star_{k_1}(s,a) - c_{k_2}(s,a) - P_{k_2,s,a}V^\star_{k_1}) \\
&\leq \sum_{s,a} q^\star_{k_2}(s,a)(c_{k_1}(s,a) - c_{k_2}(s,a) + (P_{k_1,s,a} - P_{k_2,s,a})V^\star_{k_1}) \leq (\Delta_c + B_\star\Delta_P)T_\star.
\end{aligned}$$

where in the last inequality we apply $\|c_{k_1} - c_{k_2}\|_\infty \leq \Delta_c$,$(P_{k_1,s,a} - P_{k_2,s,a})V^\star_{k_1} \leq \max_{s,a}\|P_{k_1,s,a} - P_{k_2,s,a}\|_1 \left\|V^\star_{k_1}\right\|_\infty \leq B_\star\Delta_P$, and $\sum_{s,a} q^\star_{k_2}(s,a) \leq T_\star$. □

We also give an example showing that the bound in Lemma 14 is tight up to a multiplication factor. Consider an SSP instance with only one state $s_{\text{init}}$ and one action $a_g$, such that $c(s_{\text{init}}, a_g) = \frac{B_\star}{T_\star}$, $P(g|s_{\text{init}}, a_g) = \frac{1}{T_\star}$, and $P(s_{\text{init}}|s_{\text{init}}, a_g) = 1 - P(g|s_{\text{init}}, a_g)$ with $1 \leq B_\star \leq T_\star$. The optimal value of this instance is clearly $B_\star$. Now consider another SSP instance with perturbed cost function $c'(s_{\text{init}}, a_g) = \frac{B_\star}{T_\star} + \Delta_c$ and perturbed transition function $P'(g|s_{\text{init}}, a_g) = \frac{1}{T_\star} - \frac{\Delta_P}{2}$, $P'(s_{\text{init}}|s_{\text{init}}, a_g) = 1 - P'(g|s_{\text{init}}, a_g)$ with $\max\{\Delta_c, \Delta_P\} \leq \frac{1}{T_\star}$. The optimal value function in this instance is

$$\begin{aligned}
\frac{\frac{B_\star}{T_\star} + \Delta_c}{\frac{1}{T_\star} - \frac{\Delta_P}{2}} &= \frac{B_\star + T_\star\Delta_c}{1 - \frac{T_\star\Delta_P}{2}} \leq (B_\star + T_\star\Delta_c)(1 + T_\star\Delta_P) = B_\star + (\Delta_c + B_\star\Delta_P)T_\star + T_\star^2\Delta_c\Delta_P \\
&\leq B_\star + 2(\Delta_c + B_\star\Delta_P)T_\star,
\end{aligned}$$

where in the first inequality we apply $\frac{1}{1-x} \leq 1 + 2x$ for $x \in [0, \frac{1}{2}]$. Thus the optimal value difference between these two SSPs is of the same order of the upper bound in Lemma 14.

### B.2 Proof of Theorem 1

For the ease of analysis, in this section we consider SSP instances with different action set at different state similar to [Chen et al., 2021b]. The meaning of $SA$ is still the total number of state-action pairs in the SSP instance.

For any $B_\star, T_\star, SA, K$ with $B_\star \geq 1$, $T_\star \geq 3B_\star$, and $K \geq SA \geq 10$, we define a set of SSP instances $\{\mathcal{M}_{i,j}^K\}_{i,j}$ with $i, j \in \{0, 1, \ldots, N\}$ and $N = SA$. The instance $\mathcal{M}_{i^\star,j^\star}^K$ is constructed as follows:

- There are $N + 1$ states $\{s_{\text{init}}, s_1, \ldots, s_N\}$.
- At $s_{\text{init}}$, there are $N$ actions $a_1, \ldots, a_N$; at $s_i$ for $i \in [N]$ there is only one action $a_g$.
- $c(s_{\text{init}}, a_i) = 0$ and $c(s_i, a_g) \sim$ Bernoulli$(\frac{B_\star + \epsilon_{c,K}\mathbb{I}\{i \neq i^\star\}}{T_\star})$ for $i \in [N]$, where $\epsilon_{c,K} = \frac{1-1/N}{4}\sqrt{NB_\star/K}$.
- $P(s_i|s_{\text{init}}, a_i) = 1$, $P(g|s_j, a_g) = \frac{1+\epsilon_{P,K}\mathbb{I}\{j=j^\star\}}{T_\star}$, and $P(s_j|s_j, a_g) = 1 - P(g|s_j, a_g)$, where $\epsilon_{P,K} = \frac{1-1/N}{4}\sqrt{N/K}$.

Note that for any $\mathcal{M}_{i,j}^K$, the expected hitting time is upper bounded by $T_\star + 1$, the expected cost of optimal policy is upper bounded by $2B_\star$, and the number of state-action pairs is upper bounded by $2N$. We then use $\{\mathcal{M}_{i,j}^K\}_{i,j}$ to prove static regret lower bounds (note that static regret and dynamic regret are the same without non-stationarity, that is, $\Delta_c = \Delta_P = 0$) based on cost perturbation and transition perturbation respectively, which serve as the cornerstones of the proof of Theorem 1.

**Theorem 7.** *For any $B_\star, T_\star, SA, K$ with $B_\star \geq 1$, $T_\star \geq 3B_\star$, $K \geq SA \geq 10$, and any learner, there exists an SSP instance based on cost perturbation such that the regret of the learner after $K$ episodes is at least $\Omega(\sqrt{B_\star SAK})$.*

*Proof.* Consider a distribution of SSP instances which is uniform over $\{\mathcal{M}_{i,0}^K\}_i$ for $i \in [N]$. Let $\mathbb{E}_i$ be the expectation w.r.t $\mathcal{M}_{i,0}^K$, $P_i$ be the distribution of learner's observations w.r.t $\mathcal{M}_{i,0}^K$, and $K_i$ the number of visits to state $i$ in $K$ episodes. Also let $\epsilon_c = \epsilon_{c,K}$. The expected regret over this distribution of SSPs can be lower bounded as

$$\mathbb{E}[R_K] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_i[R_K] \geq \frac{1}{N}\sum_{i=1}^N \mathbb{E}_i[K - K_i]\epsilon_c = \epsilon_c\left(K - \frac{1}{N}\sum_{i=1}^N \mathbb{E}_i[K_i]\right).$$

Note that $\mathcal{M}_{0,0}^K$ has no "good" state. By Pinsker's inequality:

$$\mathbb{E}_i[K_i] - \mathbb{E}_0[K_i] \leq K\left\|P_i - P_0\right\|_1 \leq K\sqrt{2\text{KL}(P_0, P_i)}.$$

By the divergence decomposition lemma [Lattimore and Szepesvári, 2020, Lemma 15.1], we have:

$$\begin{aligned}\text{KL}(P_0, P_i) &= \mathbb{E}_0[K_i] \cdot T_\star \cdot \text{KL}(\text{Bernoulli}((B_\star + \epsilon_c)/T_\star), \text{Bernoulli}(B_\star/T_\star)) \\ &\leq \mathbb{E}_0[K_i] \cdot T_\star \cdot \frac{\epsilon_c^2/T_\star^2}{\frac{B_\star}{T_\star}(1 - \frac{B_\star}{T_\star})} \leq \frac{2\epsilon_c^2}{B_\star}\mathbb{E}_0[K_i]. \\ &\hspace{6cm}\text{([Gerchinovitz and Lattimore, 2016, Lemma 6])}\end{aligned}$$

Therefore, by Cauchy-Schwarz inequality,

$$\sum_{i=1}^N \mathbb{E}_i[K_i] \leq \sum_{i=1}^N \left(\mathbb{E}_0[K_i] + 2\epsilon_c K\sqrt{\mathbb{E}_0[K_i]/B_\star}\right) \leq K + 2\epsilon_c K\sqrt{NK/B_\star}.$$

Plugging this back and by the definition of $\epsilon_c$, we obtain

$$\mathbb{E}[R_K] \geq \epsilon_c K\left(1 - \frac{1}{N} - 2\epsilon_c\sqrt{\frac{K}{NB_\star}}\right) = \frac{(1 - 1/N)^2}{8}\sqrt{B_\star NK} = \Omega(\sqrt{B_\star SAK}).$$

This completes the proof. $\square$

**Theorem 8.** *For any $B_\star, T_\star, SA, K$ with $B_\star \geq 1$, $T_\star \geq 3B_\star$, $K \geq SA \geq 10$, and any learner, there exists an SSP instance based on transition perturbation such that the regret of the learner after $K$ episodes is at least $\Omega(B_\star\sqrt{SAK})$.*

*Proof.* Consider a distribution of SSP instances which is uniform over $\{\mathcal{M}^K_{0,j}\}_j$ for $j \in [N]$. Let $\mathbb{E}_j$ be the expectation w.r.t $\mathcal{M}^K_{0,j}$, $P_j$ be the distribution of learner's observations w.r.t $\mathcal{M}^K_{0,j}$, and $K_j$ the number of visits to state $j$ in $K$ episodes. Also let $\epsilon_P = \epsilon_{P,K}$. The expected regret over this distribution of SSPs can be lower bounded as

$$\begin{aligned}\mathbb{E}[R_K] &= \frac{1}{N}\sum_{j=1}^N \mathbb{E}_j[R_K] \geq \frac{1}{N}\sum_{j=1}^N \mathbb{E}_j[K-K_j]\cdot B_\star\left(1-\frac{1}{1+\epsilon_P}\right)\\ &\geq \frac{B_\star\epsilon_P}{2}\left(K-\frac{1}{N}\sum_{j=1}^N\mathbb{E}_j[K_j]\right).\end{aligned}$$

Note that $\mathcal{M}^K_{0,0}$ has no "good" state. By Pinsker's inequality:

$$\mathbb{E}_j[K_j]-\mathbb{E}_0[K_j] \leq K\left\|P_j-P_0\right\|_1 \leq K\sqrt{2\text{KL}(P_0,P_j)}.$$

By the divergence decomposition lemma [Lattimore and Szepesvári, 2020, Lemma 15.1], we have:

$$\begin{aligned}\text{KL}(P_0,P_j) &= \mathbb{E}_0[K_j]\cdot\text{KL}(\text{Geometric}(1/T_\star),\text{Geometric}((1+\epsilon_P)/T_\star))\\ &= \mathbb{E}_0[K_j]\cdot T_\star\cdot\text{KL}(\text{Bernoulli}(1/T_\star),\text{Bernoulli}((1+\epsilon_P)/T_\star))\\ &\leq \mathbb{E}_0[K_j]\cdot T_\star\cdot\frac{\epsilon_P^2/T_\star^2}{\frac{1+\epsilon_P}{T_\star}(1-\frac{1+\epsilon_P}{T_\star})} \leq 2\epsilon_P^2\mathbb{E}_0[K_j].\end{aligned}$$

([Gerchinovitz and Lattimore, 2016, Lemma 6] and $\epsilon_P \leq \frac{1}{4}$)

Therefore, by Cauchy-Schwarz inequality,

$$\sum_{j=1}^N\mathbb{E}_j[K_j] \leq \sum_{j=1}^N\left(\mathbb{E}_0[K_j]+2\epsilon_P K\sqrt{\mathbb{E}_0[K_j]}\right) \leq K+2\epsilon_P K\sqrt{NK}.$$

Plugging this back and by the definition of $\epsilon_P$, we obtain

$$\mathbb{E}[R_K] \geq \frac{B_\star\epsilon_P K}{2}\left(1-\frac{1}{N}-2\epsilon_P\sqrt{\frac{K}{N}}\right) \geq \frac{(1-1/N)^2}{16}B_\star\sqrt{NK} = \Omega(B_\star\sqrt{SAK}).$$

This completes the proof. □

Now we are ready to prove Theorem 1.

*Proof of Theorem 1.* We construct a hard non-stationary SSP instance as follows: we divide $K$ episodes into $L = L_c + L_P$ epochs. Each of the first $L_c$ epochs has length $\frac{K}{2L_c}$, and the corresponding SSP is uniformly sampled from $\{\mathcal{M}^{K/(2L_c)}_{i,0}\}_{i\in[N]}$ independently; each of the last $L_P$ epochs has length $\frac{K}{2L_P}$, and the corresponding SSP is uniformly sampled from $\{\mathcal{M}^{K/(2L_P)}_{0,j}\}_{j\in[N]}$ independently. By Theorem 7 and Theorem 8, the regrets in each of the first $L_c$ epochs and each of the last $L_P$ epochs are of order $\Omega(\sqrt{B_\star SAK/L_c})$ and $\Omega(B_\star\sqrt{SAK/L_P})$ respectively. Moreover, the total change in cost and transition functions are upper bounded by $\frac{\epsilon_c L_c}{T_\star}$ and $\frac{2\epsilon_P L_P}{T_\star}$ respectively with $\epsilon_c = \epsilon_{c,\frac{K}{2L_c}}$ and $\epsilon_P = \epsilon_{P,\frac{K}{2L_P}}$. Now let $\frac{\epsilon_c L_c}{T_\star} = \Delta_c$ and $\frac{2\epsilon_P L_P}{T_\star} = \Delta_P$, we have $L_c = (\frac{4\Delta_c T_\star}{1-1/N})^{2/3}(\frac{K}{2NB_\star})^{1/3}$ and $L_P = (\frac{2\Delta_P T_\star}{1-1/N})^{2/3}(\frac{K}{2N})^{1/3}$, and the dynamic regret is of order $\Omega(L_c\cdot\sqrt{B_\star SAK/L_c}+L_P\cdot B_\star\sqrt{SAK/L_P}) = \Omega((B_\star SAT_\star(\Delta_c+B_\star^2\Delta_P))^{1/3}K^{2/3})$. □

## C Omitted Details in Section 4

**Notations** Under the protocol of Algorithm 1, for any $k \in [K]$, denote by $M_k$ the number of intervals in the first $k$ episodes. Clearly, $M = M_K$.

The following lemma is a more general version of Lemma 1.

**Lemma 15.** *For any $K' \in [K]$, $R_{K'} \leq \mathring{R}_{M_{K'}} + B_\star$.*

*Proof.* Let $\mathcal{I}_k$ be the set of intervals in episode $k$. Then the regret in episode $k$ satisfies

$$\begin{aligned}\sum_{m\in\mathcal{I}_k}\sum_{h=1}^{H_m} c_h^m - V_k^\star(s_1^k) &= \sum_{m\in\mathcal{I}_k}\left(\sum_{h=1}^{H_m} c_h^m - V_1^{\pi^\star,m}(s_1^m)\right) + \sum_{m\in\mathcal{I}_k} V_1^{\pi^\star,m}(s_1^m) - V_k^\star(s_1^k)\\ &\leq \sum_{m\in\mathcal{I}_k}(C^m - V_1^{\pi^\star,m}(s_1^m)) + \frac{B_\star}{2K},\end{aligned}$$

where the last step is by the definition of $c_{H_m+1}^m$ and $V_1^{\pi^\star,m}(s_1^m) \leq V_k^\star(s_1^m) + \frac{B_\star}{2K} \leq \frac{3}{2}B_\star$ by Lemma 46. Summing up over $k$ completes the proof. $\square$

**Lemma 16.** *Suppose algorithm $\mathfrak{A}$ ensures $\mathring{R}_{M'} = \tilde{\mathcal{O}}(\gamma_0 + \gamma_1 {M'}^{1/3} + \gamma_{\frac{1}{2}} {M'}^{1/2} + \gamma_2 {M'}^{2/3})$ for any number of intervals $M' \leq M$ with cetain probability. Then with the same probability, $M_{K'} = \tilde{\mathcal{O}}(K' + \gamma_0/B_\star + (\gamma_1/B_\star)^{3/2} + (\gamma_{\frac{1}{2}}/B_\star)^2 + (\gamma_2/B_\star)^3)$ and $\mathring{R}_{M_{K'}} = \tilde{\mathcal{O}}(\gamma_1 {K'}^{1/3} + \gamma_{\frac{1}{2}}{K'}^{1/2} + \gamma_2 {K'}^{2/3} + \gamma_1^{3/2}/B_\star^{1/2} + \gamma_{\frac{1}{2}}^2/B_\star + \gamma_2^3/B_\star^2 + \gamma_0)$ for any $K' \in [K]$.*

*Proof.* Fix a $K' \in [K]$. For any $M' \leq M_{K'}$, let $\mathcal{C}_g = \{m \in [M'] : s_{H_m+1}^m = g\}$. Then,

$$\begin{aligned}\mathring{R}_{M'} &= \sum_{m\in\mathcal{C}_g}(C^m - V_1^{\pi^\star,m}(s_1^m)) + \sum_{m\notin\mathcal{C}_g}(C^m - V_1^{\pi^\star,m}(s_1^m))\\ &= \tilde{\mathcal{O}}\left(\gamma_0 + \gamma_1 {M'}^{1/3} + \gamma_{\frac{1}{2}}{M'}^{1/2} + \gamma_2 {M'}^{2/3}\right).\end{aligned} \tag{4}$$

Note that $V_1^{\pi^\star,m}(s_1^m) \leq V_{k(m)}^\star(s_1^m) + \frac{B_\star}{2K} \leq \frac{3}{2}B_\star$ by Lemma 46. Moreover, $C^m \geq 2B_\star$ when $m \notin \mathcal{C}_g$. Therefore, $C^m - V_1^{\pi^\star,m}(s_1^m) \geq -\frac{3B_\star}{2}$ for $m \in \mathcal{C}_g$ and $C^m - V_1^{\pi^\star,m}(s_1^m) \geq \frac{B_\star}{2}$ for $m \notin \mathcal{C}_g$. Reorganizing terms and by $|\mathcal{C}_g| \leq K'$, we get:

$$\frac{B_\star M'}{2} \leq 2B_\star K' + \tilde{\mathcal{O}}\left(\gamma_0 + \gamma_1 {M'}^{1/3} + \gamma_{\frac{1}{2}}{M'}^{1/2} + \gamma_2 {M'}^{2/3}\right).$$

Solving a quadratic inequality w.r.t. $M'$, we get $M' = \tilde{\mathcal{O}}(K' + \gamma_0/B_\star + (\gamma_1/B_\star)^{3/2} + (\gamma_{\frac{1}{2}}/B_\star)^2 + (\gamma_2/B_\star)^3)$. Define $\gamma = \gamma_0/B_\star + (\gamma_1/B_\star)^{3/2} + (\gamma_{\frac{1}{2}}/B_\star)^2 + (\gamma_2/B_\star)^3$. Plugging the bound on $M'$ back to Eq. (4), we have

$$\begin{aligned}\mathring{R}_{M'} &= \tilde{\mathcal{O}}\left(\gamma_0 + \gamma_1{K'}^{1/3} + \gamma_{\frac{1}{2}}{K'}^{1/2} + \gamma_2{K'}^{2/3} + \gamma_1\gamma^{1/3} + \gamma_{\frac{1}{2}}\gamma^{1/2} + \gamma_2\gamma^{2/3}\right)\\ &= \tilde{\mathcal{O}}\left(\gamma_0 + \gamma_1{K'}^{1/3} + \gamma_{\frac{1}{2}}{K'}^{1/2} + \gamma_2{K'}^{2/3} + \gamma_1^{3/2}/B_\star^{1/2} + \gamma_{\frac{1}{2}}^2/B_\star + \gamma_2^3/B_\star^2 + B_\star\gamma\right)\\ &= \tilde{\mathcal{O}}\left(\gamma_0 + \gamma_1{K'}^{1/3} + \gamma_{\frac{1}{2}}{K'}^{1/2} + \gamma_2{K'}^{2/3} + \gamma_1^{3/2}/B_\star^{1/2} + \gamma_{\frac{1}{2}}^2/B_\star + \gamma_2^3/B_\star^2\right),\end{aligned}$$

where in the second last step we apply Young's inequality for product ($xy \leq x^p/p + y^q/q$ for $x \geq 0$, $y \geq 0$, $p > 1$, $q > 1$, and $\frac{1}{p} + \frac{1}{q} = 1$). Putting everything together and setting $M' = M_{K'}$ completes the proof. $\square$

## D Omitted Details in Section 5

**Extra Notations** Let $Q_h^m$, $V_h^m$, $x_m$ be the value of $Q_h$, $V_h$, and $x$ at the beginning of interval $m$, and $Q_{H+1}^m(s,a) = V_{H+1}^m(s)$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$.

### D.1 Proof of Theorem 2

We first prove two lemmas related to the optimism of $Q_h^m$. Define the following reference value function: $\mathring{Q}_h^m(s,a) = (\widehat{c}^m(s,a) + \bar{P}_{s,a}^m \mathring{V}_{h+1}^m - b^m(s,a,\mathring{V}_{h+1}^m) - \mathring{x}_m)_+$ for $h \in [H]$, where $\mathring{V}_h^m(s) = \operatorname{argmin}_a \mathring{Q}_h^m(s,a)$ for $h \in [H]$, $\mathring{V}_{H+1}^m = c_f$, $\mathring{Q}_{H+1}^m(s,a) = \mathring{V}_{H+1}^m(s)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $\mathring{x}_m = \Delta_{c,m} + 4B_\star\Delta_{P,m}$.

**Lemma 17.** *With probability at least* $1 - 2\delta$, $\mathring{Q}_h^m(s,a) \leq Q_h^{\star,m}(s,a)$ *for* $m \leq M$.

*Proof.* We prove this by induction on $h$. The base case of $h = H+1$ is clearly true. For $h \leq H$, by Lemma 48, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned}
\mathring{Q}_h^m(s,a) &= \widehat{c}^m(s,a) + \bar{P}_{s,a}^m \mathring{V}_{h+1}^m - b^m(s,a,\mathring{V}_{h+1}^m) - \mathring{x}_m \\
&\leq \widehat{c}^m(s,a) + \bar{P}_{s,a}^m V_{h+1}^{\star,m} - b^m(s,a,V_{h+1}^{\star,m}) - \mathring{x}_m \qquad \text{(by the induction step)} \\
&= \widehat{c}^m(s,a) + \widetilde{P}_{s,a}^m V_{h+1}^{\star,m} + (\bar{P}_{s,a}^m - \widetilde{P}_{s,a}^m) V_{h+1}^{\star,m} - b^m(s,a,V_{h+1}^{\star,m}) - \mathring{x}_m \\
&\overset{\text{(i)}}{\leq} \widehat{c}^m(s,a) + \widetilde{P}_{s,a}^m V_{h+1}^{\star,m} - \mathring{x}_m \overset{\text{(ii)}}{\leq} c^m(s,a) + P_{s,a}^m V_{h+1}^{\star,m} = Q_h^{\star,m}(s,a),
\end{aligned}$$

where in (i) we apply Lemma 8 with $|\{V_h^{\star,m}\}_{m,h}| \leq HK+1$ to obtain $(\bar{P}_{s,a}^m - \widetilde{P}_{s,a}^m)V_{h+1}^{\star,m} - b^m(s,a,V_{h+1}^{\star,m}) \leq 0$; in (ii) we apply Lemma 5, Lemma 2, and the definition of $\mathring{x}_m$. □

**Lemma 18.** *With probability at least* $1-2\delta$, $Q_h^m(s,a) \leq Q_h^{\star,m}(s,a) + (\Delta_{c,m} + 4B_\star\Delta_{P,m})(H-h+1)$ *and* $x_m \leq \max\{\frac{1}{mH}, 2(\Delta_{c,m} + 4B_\star\Delta_{P,m})\}$.

*Proof.* The second statement simply follows from Lemma 17, $Q_h^{\star,m}(s,a) \leq Q_h^{\pi^\star,m}(s,a) \leq 4B_\star = B/4$ by Lemma 2, and the computing procedure of $x_m$. We now prove $Q_h^m(s,a) \leq \mathring{Q}_h^m(s,a) + (\Delta_{c,m} + 4B_\star\Delta_{P,m})(H-h+1)$ by induction on $h$, and the first statement simply follows from $\mathring{Q}_h^m(s,a) \leq Q_h^{\star,m}(s,a)$ (Lemma 17). The statement is clearly true for $h = H+1$. For $h \leq H$, by the induction step and $\left\|V_{h+1}^m\right\|_\infty \leq B/4$ from the update rule, we have $V_{h+1}^m(s) \leq \min\{B/4, \mathring{V}_{h+1}^m(s) + (\Delta_{c,m} + 4B_\star\Delta_{P,m})(H-h)\} \leq \mathring{V}_{h+1}^m(s) + y_{h+1}^m \leq B$ for any $s \in \mathcal{S}_+$, where $y_h^m = \min\{B/4, (\Delta_{c,m} + 4B_\star\Delta_{P,m})(H-h+1)\}$. Thus,

$$\begin{aligned}
\bar{P}_{s,a}^m V_{h+1}^m - b^m(s,a,V_{h+1}^m) - x_m &\leq \bar{P}_{s,a}^m(\mathring{V}_{h+1}^m + y_{h+1}^m) - b^m(s,a,\mathring{V}_{h+1}^m + y_{h+1}^m) \\
&\hspace{10em} \text{(Lemma 48 and } x_m \geq 0) \\
&\leq \bar{P}_{s,a}^m \mathring{V}_{h+1}^m - b^m(s,a,\mathring{V}_{h+1}^m) - \mathring{x}_m + (\Delta_{c,m} + 4B_\star^m\Delta_{P,m})(H-h+1),
\end{aligned}$$

where in the last inequality we apply definition of $\mathring{x}_m$ and $b^m(s,a,\mathring{V}_{h+1}^m + y_{h+1}^m) = b^m(s,a,\mathring{V}_{h+1}^m)$ since constant offset does not change the variance. Then, $Q_h^m(s,a) \leq \mathring{Q}_h^m(s,a) + (\Delta_{c,m} + 4B_\star\Delta_{P,m})(H-h+1)$ by the update rule of $Q_h^m$ and the definition of $\mathring{Q}_h^m$. □

We are now ready to prove the main theorem, from which Theorem 2 is a simple corollary.

**Theorem 9.** *Algorithm 2 ensures with probability at least* $1 - 22\delta$, *for any* $M' \leq M$, $\mathring{R}_{M'} = \tilde{O}(\sqrt{B_\star SAL_{c,M'}M'} + B_\star\sqrt{SAL_{P,M'}M'} + B_\star SAL_{c,M'} + B_\star S^2AL_{P,M'} + \sum_{m=1}^{M'}(\Delta_{c,m} + B_\star\Delta_{P,m})H)$.

*Proof.* Note that with probability at least $1-2\delta$:

$$\begin{aligned}
\mathring{R}_{M'} &\leq \sum_{m=1}^{M'}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - V_1^{\star,m}(s_1^m)\right) && (V_1^{\star,m}(s_1^m)\leq V_1^{\pi^\star,m}(s_1^m))\\
&\leq \sum_{m=1}^{M'}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - V_1^{m}(s_1^m)\right) + \sum_{m=1}^{M'}(\Delta_{c,m}+4B_\star\Delta_{P,m})H && (\text{Lemma 18})\\
&\leq \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(c_h^m + V_{h+1}^m(s_{h+1}^m) - V_h^m(s_h^m)\right) + \sum_{m=1}^{M'}(\Delta_{c,m}+4B_\star\Delta_{P,m})H + \tilde{\mathcal{O}}\left(B_\star SAL_{M'}\right)\\
& && (c_{H_m+1}^m = \tilde{\mathcal{O}}(B_\star) \text{ and Lemma 12})\\
&\leq \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left((c_h^m - \widehat{c}_h^m) + (V_{h+1}^m(s_{h+1}^m) - P_h^m V_{h+1}^m) + (P_h^m - \bar{P}_h^m)V_{h+1}^m + b_h^m\right)\\
&\quad + 2\sum_{m=1}^{M'}(\Delta_{c,m}+4B_\star\Delta_{P,m})H + \tilde{\mathcal{O}}\left(B_\star SAL_{M'}\right),
\end{aligned}$$

where the last step is by the definitions of $V_h^m(s_h^m)$, $x_m \leq \max\{\frac{1}{mH}, 2(\Delta_{c,m}+4B_\star\Delta_{P,m})\}$ (Lemma 18), $\max\{a,b\}\leq\frac{a+b}{2}$, and $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{1}{mH} = \tilde{\mathcal{O}}(1)$. Now we bound the first three sums separately. For the first term, with probability at least $1-4\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c_h^m - \widehat{c}_h^m) = \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c_h^m - c^m(s_h^m,a_h^m)) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c^m(s_h^m,a_h^m) - \widehat{c}_h^m)\\
&\leq \tilde{\mathcal{O}}\left(\sqrt{C_{M'}} + \sqrt{SAL_{c,M'}C_{M'}} + SAL_{c,M'}\right) + 2\sum_{m=1}^{M'}\Delta_{c,m}H. \qquad (\text{Lemma 49 and Lemma 3})
\end{aligned}$$

For the second term, by Lemma 49, with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(V_{h+1}^m(s_{h+1}^m) - P_h^m V_{h+1}^m) = \tilde{\mathcal{O}}\left(\sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^m)} + B_\star\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m})} + \sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} - V_{h+1}^m)} + B_\star\right),\\
& \qquad (\textsc{Var}[X+Y]\leq 2(\textsc{Var}[X]+\textsc{Var}[Y]) \text{ and } \sqrt{a+b}\leq\sqrt{a}+\sqrt{b})
\end{aligned}$$

which is dominated by the upper bound of the third term below. For the third term, by $P_h^m V_{h+1}^m \leq \widetilde{P}_h^m V_{h+1}^m + 4B_\star(\Delta_{P,m} + \mathbf{n}_h^m)$, with probability at least $1-2\delta$,

$$\begin{aligned}
\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(P_h^m - \bar{P}_h^m)V_{h+1}^m &\leq \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\widetilde{P}_h^m - \bar{P}_h^m)V_{h+1}^m + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}4B_\star(\Delta_{P,m} + \mathbf{n}_h^m)\\
&\leq \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left((\widetilde{P}_h^m - \bar{P}_h^m)V_{h+1}^{\star,m} + (\widetilde{P}_h^m - \bar{P}_h^m)(V_{h+1}^m - V_{h+1}^{\star,m}) + 4B_\star\mathbf{n}_h^m\right) + \sum_{m=1}^{M'}4B_\star\Delta_{P,m}H\\
&= \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\mathbb{V}(P_h^m, V_{h+1}^{\star,m})}{\mathbf{N}_h^m}} + \frac{SB_\star}{\mathbf{N}_h^m} + \sqrt{\frac{S\mathbb{V}(P_h^m, V_{h+1}^m - V_{h+1}^{\star,m})}{\mathbf{N}_h^m}}\right) + \sum_{m=1}^{M'}B_\star\Delta_{P,m}H\right)\\
&\qquad \left(\mathbf{n}_h^m \leq \tfrac{1}{\mathbf{N}_h^m}, \text{ Lemma 8 with } |\{V_{h+1}^{\star,m}\}_{m,h}| \leq HK+1, \text{ and Lemma 6}\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m})} + \sqrt{S^2AL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} - V_{h+1}^m)}\right)\\
&\qquad + \tilde{\mathcal{O}}\left(B_\star S^2AL_{P,M'} + \sum_{m=1}^{M'}B_\star\Delta_{P,m}H\right). \qquad \text{(Cauchy-Schwarz inequality and Lemma 11)}
\end{aligned}$$

Moreover, by Lemma 10, with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{m=1}^{M'}\sum_{h=1}^{H_m}b_h^m &= \tilde{\mathcal{O}}\left(\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^m)} + B_\star S^{1.5}AL_{P,M'} + B_\star\sqrt{SAHL_{P,M'}\sum_{m=1}^{M'}\Delta_{P,m}}\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m})} + \sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^m - V_{h+1}^{\star,m})}\right)\\
&\qquad + \tilde{\mathcal{O}}\left(B_\star S^{1.5}AL_{P,M'} + \sum_{m=1}^{M'}B_\star\Delta_{P,m}H\right).\\
&\qquad (\textsc{Var}[X+Y] \leq 2\textsc{Var}[X] + 2\textsc{Var}[Y], \sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \text{ and AM-GM inequality})
\end{aligned}$$

which is dominated by the upper bound of the third term above. Putting everything together, we have with probability at least $1-11\delta$,

$$\begin{aligned}
\mathring{R}_{M'} &= \tilde{\mathcal{O}}\left(\sqrt{SAL_{c,M'}C_{M'}} + B_\star SAL_{c,M'} + \sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m})} + B_\star S^2AL_{P,M'}\right)\\
&\qquad + \tilde{\mathcal{O}}\left(\sqrt{S^2AL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} - V_{h+1}^m)} + \sum_{m=1}^{M'}(\Delta_{c,m} + B_\star\Delta_{P,m})H\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{SAL_{c,M'}C_{M'}} + \sqrt{B_\star SAL_{P,M'}C_{M'}} + B_\star SAL_{c,M'} + B_\star S^2AL_{P,M'}\right)\\
&\qquad + \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}(\Delta_{c,m} + B_\star\Delta_{P,m})H\right). \quad \text{(Lemma 19, Lemma 20 and AM-GM inequality)}
\end{aligned}$$

Note that $\mathring{R}_{M'} = \sum_{m=1}^{M'}(C^m - V_1^{\pi^\star,m}(s_1^m)) \geq C_{M'} - 4B_\star M'$ (Lemma 2). Reorganizing terms and solving a quadratic inequality (Lemma 45) w.r.t $C_{M'}$ gives $C_{M'} = \tilde{\mathcal{O}}(B_\star M')$ ignoring lower order terms. Plugging this back completes the proof. □

*Proof of Theorem 2.* Note that by by Line 3 and Line 4 of Algorithm 2, we have $L_c \leq \lceil\frac{M'}{W_c}\rceil$, $L_P \leq \lceil\frac{M'}{W_P}\rceil$, and the number of intervals between consecutive resets of $\mathbf{M}$ (or $\mathbf{N}$) are upper bounded by $W_c$ (or $W_P$), which gives

$$\sum_{m=1}^{M'}(\Delta_{c,m}+B_\star\Delta_{P,m})H \le \sum_{m=1}^{M'}(\Delta_{c,f^c(m)}+B_\star\Delta_{P,f^P(m)})H \le (W_c\Delta_c+B_\star W_P\Delta_P)H$$

Applying Theorem 9 completes the proof. □

### D.2 Proof of Theorem 3

We first show that Algorithm 3 ensures an anytime regret bound in $\mathring{\mathcal{M}}$.

**Theorem 10.** *With probability at least $1-22\delta$, Algorithm 3 ensures for any $M' \le M$, $\mathring{R}_{M'} = \tilde{\mathcal{O}}((B_\star SAT_{\max}\Delta_c)^{1/3}M'^{2/3} + B_\star(SAT_{\max}\Delta_P)^{1/3}M'^{2/3} + (B_\star SAT_{\max}\Delta_c)^{2/3}M'^{1/3} + B_\star(S^{2.5}AT_{\max}\Delta_P)^{2/3}M'^{1/3} + (\Delta_c + B_\star\Delta_P)T_{\max})$.*

*Proof.* It suffices to prove the desired inequality for $M' \in \{2^n-1\}_{n\in\mathbb{N}_+}$. Suppose $M' = 2^N - 1$ for some $N \ge 1$. By the doubling scheme, we run Algorithm 2 on intervals $[2^{n-1}, 2^n-1]$ for $n = 1, \ldots, N$, and the regret on intervals $[2^{n-1}, 2^n-1]$ is of order $\tilde{\mathcal{O}}((B_\star SAT_{\max}\Delta_c)^{1/3}(2^{n-1})^{2/3} + B_\star(SAT_{\max}\Delta_P)^{1/3}(2^{n-1})^{2/3} + (B_\star SAT_{\max}\Delta_c)^{2/3}(2^{n-1})^{1/3} + B_\star(S^{2.5}AT_{\max}\Delta_c)^{2/3}(2^{n-1})^{1/3} + (\Delta_c + B_\star\Delta_P)T_{\max})$ by Theorem 2 and the choice of $W_c$ and $W_P$. Summing over $n$ completes the proof. □

*Proof of Theorem 3.* By Lemma 16 and Theorem 10 with $\gamma_0 = (\Delta_c + B_\star\Delta_P)T_{\max}$, $\gamma_1 = (B_\star SAT_{\max}\Delta_c)^{2/3} + B_\star(S^{2.5}AT_{\max}\Delta_P)^{2/3}$, $\gamma_{\frac{1}{2}} = 0$, and $\gamma_2 = (B_\star SAT_{\max}\Delta_c)^{1/3} + B_\star(SAT_{\max}\Delta_P)^{1/3}$, we have $\gamma_1^{3/2}/B_\star^{1/2} = \tilde{\mathcal{O}}(B_\star^{1/2}SAT_{\max}\Delta_c + B_\star S^{2.5}AT_{\max}\Delta_P)$, $\gamma_2^3/B_\star^2 = \tilde{\mathcal{O}}(SAT_{\max}\Delta_c/B_\star^2 + B_\star SAT_{\max}\Delta_P)$, and thus $\mathring{R}_{M_{K'}} = \tilde{\mathcal{O}}((B_\star SAT_{\max}\Delta_c)^{1/3}K'^{2/3} + B_\star(SAT_{\max}\Delta_P)^{1/3}K'^{2/3} + (B_\star SAT_{\max}\Delta_c)^{2/3}K'^{1/3} + B_\star(S^{2.5}AT_{\max}\Delta_P)^{2/3}K'^{1/3} + B_\star^{1/2}SAT_{\max}\Delta_c + B_\star S^{2.5}AT_{\max}\Delta_P)$ for any $K' \in [K]$. Then by Lemma 15, we obtain the same bound as $\mathring{R}_{M_{K'}}$ for $R_{K'}$. □

### D.3 Auxiliary Lemmas

**Lemma 19.** *With probability at least $1-2\delta$, $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m}) = \tilde{\mathcal{O}}\left(B_\star C_{M'} + B_\star^2\right)$ for any $M' \le M$.*

*Proof.* Applying Lemma 9 with $\left\|V_h^{\star,m}\right\|_\infty \le 4B_\star$ (Lemma 2), with probability at least $1-2\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m}) \\
&= \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}V_{H_m+1}^{\star,m}(s_{H_m+1}^m)^2 + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}B_\star(V_h^{\star,m}(s_h^m) - P_h^mV_{h+1}^{\star,m})_+ + B_\star^2\right) \\
&= \tilde{\mathcal{O}}\left(B_\star C_{M'} + B_\star^2\right),
\end{aligned}$$

where in the last step we apply

$$(V_h^{\star,m}(s_h^m) - P_h^mV_{h+1}^{\star,m})_+ \le (Q_h^{\star,m}(s_h^m,a_h^m) - P_h^mV_{h+1}^{\star,m})_+ \le c^m(s_h^m,a_h^m),$$

and $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}c^m(s_h^m,a_h^m) = \tilde{\mathcal{O}}(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}c_h^m)$ by Lemma 50. □

**Lemma 20.** *With probability at least $1-9\delta$, for any $M' \le M$, $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} - V_{h+1}^m) = \tilde{\mathcal{O}}(B_\star\sqrt{B_\star SAL_{P,M'}C_{M'}} + B_\star\sqrt{SAL_{c,M'}C_{M'}} + B_\star^2S^2AL_{P,M'} + B_\star^2SAL_{c,M'} + \sum_{m=1}^{M'}B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H)$.*

*Proof.* Let $z_h^m = \min\{B/4, (\Delta_{c,m}+4B_\star\Delta_{P,m})H\}\mathbb{I}\{h \leq H\}$. By Lemma 18 and $\|V_h^m\|_\infty \leq B/4$, we have $V_h^{\star,m}(s) + z_h^m \geq V_h^m(s)$ for all $s \in \mathcal{S}_+$. Moreover, by Lemma 12,

$$
\begin{aligned}
&\sum_{m=1}^{M'} (V_{H_m+1}^{\star,m}(s_{H_m+1}^m) + z_{H_m+1}^m - V_{H_m+1}^m(s_{H_m+1}^m))^2 \\
&\leq \sum_{m=1}^{M'} (z_{H_m+1}^m)^2 \mathbb{I}\{s_{H_m+1}^m = g\} + \tilde{\mathcal{O}}\left(B_\star^2 \sum_{m=1}^{M'} \mathbb{I}\{H_m < H, s_{H_m+1}^m \neq g\}\right) \\
&= 4B_\star \sum_{m=1}^{M'} (\Delta_{c,m} + 4B_\star\Delta_{P,m})H + \tilde{\mathcal{O}}\left(B_\star^2 SAL_{M'}\right).
\end{aligned}
$$

Also note that

$$
\begin{aligned}
(*) &= \sum_{m=1}^{M'} B_\star \sum_{h=1}^{H_m} (V_h^{\star,m}(s_h^m) - V_h^m(s_h^m) - P_h^m V_{h+1}^{\star,m} + P_h^m V_{h+1}^m + z_h^m - z_{h+1}^m)_+ \\
&\leq \sum_{m=1}^{M'} B_\star \sum_{h=1}^{H_m} \left(c^m(s_h^m, a_h^m) + \widetilde{P}_h^m V_{h+1}^m - V_h^m(s_h^m) + 4B_\star \mathbf{n}_h^m\right)_+ + 2\sum_{m=1}^{M'} B_\star(\Delta_{c,m} + 4B_\star\Delta_{P,m})H \\
&\quad (V_h^{\star,m}(s_h^m) \leq Q_h^{\star,m}(s_h^m, a_h^m),\ z_h^m \geq z_{h+1}^m, \text{ and } P_h^m V_{h+1}^m \leq \widetilde{P}_h^m V_{h+1}^m + 4B_\star(\mathbf{n}_h^m + \Delta_{P,m})) \\
&\leq \sum_{m=1}^{M'} B_\star \sum_{h=1}^{H_m} (c^m(s_h^m, a_h^m) - \widehat{c}_h^m + (\widetilde{P}_h^m - \bar{P}_h^m) V_{h+1}^{\star,m} + (\widetilde{P}_h^m - \bar{P}_h^m)(V_{h+1}^m - V_{h+1}^{\star,m}) + b_h^m)_+ \\
&\quad + 3\sum_{m=1}^{M'} B_\star(\Delta_{c,m} + 4B_\star\Delta_{P,m})H + 4B_\star^2 \sum_{m=1}^{M'}\sum_{h=1}^{H_m} \mathbf{n}_h^m + \tilde{\mathcal{O}}(B_\star),
\end{aligned}
$$

where the last step is by the definitions of $V_h^m(s_h^m)$, $x_m \leq \max\{\frac{1}{mH}, 2(\Delta_{c,m} + 4B_\star\Delta_{P,m})\}$ (Lemma 18), $\max\{a,b\} \leq \frac{a+b}{2}$, and $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{1}{mH} = \tilde{\mathcal{O}}(1)$. Now by Lemma 3, Lemma 8, Lemma 6, and $\mathbf{n}_h^m \leq \frac{1}{\mathbf{N}_h^m}$, we continue with

$$
\begin{aligned}
(*) &= \tilde{\mathcal{O}}\left(B_\star\left(\sqrt{SAL_{c,M'}C_{M'}} + SAL_{c,M'} + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\mathbb{V}(P_h^m, V_{h+1}^{\star,m})}{\mathbf{N}_h^m}} + \sqrt{\frac{S\mathbb{V}(P_h^m, V_{h+1}^m - V_{h+1}^{\star,m})}{\mathbf{N}_h^m}}\right)\right)\right) \\
&\quad + \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B_\star^2 S}{\mathbf{N}_h^m} + \sum_{m=1}^{M'} B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H + B_\star\sum_{m=1}^{M'}\sum_{h=1}^{H_m} b_h^m\right) \\
&= \tilde{\mathcal{O}}\left(B_\star\sqrt{SAL_{c,M'}C_{M'}} + B_\star SAL_{c,M'} + B_\star\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m})} + B_\star^2 S^2 AL_{P,M'}\right) \\
&\quad + \tilde{\mathcal{O}}\left(B_\star\sqrt{S^2AL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} - V_{h+1}^m)} + \sum_{m=1}^{M'} B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H\right),
\end{aligned}
$$

where in the last step we apply Cauchy-Schwarz inequality, Lemma 11, Lemma 10, $\textsc{Var}[X+Y] \leq 2\textsc{Var}[X] + 2\textsc{Var}[Y]$, and AM-GM inequality. Finally, by Lemma 19, we continue with

$$
\begin{aligned}
(*) &= \tilde{\mathcal{O}}\left(B_\star\sqrt{SAL_{c,M'}C_{M'}} + B_\star SAL_{c,M'} + B_\star\sqrt{B_\star SAL_{P,M'}C_{M'}} + B_\star^2 S^2 AL_{P,M'}\right) \\
&\quad + \tilde{\mathcal{O}}\left(B_\star\sqrt{S^2AL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} - V_{h+1}^m)} + \sum_{m=1}^{M'} B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H\right).
\end{aligned}
$$

Applying Lemma 9 on value functions $\{V_h^{\star,m} + z_h^m - V_h^m\}_{m,h}$ (constant offset does not change the variance) and plugging in the bounds above, we have

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} - V_{h+1}^m) = \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} + z_h^m - V_{h+1}^m)$$
$$= \tilde{\mathcal{O}}\left(B_\star\sqrt{SAL_{c,M'}C_{M'}} + B_\star^2 SAL_{c,M'} + B_\star\sqrt{B_\star SAL_{P,M'}C_{M'}} + B_\star^2 S^2 AL_{P,M'}\right)$$
$$+ \tilde{\mathcal{O}}\left(B_\star\sqrt{S^2AL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} - V_{h+1}^m)} + \sum_{m=1}^{M'}B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H\right).$$

Then solving a quadratic inequality w.r.t $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\star,m} - V_{h+1}^m)$ (Lemma 45) completes the proof. □

### D.4 Minimax Optimal Bound in Finite-Horizon MDP

Here we give a high level arguments on why Algorithm 2 implies a minimax optimal dynamic regret bound in the finite-horizon setting. To adapt Algorithm 2 to the non-homogeneous finite-horizon setting, we maintain empirical cost and transition functions for each layer $h \in [H]$ and let $c_f(s) = 0$. Following similar arguments and substituting $B_\star, T_{\max}$ by horizon $H$, Theorem 2 implies (ignoring lower order terms)

$$\mathring{R}_{M'} = \tilde{\mathcal{O}}\left(\sqrt{SAH^2/W_c}M' + \sqrt{SAH^3/W_P}M' + (\Delta_c W_c + H\Delta_P W_P)H\right)$$
$$= \tilde{\mathcal{O}}\left(H(SA\Delta_c)^{1/3}M'^{2/3} + (SAH^5\Delta_P)^{1/3}M'^{2/3}\right),$$

where the extra $\sqrt{H}$ dependency in the first two terms comes from estimating the cost and transition functions of each layer independently, and we set $W_c = (SA)^{1/3}(M'/\Delta_c)^{2/3}$, $W_P = (SA/H)^{1/3}(M'/\Delta_P)^{2/3}$. Note that the lower bound construction in [Mao et al., 2021] only make use of non-stationary transition. The lower bound they prove is $\Omega((SA\Delta)^{1/3}(HT)^{2/3})$ (their Theorem 5), which actually matches our upper bound $\tilde{\mathcal{O}}((SAH^5\Delta_P)^{1/3}M'^{2/3})$ for non-stationary transition since $T = M'H$ and $\Delta = H\Delta_P$ by their definition of non-stationarity. It is also straightforward to show that the lower bound for non-stationary cost matches our upper bound following similar arguments in proving Theorem 1.

## E Omitted Details in Section 6

**Notations** Denote by $\rho_m^c$ and $\rho_m^P$ the values of $\rho^c$ and $\rho^P$ at the beginning of interval $m$ respectively, that is, $\rho_m^c = g^c(\nu_m^c)$ and $\rho_m^P = g^P(\nu_m^P)$, where $g^c(m) = \min\{\frac{c_1}{\sqrt{m}}, \frac{1}{2^8H}\}$ and $g^P(m) = \min\{\frac{c_2}{\sqrt{m}}, \frac{1}{2^8H}\}$. Denote by $\check{c}^m$ the value of $\check{c}$ at the beginning of interval $m$ and define $\check{c}_h^m = \check{c}(s_h^m, a_h^m)$. Define $\check{Q}_h^{\pi^\star,m}$ and $\check{V}_h^{\pi^\star,m}$ as the action-value function and value function w.r.t cost $c^m + 8\eta_m$, transition $P^m$, and policy $\pi^\star_{k(m)}$; and $C_{[i,j]} = \sum_{m=i}^{j}\sum_{h=1}^{H_m}C^m$. Let $\check{Q}_h^{\star,m}$ and $\check{V}_h^{\star,m}$ be the optimal value functions w.r.t cost function $c^m + 8\eta_m$ and transition function $P^m$. It is not hard to see that they can be defined recursively as follows: $\check{V}_{H+1}^{\star,m} = c_f$ and for $h \le H$,

$$\check{Q}_h^{\star,m}(s,a) = c^m(s,a) + 8\eta_m + P_{s,a}^m\check{V}_{h+1}^{\star,m}, \qquad \check{V}_h^{\star,m}(s) = \min_a Q_h^{\star,m}(s,a).$$

For notational convenience, define $\check{Q}_{H+1}^m(s,a) = \check{V}_{H+1}^m(s)$, $\check{Q}_{H+1}^{\pi^\star,m}(s,a) = \check{V}_{H+1}^{\pi^\star,m}(s)$, and $\check{Q}_{H+1}^{\star,m}(s,a) = \check{V}_{H+1}^{\star,m}(s)$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$; let $L_c = L_{c,[1,K]}$ and $L_P = L_{P,[1,K]}$.

**Proof Sketch of Theorem 4** We give a high level idea on the analysis of the main theorem and also point out the key technical challenges. We decompose the regret as follows:

$$\mathring{R}_K = \sum_{m=1}^{K}(C^m - \check{V}_1^m(s_1^m)) + \sum_{m=1}^{K}(\check{V}_1^m(s_1^m) - \check{V}_1^{\pi^\star,m}(s_1^m)) + 8T_\star\sum_{m=1}^{K}\eta_m$$
$$\lesssim \sum_{m=1}^{K}\sum_{h=1}^{H}\left(c_h^m - \widehat{c}_h^m + \check{V}_{h+1}^m(s_{h+1}^m) - \bar{P}_h^m\check{V}_{h+1}^m + b_h^m - 8\eta_m\right) \quad \text{(definition of } \check{V}_h^m(s_h^m))$$
$$+ \sum_{m=1}^{K}(\check{V}_1^m(s_1^m) - \check{V}_1^{\pi^\star,m}(s_1^m)) + 8T_\star\sum_{m=1}^{K}\eta_m.$$

We bound the three terms above separately. For the second term, we first show that $\check{V}_1^m(s_1^m) - \check{V}_1^{\pi^\star,m}(s_1^m) \leq (\Delta_{c,m} + B\Delta_{P,m})T_\star$, where $\Delta_{c,m} = \Delta_{c,[i_m^c,m]}$, $\Delta_{P,m} = \Delta_{P,[i_m^P,m]}$ are the accumulated cost and transition non-stationarity since the last reset respectively. Although proving such a bound is straightforward when $\check{V}_h^m$ is indeed a value function (similar to Lemma 14), it is non-trivial under the UCBVI update rule as the bonus term $b$ depends on the next-step value function and can not be simply treated as part of the cost function. A key step here is to make use of the monotonic property (Lemma 48) of the bonus function; see Lemma 22 for more details. Now by the periodic resets of cost and transition counters (Line 4 and Line 5), the number of intervals between consecutive resets of cost and transition estimation is upper bounded by $W_c$ and $W_P$ respectively. Thus,

$$\sum_{m=1}^{K}(\Delta_{c,m} + B\Delta_{P,m})T_\star \leq \sum_{m=1}^{K}(\Delta_{c,f^c(m)} + B\Delta_{P,f^P(m)})T_\star \leq (W_c\Delta_c + BW_P\Delta_P)T_\star$$
$$= \tilde{\mathcal{O}}\left((B_\star SAT_\star\Delta_c)^{1/3}K^{2/3} + B_\star(SAT_\star\Delta_P)^{1/3}K^{2/3} + (\Delta_c + B_\star\Delta_P)T_\star\right).$$

where the last step is simply by the chosen values of $W_c$ and $W_P$.

For the third term, we have:

$$T_\star\sum_{m=1}^{K}\eta_m \leq T_\star\sum_{m=1}^{K}\left(\frac{c_1}{\sqrt{\nu_m^c}} + \frac{Bc_2}{\sqrt{\nu_m^P}}\right) = \tilde{\mathcal{O}}\left(T_\star\left(c_1\sum_{i=1}^{L_c}\sqrt{M_i^c} + B_\star c_2\sum_{i=1}^{L_P}\sqrt{M_i^P}\right)\right)$$
$$= \tilde{\mathcal{O}}\left(T_\star(c_1\sqrt{L_cK} + B_\star c_2\sqrt{L_PK})\right) = \tilde{\mathcal{O}}\left(\sqrt{B_\star SAL_cK} + B_\star\sqrt{SAL_PK}\right),$$

where $M_i^c$ (or $M_i^P$) is the number of intervals between the $i$-th and $(i+1)$-th reset of cost (or transition) estimation, and the second last step is by Cauchy-Schwarz inequality. Finally we bound the first term, simply by **Test 1** and **Test 2**, we have (only keeping the dominating terms)

$$\sum_{m=1}^{K}\sum_{h=1}^{H}\left(c_h^m - \widehat{c}_h^m + \check{V}_{h+1}^m(s_{h+1}^m) - \bar{P}_h^m\check{V}_{h+1}^m + b_h^m - 8\eta_m\right)$$
$$= \sum_{i=1}^{L_c}\sum_{m\in\mathcal{I}_i^c}\sum_{h=1}^{H_m}(c_h^m - \widehat{c}_h^m) + \sum_{i=1}^{L_P}\sum_{m\in\mathcal{I}_i^P}\sum_{h=1}^{H_m}(\check{V}_{h+1}^m(s_{h+1}^m) - \bar{P}_h^m\check{V}_{h+1}^m) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(b_h^m - 8\eta_m)$$
$$\lesssim \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\widehat{c}_h^m}{\mathbf{M}_h^m}} + \sqrt{\frac{\mathbb{V}(\bar{P}_h^m, \check{V}_{h+1}^m)}{\mathbf{N}_h^m}}\right) = \tilde{\mathcal{O}}\left(\sqrt{B_\star SAL_cK} + B_\star\sqrt{SAL_PK}\right).$$

where $\{\mathcal{I}_i^c\}_{i=1}^{L_c}$ (or $\{\mathcal{I}_i^P\}_{i=1}^{L_P}$) is a partition of $K$ episodes such that $\mathbf{M}$ (or $\mathbf{N}$) is reseted in the last interval of each $\mathcal{I}_i^c$ (or $\mathcal{I}_i^P$) for $i < L_c$ (or $i < L_P$) and the last interval of $\mathcal{I}_{L_c}^c$ (or $\mathcal{I}_{L_P}^P$) is $K$, and in the second last step we apply the definition of $\chi_m^c$ (Lemma 24) and $\chi_m^P$ (Lemma 25). Note that the regret of non-stationarity along the learner's trajectory is cancelled out by the negative correction term $-8\eta_m$. Now it suffices to bound $L_c$ and $L_P$. It can be shown that the reset rules of the non-stationarity tests guarantee that

$$L_c = \tilde{\mathcal{O}}\left(K/W_c + B_\star K/W_P\right), \quad L_P = \tilde{\mathcal{O}}\left(K/W_P + K/(B_\star W_c)\right).$$

Details are deferred to Lemma 26. Putting everything together completes the proof.

Next, we present three lemmas related to the optimism and magnitude (**Test 3**) of estimated value function.

**Lemma 21.** *With probability at least* $1-2\delta$, *for all* $m \le K$, $\check{Q}_h^m(s,a) \le \check{Q}_h^{\star,m}(s,a) + (\Delta_{c,m} + B\Delta_{P,m})(H-h+1)$.

*Proof.* We prove this by induction on $h$. The base case of $h = H+1$ is clearly true. For $h \le H$, by **Test 3** and the induction step, we have $\check{V}_{h+1}^m(s) \le \min\{B/2, \check{V}_{h+1}^{\star,m}(s) + (\Delta_{c,m} + B\Delta_{P,m})(H-h)\} \le \check{V}_{h+1}^{\star,m}(s) + x_{h+1}^m \le B$ where $x_h^m = \min\{B/2, (\Delta_{c,m} + B\Delta_{P,m})(H-h+1)\}$ and $\check{V}_h^{\star,m}(s) \le \check{V}_h^{\pi^\star,m}(s) \le \frac{B}{4} + 8H\eta_m \le \frac{B}{3}$. Thus, with probability at least $1-2\delta$,

$$
\begin{aligned}
&\check{c}^m(s,a) + \bar{P}_{s,a}^m \check{V}_{h+1}^m - b^m(s,a,\check{V}_{h+1}^m) \\
&\le \check{c}^m(s,a) + \bar{P}_{s,a}^m(\check{V}_{h+1}^{\star,m} + x_{h+1}^m) - b^m(s,a,\check{V}_{h+1}^{\star,m} + x_{h+1}^m) && \text{(Lemma 48)} \\
&\overset{\text{(i)}}{\le} \check{c}^m(s,a) + \widetilde{P}_{s,a}^m(\check{V}_{h+1}^{\star,m} + x_{h+1}^m) && \text{(Lemma 8)} \\
&\le c^m(s,a) + 8\eta_m + \Delta_{c,m} + P_{s,a}^m(\check{V}_{h+1}^{\star,m} + x_{h+1}^m) + B\Delta_{P,m} && \text{(Lemma 5)} \\
&\le \check{Q}_h^{\star,m}(s,a) + (\Delta_{c,m} + B\Delta_{P,m})(H-h+1).
\end{aligned}
$$

Note that in (i) we use the fact that $|\{\check{V}_h^{\star,m} + x_h^m\}_{m,h}| \le (HK+1)^6$ since $|\{(c^m, P^m)\}_m| \le K$, $|\{\rho_m^c\}_m| \le K$, $|\{\rho_m^P\}_m| \le K$, $|\{\Delta_{c,m}\}_m| \le K+1$, and $|\{\Delta_{P,m}\}_m| \le K+1$ ($\Delta_{c,m} = \Delta_{P,m} = 0$ when $m$ is not the first interval of some episode). □

**Lemma 22.** *With probability at least* $1-2\delta$, *for all* $m \le K$, $\check{Q}_h^m(s,a) \le \check{Q}_h^{\pi^\star,m}(s,a) + (\Delta_{c,m} + B\Delta_{P,m})T_h^{\pi^\star,m}(s,a)$.

*Proof.* We prove this by induction on $h$. The base case of $h = H+1$ is clearly true. For $h \le H$, by **Test 3** and the induction step, we have $\check{V}_{h+1}^m(s) \le \min\{B/2, \check{V}_{h+1}^{\pi^\star,m}(s) + (\Delta_{c,m} + B\Delta_{P,m})T_{h+1}^{\pi^\star,m}(s)\} \le \check{V}_{h+1}^{\pi^\star,m}(s) + x_{h+1}^m(s) \le B$ where $x_h^m(s) = \min\{B/2, (\Delta_{c,m} + B\Delta_{P,m})T_h^{\pi^\star,m}(s)\}$ and $\check{V}_h^{\pi^\star,m}(s) \le \frac{B}{4} + 8\eta_m T_h^{\pi^\star,m}(s) \le \frac{B}{4} + 8H\eta_m \le \frac{B}{3}$. Thus, with probability at least $1-2\delta$,

$$
\begin{aligned}
&\check{c}^m(s,a) + \bar{P}_{s,a}^m \check{V}_{h+1}^m - b^m(s,a,\check{V}_{h+1}^m) \\
&\le \check{c}^m(s,a) + \bar{P}_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + x_{h+1}^m) - b^m(s,a,\check{V}_{h+1}^{\pi^\star,m} + x_{h+1}^m) && \text{(Lemma 48)} \\
&\overset{\text{(i)}}{\le} \check{c}^m(s,a) + \widetilde{P}_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + x_{h+1}^m) && \text{(Lemma 8)} \\
&\le c^m(s,a) + 8\eta_m + \Delta_{c,m} + P_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + x_{h+1}^m) + B\Delta_{P,m} && \text{(Lemma 5)} \\
&\le \check{Q}_h^{\pi^\star,m}(s,a) + (\Delta_{c,m} + B\Delta_{P,m})T_h^{\pi^\star,m}(s,a).
\end{aligned}
$$

Note that in (i) we use the fact that $|\{\check{V}_h^{\pi^\star,m} + x_h^m\}_{m,h}| \le (HK+1)^6$ since $|\{V_h^{\pi^\star,m}\}_{m,h}| \le HK+1$, $|\{\rho_m^c\}_m| \le K$, $|\{\rho_m^P\}_m| \le K$, $|\{\Delta_{c,m}\}_m| \le K+1$, $|\{\Delta_{P,m}\}_m| \le K+1$ ($\Delta_{c,m} = \Delta_{P,m} = 0$ when $m$ is not the first interval of some episode), and $|\{T_h^{\pi^\star,m}\}_{m,h}| \le HK+1$. □

**Lemma 23.** *With probability at least* $1-2\delta$, *for all* $m \le K$, *if* $\Delta_{c,m} \le \rho_m^c$ *and* $\Delta_{P,m} \le \rho_m^P$, *then* $\check{Q}_h^m(s,a) \le \check{Q}_h^{\pi^\star,m}(s,a) + \eta_m T_h^{\pi^\star,m}(s,a) \le B/2$. *Moreover, if* ***Test 3*** *fails in interval* $m$, *then* $\Delta_{c,[i_m^c,m+1]} > g^c(\nu_m^c + 1)$ *or* $\Delta_{P,[i_m^P,m+1]} > g^P(\nu_m^P + 1)$.

*Proof.* First note that $\check{Q}_h^{\pi^\star,m}(s,a) \le \frac{B}{4} + 8\eta_m T_h^{\pi^\star,m}(s,a) \le \frac{B}{4} + 8H\eta_m \le \frac{B}{3}$. We prove the first statement by induction on $h$. The base case of $h = H+1$ is clearly true. For $h \le H$, note that:

$$
\begin{aligned}
&\check{c}^m(s,a) + \bar{P}_{s,a}^m \check{V}_{h+1}^m - b^m(s,a,\check{V}_{h+1}^m) \\
&\le \check{c}^m(s,a) + \bar{P}_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + \eta_m T_{h+1}^{\pi^\star,m}) - b^m(s,a,\check{V}_{h+1}^{\pi^\star,m} + \eta_m T_{h+1}^{\pi^\star,m}) \\
&\hspace{20em} \text{(induction step and Lemma 48)} \\
&\overset{\text{(i)}}{\le} \check{c}^m(s,a) + \widetilde{P}_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + \eta_m T_{h+1}^{\pi^\star,m}) \qquad \text{(Lemma 8)} \\
&\le c^m(s,a) + 8\eta_m + \rho_m^c + P_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + \eta_m T_{h+1}^{\pi^\star,m}) + \rho_m^P(B/3 + H\eta_m) \\
&\hspace{12em} \text{(Lemma 5, } \Delta_{c,m} \le \rho_m^c \text{, and } \Delta_{P,m} \le \rho_m^P) \\
&\le \check{Q}_h^{\pi^\star,m}(s,a) + \eta_m T_h^{\pi^\star,m}(s,a). \qquad (H\eta_m \le B/12)
\end{aligned}
$$

Note that in (i) we use the fact that $|\{\check{V}_h^{\pi^\star,m} + \eta_m T_h^{\pi^\star,m}\}_{m,h}| \le (HK+1)^6$ since $|\{V_h^{\pi^\star,m}\}_{m,h}| \le HK+1$, $|\{\rho_m^c\}_m| \le K$, $\{\rho_m^P\}_m \le K$, and $|\{T_h^{\pi^\star,m}\}_{m,h}| \le HK+1$. The second statement is simply by the contraposition of the first statement. □

The next two lemmas are about **Test 1** and **Test 2**.

**Lemma 24.** *With probability at least $1-4\delta$, for any $M' \le K$, if $\Delta_{c,M'} \le \rho_{M'}^c$, then*

$$
\sum_{m=i_{M'}^c}^{M'} \sum_{h=1}^{H_m} (c_h^m - \widehat{c}_h^m) \le \tilde{\mathcal{O}}\left(\sqrt{C_{[i_{M'}^c,M']}} + \sum_{m=i_{M'}^c}^{M'} \sum_{h=1}^{H_m}\left(\sqrt{\frac{\widehat{c}_h^m}{\mathbf{M}_h^m}} + \frac{1}{\mathbf{M}_h^m}\right)\right) + \sum_{m=i_{M'}^c}^{M'} \sum_{h=1}^{H_m} \rho_m^c \triangleq \chi_{M'}^c.
$$

*Moreover, if **Test 1** fails in interval $M'$, then $\Delta_{c,M'} > \rho_{M'}^c$.*

*Proof.* Note that for any given $M' \le M$, without loss of generality, we can offset the intervals and assume $i_{M'}^c = 1$. Then with probability at least $1-4\delta$, for any $M' \le K$, assuming $i_{M'}^c = 1$ we have

$$
\begin{aligned}
\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c_h^m - \widehat{c}_h^m) &= \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c_h^m - c^m(s_h^m,a_h^m)) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c^m(s_h^m,a_h^m) - \widehat{c}_h^m) \\
&\le \tilde{\mathcal{O}}\left(\sqrt{C_{M'}}\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c^m(s_h^m,a_h^m) - \widehat{c}_h^m) \quad \text{(Lemma 49 and Lemma 50)} \\
&\le \tilde{\mathcal{O}}\left(\sqrt{C_{M'}} + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\widehat{c}_h^m}{\mathbf{M}_h^m}} + \frac{1}{\mathbf{M}_h^m}\right)\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\rho_m^c. \\
&\hspace{12em} \text{(Lemma 3, and } \Delta_{c,m} \le \Delta_{c,M'} \le \rho_{M'}^c \le \rho_m^c)
\end{aligned}
$$

The first statement is then proved by noting $i_{M'}^c = 1$. The second statement is simply by the contraposition of the first statement. □

**Lemma 25.** *With probability at least $1-16\delta$, for any $M' \le K$, if $\Delta_{c,[i_{M'}^P,M']} \le \bar{\rho}_{M'}^c \triangleq \min\{\frac{B_\star^{1.5}c_1}{\sqrt{\nu_{M'}^P}}, \frac{1}{2^8H}\}$ and $\Delta_{P,M'} \le \rho_{M'}^P$, then*

$$
\begin{aligned}
\sum_{m=i_{M'}^P}^{M'}\sum_{h=1}^{H_m}\left(\check{V}_{h+1}^m(s_{h+1}^m) - \bar{P}_h^m\check{V}_{h+1}^m\right) &\le \tilde{\mathcal{O}}\left(\sqrt{\sum_{m=i_{M'}^P}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(\bar{P}_h^m,\check{V}_{h+1}^m)} + \sum_{m=i_{M'}^P}^{M'}\sum_{h=1}^{H_m}\sqrt{\frac{\mathbb{V}(\bar{P}_h^m,\check{V}_{h+1}^m)}{\mathbf{N}_h^m}}\right) \\
&+ \tilde{\mathcal{O}}\left(\sqrt{SA(B_\star + L_{c,[i_{M'}^P,M']})C_{[i_{M'}^P,M']}} + \sqrt{B_\star SA\nu_{M'}^P} + B_\star^{2.5}S^2AHL_{c,[i_{M'}^P,M']}\right) + 4\sum_{m=i_{M'}^P}^{M'}\sum_{h=1}^{H_m}\eta_m \triangleq \chi_{M'}^P.
\end{aligned}
$$

*Moreover, if **Test 2** fails in interval $M'$, then $\Delta_{c,[i_{M'}^P,M']} > \bar{\rho}_{M'}^c$ or $\Delta_{P,M'} > \rho_{M'}^P$.*

*Proof.* For any $M' \le K$, without loss of generality, we can offset the intervals and assume $i^P_{M'} = 1$. Moreover, for any $m \le M'$, we have $\Delta_{P,m} \le \Delta_{P,M'} \le \rho^P_{M'} \le \rho^P_m$. Thus, with probability at least $1 - 2\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\check{V}^m_{h+1}(s^m_{h+1}) - \bar{P}^m_h\check{V}^m_{h+1}\right)\\
&\le \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\check{V}^m_{h+1}(s^m_{h+1}) - P^m_h\check{V}^m_{h+1}) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\widetilde{P}^m_h - \bar{P}^m_h)\check{V}^m_{h+1} + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}B(\rho^P_m + \mathbf{n}^m_h)\\
&\qquad\qquad (P^m_h\check{V}^m_{h+1} \le \widetilde{P}^m_h\check{V}^m_{h+1} + B(\Delta_{P,m} + \mathbf{n}^m_h) \text{ and } \Delta_{P,m} \le \rho^P_m)\\
&\le \tilde{\mathcal{O}}\left(\sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P^m_h, \check{V}^m_{h+1})} + B_\star SA\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\widetilde{P}^m_h - \bar{P}^m_h)\check{V}^m_{h+1} + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}B\rho^P_m\\
&\qquad\qquad (\text{Lemma 49 and } \textstyle\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbf{n}^m_h \le \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{1}{\mathbf{N}^m_h} \le SA \text{ by } L_{P,M'} = 1)\\
&\le \tilde{\mathcal{O}}\left(\sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(\bar{P}^m_h, \check{V}^m_{h+1})} + B_\star SA\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\widetilde{P}^m_h - \bar{P}^m_h)\check{V}^m_{h+1} + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}2B\rho^P_m,
\end{aligned}$$

where the last inequality is by

$$\begin{aligned}
\mathbb{V}(P^m_h, \check{V}^m_{h+1}) &\le P^m_h(\check{V}^m_{h+1} - \bar{P}^m_h\check{V}^m_{h+1})^2 \le \widetilde{P}^m_h(\check{V}^m_{h+1} - \bar{P}^m_h\check{V}^m_{h+1})^2 + B^2(\Delta_{P,m} + \mathbf{n}^m_h)\\
&\qquad\qquad (\tfrac{\sum_i p_i x_i}{\sum_i p_i} = \operatorname{argmin}_z \textstyle\sum_i p_i(x_i - z)^2)\\
&\le 2\mathbb{V}(\bar{P}^m_h, \check{V}^m_{h+1}) + \tilde{\mathcal{O}}\left(\frac{SB^2}{\mathbf{N}^m_h}\right) + B^2\rho^P_m,\\
&\qquad\qquad (\widetilde{P}^m_h(s') \le 2\bar{P}^m_h(s') + \tfrac{1}{\mathbf{N}^m_h} \text{ by Lemma 50, } \mathbf{n}^m_h \le \tfrac{1}{\mathbf{N}^m_h}, \text{ and } \Delta_{P,m} \le \rho^P_m)
\end{aligned}$$

Lemma 11, $L_{P,M'} = 1$, and AM-GM inequality. Now note that with probability at least $1 - 3\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\widetilde{P}^m_h - \bar{P}^m_h)\check{V}^m_{h+1} = \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left((\widetilde{P}^m_h - \bar{P}^m_h)\check{V}^{\star,m}_{h+1} + (\widetilde{P}^m_h - \bar{P}^m_h)(\check{V}^m_{h+1} - \check{V}^{\star,m}_{h+1})\right)\\
&\le \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\mathbb{V}(\bar{P}^m_h, \check{V}^{\star,m}_{h+1})}{\mathbf{N}^m_h}} + \frac{SB_\star}{\mathbf{N}^m_h}\right) + \sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P^m_h, \check{V}^m_{h+1} - \check{V}^{\star,m}_{h+1})}\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B\rho^P_m}{32}\\
&\qquad\qquad (\text{Lemma 8, Lemma 6, Cauchy-Schwarz inequality, Lemma 11, and } \Delta_{P,m} \le \rho^P_m)\\
&\le \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\mathbb{V}(\bar{P}^m_h, \check{V}^m_{h+1})}{\mathbf{N}^m_h}} + \frac{SB_\star}{\mathbf{N}^m_h}\right) + \sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P^m_h, \check{V}^m_{h+1} - \check{V}^{\star,m}_{h+1})}\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B\rho^P_m}{16},
\end{aligned}$$

where in the last step we apply

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\sqrt{\frac{\mathbb{V}(\bar{P}^m_h, \check{V}^{\star,m}_{h+1})}{\mathbf{N}^m_h}} \le \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\mathbb{V}(\bar{P}^m_h, \check{V}^m_{h+1})}{\mathbf{N}^m_h}} + \sqrt{\frac{\mathbb{V}(\bar{P}^m_h, \check{V}^m_{h+1} - \check{V}^{\star,m}_{h+1})}{\mathbf{N}^m_h}}\right)$$

by $\sqrt{\text{VAR}[X+Y]} \le \sqrt{\text{VAR}[X]} + \sqrt{\text{VAR}[Y]}$ [Cohen et al., 2021, Lemma E.3] and

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\sqrt{\frac{\mathbb{V}(\bar{P}_h^m, \check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m})}{\mathbf{N}_h^m}} \le \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\sqrt{\frac{\bar{P}_h^m((\check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m}) - P_h^m(\check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m}))^2}{\mathbf{N}_h^m}}$$
$$\left(\tfrac{\sum_i p_i x_i}{\sum_i p_i} = \operatorname{argmin}_z \textstyle\sum_i p_i(x_i - z)^2\right)$$
$$\le \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\sqrt{\frac{2\widetilde{P}_h^m((\check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m}) - P_h^m(\check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m}))^2}{\mathbf{N}_h^m}} + \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B\sqrt{S}}{\mathbf{N}_h^m}\right)$$
$$\left(\bar{P}_h^m(s') \le 2\widetilde{P}_h^m(s') + \tilde{\mathcal{O}}\left(\tfrac{1}{\mathbf{N}_h^m}\right) \text{ by Lemma 50}\right)$$
$$\le \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\sqrt{\frac{2\mathbb{V}(P_h^m, \check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m})}{\mathbf{N}_h^m}} + \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B\sqrt{S}}{\mathbf{N}_h^m} + \sum_{m=1}^{M'}\sum_{h=1}^{H_m} B\sqrt{\frac{\Delta_{P,m}}{\mathbf{N}_h^m}}\right)$$
$$\le \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m})} + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B\sqrt{S}}{\mathbf{N}_h^m}\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B\rho_m^P}{32}.$$
(Cauchy-Schwarz inequality, Lemma 11, $L_{P,M'} = 1$, AM-GM inequality, and $\Delta_{P,m} \le \rho_m^P$)

Now by Lemma 28, $L_{P,M'} = 1$, and AM-GM inequality, we have with probability $1 - 10\delta$,

$$\sqrt{S^2 A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m})} \le \tilde{\mathcal{O}}\left(\sqrt{SAL_{c,M'}C_{M'}} + \sqrt{B_\star SA(C_{M'} + M')}\right)$$
$$+ \tilde{\mathcal{O}}\left(B_\star S^2 A + B_\star S^{1.5} A L_{c,M'} + \sqrt{B_\star S^2 A \sum_{m=1}^{M'}(\Delta_{c,m} + B_\star \Delta_{P,m})H}\right).$$

Moreover, by $i_m^c \ge i_m^P$ and $\nu_m^c \le \nu_m^P$ due to the reset rules, we have $\Delta_{c,m} \le \Delta_{c,[i_{M'}^P,m]} \le \Delta_{c,[i_{M'}^P,M']} \le \bar{\rho}_{M'}^c \le \bar{\rho}_m^c \le B_\star^{1.5}\min\{\frac{c_1}{\sqrt{\nu_m^P}}, \frac{1}{2^8 H}\} \le B_\star^{1.5}\min\{\frac{c_1}{\sqrt{\nu_m^c}}, \frac{1}{2^8 H}\} \le B_\star^{1.5}\rho_m^c$. Therefore, by $\Delta_{P,m} \le \rho_m^P$ and AM-GM inequality,

$$\sqrt{B_\star S^2 A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\Delta_{c,m} + B_\star\Delta_{P,m})} \le \sqrt{B_\star^{2.5} S^2 AH \sum_{m=1}^{M'}(\rho_m^c + B_\star\rho_m^P)} \le B_\star^{2.5}S^2 AH + \sum_{m=1}^{M'}\eta_m.$$

Plugging these back, and by Lemma 11, $L_{P,M'} = 1$, we obtain

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\widetilde{P}_h^m - \bar{P}_h^m)\check{V}_{h+1}^m \le \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\mathbb{V}(\bar{P}_h^m, \check{V}_{h+1}^m)}{\mathbf{N}_h^m}}\right) + \sqrt{B_\star SA(C_{M'} + M')}\right)$$
$$+ \tilde{\mathcal{O}}\left(\sqrt{SAL_{c,M'}C_{M'}} + B_\star S^{1.5} A L_{c,M'} + B_\star^{2.5} S^2 AH\right) + 2\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\eta_m.$$

Plugging this back and noting $i_{M'}^P = 1$ completes the proof of the first statement. The second statement is simply by the contraposition of the first statement. $\square$

### E.1 Proof of Theorem 4

*Proof.* By $s_1^m = s_{\text{init}}$, we decompose the regret as follows, with probability at least $1-2\delta$,

$$
\begin{aligned}
\mathring{R}_K &= \sum_{m=1}^{K}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - V_1^{\pi^\star,m}(s_1^m)\right)\\
&= \sum_{m=1}^{K}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - \check{V}_1^m(s_1^m)\right) + \sum_{m=1}^{K}\left(\check{V}_1^m(s_1^m) - \check{V}_1^{\pi^\star,m}(s_1^m)\right) + 8T_\star\sum_{m=1}^{K}\eta_m\\
&\leq \sum_{m=1}^{K}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - \check{V}_1^m(s_1^m)\right) + \sum_{m=1}^{K}(\Delta_{c,m} + B\Delta_{P,m})T_\star + 8T_\star\sum_{m=1}^{K}\eta_m
\end{aligned}
$$

(Lemma 22)

We first bound the first and the third term above separately. For the third term, we have:

$$
\begin{aligned}
T_\star\sum_{m=1}^{K}\eta_m &\leq T_\star\sum_{m=1}^{K}\left(\frac{c_1}{\sqrt{\nu_m^c}} + \frac{Bc_2}{\sqrt{\nu_m^P}}\right) = \tilde{\mathcal{O}}\left(T_\star\left(c_1\sum_{i=1}^{L_c}\sqrt{M_i^c} + B_\star c_2\sum_{i=1}^{L_P}\sqrt{M_i^P}\right)\right)\\
&\qquad\qquad\left(\textstyle\sum_{i=1}^{j}\frac{1}{\sqrt{i}} = \mathcal{O}(\sqrt{j})\right)\\
&= \tilde{\mathcal{O}}\left(T_\star(c_1\sqrt{L_cK} + B_\star c_2\sqrt{L_PK})\right) = \tilde{\mathcal{O}}\left(\sqrt{B_\star SAL_cK} + B_\star\sqrt{SAL_PK}\right),
\end{aligned}
$$

where $M_i^c$ (or $M_i^P$) is the number of intervals between the $i$-th and $(i+1)$-th reset of cost (or transition) estimation, and the second last step is by Cauchy-Schwarz inequality. For the first term, define $\{\mathcal{I}_i^c\}_{i=1}^{L_c}$ (or $\{\mathcal{I}_i^P\}_{i=1}^{L_P}$) as a partition of $K$ episodes such that $\mathbf{M}$ (or $\mathbf{N}$) is reset in the last interval of each $\mathcal{I}_i^c$ (or $\mathcal{I}_i^P$) for $i < L_c$ (or $i < L_P$) and the last interval of $\mathcal{I}_{L_c}^c$ (or $\mathcal{I}_{L_P}^P$) is $K$. Also let $L = L_c + L_P$. Then with probability at least $1-20\delta$,

$$
\sum_{m=1}^{K}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - \check{V}_1^m(s_1^m)\right) \leq \sum_{m=1}^{K}\sum_{h=1}^{H_m}\left(c_h^m + \check{V}_{h+1}^m(s_{h+1}^m) - \check{V}_h^m(s_h^m)\right) + \tilde{\mathcal{O}}(B_\star SAL)
$$

(Lemma 12)

$$
\leq \sum_{m=1}^{K}\sum_{h=1}^{H_m}\left(c_h^m - \widehat{c}_h^m + \check{V}_{h+1}^m(s_{h+1}^m) - \bar{P}_h^m\check{V}_{h+1}^m + b_h^m - 8\eta_m\right) + \tilde{\mathcal{O}}(B_\star SAL)
$$

(definition of $\check{V}_h^m(s_h^m)$)

$$
\begin{aligned}
&= \sum_{i=1}^{L_c}\sum_{m\in\mathcal{I}_i^c}\sum_{h=1}^{H_m}(c_h^m - \widehat{c}_h^m) + \sum_{i=1}^{L_P}\sum_{m\in\mathcal{I}_i^P}\sum_{h=1}^{H_m}(\check{V}_{h+1}^m(s_{h+1}^m) - \bar{P}_h^m\check{V}_{h+1}^m) + \sum_{m=1}^{K}\sum_{h=1}^{H_m}(b_h^m - 8\eta_m) + \tilde{\mathcal{O}}(B_\star SAL)\\
&= \tilde{\mathcal{O}}\left(\sqrt{L_cC_K} + \sum_{m=1}^{K}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\bar{c}_h^m}{\mathbf{M}_h^m}} + \frac{1}{\mathbf{M}_h^m}\right) + \sqrt{L_P\sum_{m=1}^{K}\sum_{h=1}^{H_m}\mathbb{V}(\bar{P}_h^m, \check{V}_{h+1}^m)} + \sum_{m=1}^{K}\sum_{h=1}^{H_m} b_h^m\right)\\
&\quad + \tilde{\mathcal{O}}\left(B_\star^{2.5}S^2AHL_c + \sqrt{B_\star SAL_P(C_K+K)} + \sqrt{SAL_cC_K} + HL_c + B_\star HL_P\right),
\end{aligned}
$$

(**Test 1** (Lemma 24), **Test 2** (Lemma 25), and Cauchy-Schwarz inequality)

where $\tilde{\mathcal{O}}(HL_c + B_\star HL_P)$ is upper bound of the costs in intervals where **Test 1** fails or **Test 2** fails. By Lemma 3 and AM-GM inequality, with probability at least $1-3\delta$,

$$
\sum_{m=1}^{K}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\bar{c}_h^m}{\mathbf{M}_h^m}} + \frac{1}{\mathbf{M}_h^m}\right) = \tilde{\mathcal{O}}\left(SAHL_c + \sqrt{SAL_cC_K}\right) + \sum_{m=1}^{K}\Delta_{c,m}.
$$

Following the proof of Lemma 10, we have $\sqrt{L_P \sum_{m=1}^{K}\sum_{h=1}^{H_m} \mathbb{V}(\bar{P}_h^m, \check{V}_{h+1}^m)}$ is dominated by the upper bound of $\sum_{m=1}^{M'}\sum_{h=1}^{H_m} b_h^m$. Thus with probability at least $1-\delta$,

$$
\begin{aligned}
&\sqrt{L_P \sum_{m=1}^{K}\sum_{h=1}^{H_m} \mathbb{V}(\bar{P}_h^m, \check{V}_{h+1}^m)} + \sum_{m=1}^{K}\sum_{h=1}^{H_m} b_h^m \\
&= \tilde{\mathcal{O}}\left(\sqrt{SAL_P \sum_{m=1}^{K}\sum_{h=1}^{H_m} \mathbb{V}(P_h^m, \check{V}_{h+1}^m)} + B_\star S^{1.5} AL_P + B_\star \sqrt{SAL_P \sum_{m=1}^{K}\sum_{h=1}^{H_m} \Delta_{P,m}}\right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{B_\star SAL_P(C_K+K)} + \sqrt{SAL_c C_K} + B_\star S^{1.5} AHL\right) + \sum_{m=1}^{K}(\Delta_{c,m} + B_\star \Delta_{P,m}),
\end{aligned}
$$

where in the last inequality we apply AM-GM inequality on $B_\star\sqrt{SAL_P \sum_{m=1}^{K}\sum_{h=1}^{H_m} \Delta_{P,m}}$, and note that with probability at least $1-11\delta$,

$$
\begin{aligned}
&\sqrt{SAL_P \sum_{m=1}^{K}\sum_{h=1}^{H_m} \mathbb{V}(P_h^m, \check{V}_{h+1}^m)} \\
&= \tilde{\mathcal{O}}\left(\sqrt{SAL_P \sum_{m=1}^{K}\sum_{h=1}^{H_m} \mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m})} + \sqrt{SAL_P \sum_{m=1}^{K}\sum_{h=1}^{H_m} \mathbb{V}(P_h^m, \check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m})}\right) \\
&\qquad\qquad (\textsc{Var}[X+Y] \le 2\textsc{Var}[X] + 2\textsc{Var}[Y] \text{ and } \sqrt{a+b} \le \sqrt{a} + \sqrt{b}) \\
&= \tilde{\mathcal{O}}\left(\sqrt{B_\star SAL_P(C_K+K)} + \sqrt{SAL_c C_K} + B_\star S^{1.5} AHL\right) + \sum_{m=1}^{M'}(\Delta_{c,m} + B_\star \Delta_{P,m}). \\
&\qquad\qquad (\text{Lemma 27, Lemma 28, and AM-GM inequality})
\end{aligned}
$$

Putting everything together, we have

$$
\mathring{R}_K = \tilde{\mathcal{O}}\left(\sqrt{SA(L_c + B_\star L_P)(C_K + B_\star K)} + B_\star^{2.5} S^2 AHL + \sum_{m=1}^{K}(\Delta_{c,m} + B_\star \Delta_{P,m})T_\star\right).
$$

Now by $\mathring{R}_K \ge C_K - 4B_\star K$, solving a quadratic inequality (Lemma 45) w.r.t $C_K$ and plugging the bound on $C_K$ back, we obtain

$$
\mathring{R}_K = \tilde{\mathcal{O}}\left(\sqrt{B_\star SAL_c K} + B_\star\sqrt{SAL_P K} + B_\star^{2.5} S^2 AHL + \sum_{m=1}^{K}(\Delta_{c,m} + B_\star \Delta_{P,m})T_\star\right).
$$

It suffices to bound the last term above. By the periodic resets of $\mathbf{M}$ and $\mathbf{N}$ (Line 4 and Line 5 of Algorithm 4), the number of intervals between consecutive resets of $\mathbf{M}$ and $\mathbf{N}$ are upper bounded by $W_c$ and $W_P$ respectively. Thus,

$$
\begin{aligned}
&\sum_{m=1}^{K}(\Delta_{c,m} + B_\star \Delta_{P,m})T_\star \le \sum_{m=1}^{K}(\Delta_{c,f^c(m)} + B_\star \Delta_{P,f^P(m)})T_\star \le (W_c\Delta_c + B_\star W_P \Delta_P)T_\star \\
&= \tilde{\mathcal{O}}\left((B_\star SAT_\star\Delta_c)^{1/3}K^{2/3} + B_\star(SAT_\star\Delta_P)^{1/3}K^{2/3} + (\Delta_c + B_\star\Delta_P)T_\star\right),
\end{aligned}
$$

where the last step is simply by the chosen values of $W_c$ and $W_P$. Plugging this back and applying Lemma 26 completes the proof. □

**Lemma 26.** *With probability at least* $1-2\delta$*, Algorithm 4 with* $p = 1/B_\star$ *ensures*

$$
\begin{aligned}
L_c &= \tilde{\mathcal{O}}\left((B_\star SA)^{-1/3}(T_\star\Delta_c)^{2/3}K^{1/3} + B_\star(SA)^{-1/3}(T_\star\Delta_P)^{2/3}K^{1/3} + H(\Delta_c + B_\star\Delta_P)\right), \\
L_P &= \tilde{\mathcal{O}}\left((B_\star SA)^{-1/3}(T_\star\Delta_c)^{2/3}K^{1/3}/B_\star + (SA)^{-1/3}(T_\star\Delta_P)^{2/3}K^{1/3} + H(\Delta_c + \Delta_P)\right).
\end{aligned}
$$

*Proof.* We consider the number of resets of $\mathbf{M}$ and $\mathbf{N}$ from each test separately. By Lemma 24 and Lemma 13, there are at most $\tilde{\mathcal{O}}((c_1^{-1}\Delta_c)^{2/3}K^{1/3} + H\Delta_c)$ resets of $\mathbf{M}$ triggered by **Test 1**. By Lemma 25 and Lemma 13, there are at most $\tilde{\mathcal{O}}(((B_\star^{-1.5}c_1^{-1}\Delta_c)^{2/3} + (c_2^{-1}\Delta_P)^{2/3})K^{1/3} + H(\Delta_c + \Delta_P))$ resets of $\mathbf{M}$ and $\mathbf{N}$ triggered by **Test 2**.

Next, we consider **Test 3**. Define $\mathbb{I}_m^c = \mathbb{I}\{\Delta_{c,[i_m^c,m+1]} > g^c(\nu_m^c + 1)\}$ and $\mathbb{I}_m^P = \mathbb{I}\{\Delta_{P,[i_m^P,m+1]} > g^P(\nu_m^P + 1)\}$. Note that whenever **Test 3** fails in interval $m$, we have $\mathbb{I}_m^c = 1$ or $\mathbb{I}_m^P = 1$ by Lemma 23. We partition $K$ intervals into segments $\mathcal{I}_1, \ldots, \mathcal{I}_{N_c}$, such that in the last interval of each $\mathcal{I}_i$ with $i < N_c$ denoted by $m$, **Test 3** fails and $\mathbb{I}_m^c = 1$. Since $\nu^c$ is reset whenever **Test 3** fails, we have $\Delta_{\mathcal{I}_i\cup\{m+1\}} \geq \Delta_{[i_m^c,m+1]} > g^c(\nu_m^c + 1) \geq g^c(|\mathcal{I}_i| + 1)$. By Lemma 13, we obtain $N_c = \tilde{\mathcal{O}}((c_1^{-1}\Delta_c)^{2/3}K^{1/3} + H\Delta_c)$.

Now define $\mathbb{A}_m$ as the indicator that **Test 3** fails in interval $m$ and $\mathbb{I}_m^P = 1$. Also define $\mathbb{A}'_m$ as the indicator that **Test 3** fails and $\mathbf{N}$ is reset in interval $m$, and $\mathbb{I}_m^P = 1$. We then partition $K$ intervals into segments $\mathcal{I}'_1, \ldots, \mathcal{I}'_{N_P}$, such that in the last interval of each $\mathcal{I}'_i$ with $i < N_P$ denoted by $m$, $\mathbb{A}'_m = 1$. Since $\nu^P$ is reset in interval $m$ when $\mathbb{A}'_m = 1$, we have $\Delta_{\mathcal{I}'_i\cup\{m+1\}} \geq \Delta_{[i_m^P,m+1]} > g^P(\nu_m^P + 1) \geq g^P(|\mathcal{I}'_i|+1)$. By Lemma 13, we have $N_P = \tilde{\mathcal{O}}((c_2^{-1}\Delta_P)^{2/3}K^{1/3} + H\Delta_P)$. Moreover, by Lemma 50 and the reset rule of **Test 3**, we have $p\sum_m \mathbb{A}_m = \tilde{\mathcal{O}}(\sum_m \mathbb{A}'_m)$ with probability at least $1-\delta$, which gives $\sum_m \mathbb{A}_m = \tilde{\mathcal{O}}(N_P/p)$.

Since $\mathbb{I}_m^c = 1$ or $\mathbb{I}_m^P = 1$ when **Test 3** fails in interval $m$, the total number of times that **Test 3** fails $N_3 \leq N_c + \sum_m \mathbb{A}_m = \tilde{\mathcal{O}}((c_1^{-1}\Delta_c)^{2/3}K^{1/3} + B_\star(c_2^{-1}\Delta_P)^{2/3}K^{1/3} + H(\Delta_c + B_\star\Delta_P))$. Now by the reset rule of **Test 3**, the number of times $\mathbf{M}$ is reset due to **Test 3** is upper bounded by $N_3$, and the number of times $\mathbf{N}$ is reset due to **Test 3** is upper bounded by $\tilde{\mathcal{O}}(pN_3)$ with probability at least $1-\delta$ by Lemma 50. Finally, by Line 4 and Line 5 of Algorithm 4, there are at most $\frac{K}{W_c}$ resets of $\mathbf{M}$ and $\frac{K}{W_P}$ resets of $\mathbf{N}$ respectively due to periodic restarts. Putting all cases together, we have

$$\begin{aligned} L_c &= \tilde{\mathcal{O}}\left((c_1^{-1}\Delta_c)^{2/3}K^{1/3} + B_\star(c_2^{-1}\Delta_P)^{2/3})K^{1/3} + H(\Delta_c + B_\star\Delta_P) + K/W_c\right) \\ &= \tilde{\mathcal{O}}\left((B_\star SA)^{-1/3}(T_\star\Delta_c)^{2/3}K^{1/3} + B_\star(SA)^{-1/3}(T_\star\Delta_P)^{2/3}K^{1/3} + H(\Delta_c + B_\star\Delta_P)\right), \end{aligned}$$

and

$$\begin{aligned} L_P &= \tilde{\mathcal{O}}\left(\frac{1}{B_\star}(c_1^{-1}\Delta_c)^{2/3}K^{1/3} + (c_2^{-1}\Delta_P)^{2/3}K^{1/3} + H(\Delta_c + \Delta_P) + K/W_P\right) \\ &= \tilde{\mathcal{O}}\left(\frac{(B_\star SA)^{-1/3}(T_\star\Delta_c)^{2/3}K^{1/3}}{B_\star} + (SA)^{-1/3}(T_\star\Delta_P)^{2/3}K^{1/3} + H(\Delta_c + \Delta_P)\right). \end{aligned}$$

This completes the proof. $\square$

### E.2 Auxiliary Lemmas

**Lemma 27.** *With probability at least $1-\delta$, for any $M' \leq K$, $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m}) = \tilde{\mathcal{O}}\left(B_\star C_{M'} + B_\star M' + B_\star^2\right)$.*

*Proof.* Applying Lemma 9 with $\left\|\check{V}_h^{\star,m}\right\|_\infty \leq B$, with probability at least $1-\delta$,

$$\begin{aligned} &\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m}) \\ &= \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\check{V}_{H_m+1}^{\star,m}(s_{H_m+1}^m)^2 + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}B_\star(\check{V}_h^{\star,m}(s_h^m) - P_h^m\check{V}_{h+1}^{\star,m})_+ + B_\star^2\right) \\ &= \tilde{\mathcal{O}}\left(B_\star C_{M'} + B_\star M' + B_\star^2\right), \end{aligned}$$

where in the last step we apply

$$\begin{aligned} (\check{V}_h^{\star,m}(s_h^m) - P_h^m\check{V}_{h+1}^{\star,m})_+ &\leq (\check{Q}_h^{\star,m}(s_h^m, a_h^m) - P_h^m\check{V}_{h+1}^{\star,m})_+ \leq c^m(s_h^m, a_h^m) + 8\eta_m \\ &\leq c^m(s_h^m, a_h^m) + 1/H, \end{aligned}$$

and also Lemma 50. $\square$

**Lemma 28.** *With probability at least* $1-10\delta$, *for any* $M' \leq K$, $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m} - \check{V}_{h+1}^m) = \tilde{\mathcal{O}}(B_\star\sqrt{SAL_{c,M'}C_{M'}} + B_\star\sqrt{B_\star SAL_{P,M'}(C_{M'}+M')} + B_\star^2 S^2 AL_{P,M'} + B_\star^2 SAL_{c,M'} + \sum_{m=1}^{M'} B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H)$.

*Proof.* Let $z_h^m = \min\{B/2, (\Delta_{c,m} + B\Delta_{P,m})H\}\mathbb{I}\{h \leq H\}$. By Lemma 21, we have $\check{V}_h^{\star,m}(s) + z_h^m \geq \check{V}_h^m(s)$. Moreover, by Lemma 12, we have

$$\begin{aligned}
&\sum_{m=1}^{M'}(\check{V}_{H_m+1}^{\star,m}(s_{H_m+1}^m) + z_{H_m+1}^m - \check{V}_{H_m+1}^m(s_{H_m+1}^m))^2 \\
&\leq \sum_{m=1}^{M'}(z_{H_m+1}^m)^2\mathbb{I}\{s_{H_m+1}^m = g\} + 64B_\star^2\sum_{m=1}^{M'}\mathbb{I}\{H_m < H, s_{H_m+1}^m \neq g\} \\
&= \tilde{\mathcal{O}}\left(B_\star\sum_{m=1}^{M'}(\Delta_{c,m} + B_\star\Delta_{P,m})H + B_\star^2 SAL_{M'}\right).
\end{aligned}$$

and

$$\begin{aligned}
(*) &= \sum_{m=1}^{M'} B_\star\sum_{h=1}^{H_m}(\check{V}_h^{\star,m}(s_h^m) - \check{V}_h^m(s_h^m) - P_h^m\check{V}_{h+1}^{\star,m} + P_h^m\check{V}_{h+1}^m + z_h^m - z_{h+1}^m)_+ \\
&\leq \sum_{m=1}^{M'} B_\star\sum_{h=1}^{H_m}\left(c^m(s_h^m, a_h^m) + 8\eta_m + \widetilde{P}_h^m\check{V}_{h+1}^m - \check{V}_h^m(s_h^m) + B(\Delta_{P,m} + \mathbf{n}_h^m)\right)_+ + B_\star\sum_{m=1}^{M'}(z_1^m - z_{H_m+1}^m) \\
&\qquad (\check{V}_h^{\star,m}(s_h^m) \leq \check{Q}_h^{\star,m}(s_h^m, a_h^m),\ z_h^m \geq z_{h+1}^m, \text{ and } P_{h+1}^m\check{V}_{h+1}^m \leq \widetilde{P}_{h+1}^m\check{V}_{h+1}^m + B(\Delta_{P,m} + \mathbf{n}_h^m)) \\
&\leq \sum_{m=1}^{M'} B_\star\sum_{h=1}^{H_m}(c^m(s_h^m, a_h^m) - \widehat{c}_h^m + (\widetilde{P}_h^m - \bar{P}_h^m)\check{V}_{h+1}^{\star,m} + (\widetilde{P}_h^m - \bar{P}_h^m)(\check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m}) + b_h^m)_+ \\
&\quad + \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'} B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H + B_\star^2\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbf{n}_h^m\right). \qquad \text{(definition of } \check{V}_h^m(s_h^m))
\end{aligned}$$

Now by Lemma 3, Lemma 8, Lemma 6, and $\mathbf{n}_h^m \leq \frac{1}{\mathbf{N}_h^m}$, we continue with

$$\begin{aligned}
(*) &= \tilde{\mathcal{O}}\left(B_\star\left(\sqrt{SAL_{c,M'}C_{M'}} + SAL_{c,M'} + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m})}{\mathbf{N}_h^m}} + \sqrt{\frac{S\mathbb{V}(P_h^m, \check{V}_{h+1}^m - \check{V}_{h+1}^{\star,m})}{\mathbf{N}_h^m}}\right)\right)\right) \\
&\quad + \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B_\star^2 S}{\mathbf{N}_h^m} + \sum_{m=1}^{M'} B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H + B_\star\sum_{m=1}^{M'}\sum_{h=1}^{H_m} b_h^m\right) \\
&= \tilde{\mathcal{O}}\left(B_\star\sqrt{SAL_{c,M'}C_{M'}} + B_\star SAL_{c,M'} + B_\star\sqrt{SAL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m})} + B_\star^2 S^2 AL_{P,M'}\right) \\
&\quad + \tilde{\mathcal{O}}\left(B_\star\sqrt{S^2 AL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m} - \check{V}_{h+1}^m)} + \sum_{m=1}^{M'} B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H\right),
\end{aligned}$$

where in the last step we apply Cauchy-Schwarz inequality, Lemma 11, Lemma 10, $\textsc{Var}[X+Y] \leq 2\textsc{Var}[X] + 2\textsc{Var}[Y]$, and AM-GM inequality. Finally, by Lemma 27, we continue with

$$\begin{aligned}
(*) &= \tilde{\mathcal{O}}\left(B_\star\sqrt{SAL_{c,M'}C_{M'}} + B_\star SAL_{c,M'} + B_\star\sqrt{B_\star SAL_{P,M'}(C_{M'} + M')} + B_\star^2 S^2 AL_{P,M'}\right) \\
&\quad + \tilde{\mathcal{O}}\left(B_\star\sqrt{S^2 AL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m} - \check{V}_{h+1}^m)} + \sum_{m=1}^{M'} B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H\right).
\end{aligned}$$

Applying Lemma 9 on value functions $\{\check{V}_h^{\star,m} + z_h^m - \check{V}_h^m\}_{m,h}$ (constant offset does not change the variance) and plugging in the bounds above, we have

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m} - \check{V}_{h+1}^m) = \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m} + z_h^m - \check{V}_{h+1}^m)$$
$$= \tilde{\mathcal{O}}\left(B_\star\sqrt{SAL_{c,M'}C_{M'}} + B_\star^2 SAL_{c,M'} + B_\star\sqrt{B_\star SAL_{P,M'}(C_{M'}+M')} + B_\star^2 S^2 AL_{P,M'}\right)$$
$$+ \tilde{\mathcal{O}}\left(B_\star\sqrt{S^2AL_{P,M'}\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m} - \check{V}_{h+1}^m) + \sum_{m=1}^{M'} B_\star(\Delta_{c,m} + B_\star\Delta_{P,m})H}\right).$$

Then solving a quadratic inequality w.r.t $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^{\star,m} - \check{V}_{h+1}^m)$ (Lemma 45) completes the proof. □

### E.3 Proof of Theorem 5

We first prove a general regret guarantee of Algorithm 5, from which Theorem 5 is a direct corollary.

**Theorem 11.** *Suppose $\mathfrak{A}_1$ ensures $\mathring{R}_K \le R^1$ when $s_1^m = s_{\text{init}}$ for $m \le K$, and $\mathfrak{A}_2$ ensures $R_{K'} \le R^2(K')$ for any $K' \le K$ such that $R^2(k)$ is sub-linear w.r.t $k$. Then Algorithm 5 ensures $R_K = \tilde{\mathcal{O}}(R^1)$ (ignoring lower order terms).*

*Proof.* Let $\mathcal{I}_k$ be the set of intervals in episode $k$, and $m_i^k$ be the $i$-th interval of episode $k$ (if exists). The regret is decomposed as:

$$R_K = \sum_{k=1}^K\left[\sum_{h=1}^{H_{m_1^k}} c_h^{m_1^k} + c_{H_{m_1^k}+1}^{m_1^k} - V_k^\star(s_1^k)\right] + \sum_{k=1}^K\left[\sum_{m\in\mathcal{I}_k\setminus\{m_1^k\}}\sum_{h=1}^{H_m} c_h^m - c_{H_{m_1^k}+1}^{m_1^k}\right].$$

Note that $V_1^{\pi^\star,m_1^k}(s_1^{m_1^k}) \le V_k^\star(s_1^k) + B_\star/K$ by Lemma 46. Therefore,

$$\sum_{k=1}^K\left[\sum_{h=1}^{H_{m_1^k}} c_h^{m_1^k} + c_{H_{m_1^k}+1}^{m_1^k} - V_k^\star(s_1^k)\right] \le \sum_{k=1}^K\left[\sum_{h=1}^{H_{m_1^k}} c_h^{m_1^k} + c_{H_{m_1^k}+1}^{m_1^k} - V_1^{\pi^\star,m_1^k}(s_1^{m_1^k})\right] + B_\star$$
$$\le R^1 + B_\star.$$

For the second term, note that $c_{H_{m_1^k}+1}^{m_1^k} = 2B_\star$ if $s_1^{m_2^k}$ exists. Define $K_f = \sum_{k=1}^K \mathbb{I}\{|\mathcal{I}_k| > 1\}$, we have (define $s_1^{m_2^k} = g$ if $m_2^k$ does not exist)

$$\sum_{k=1}^K\left[\sum_{m\in\mathcal{I}_k\setminus\{m_1^k\}}\sum_{h=1}^{H_m} c_h^m - c_{H_{m_1^k}+1}^{m_1^k}\right] \le \sum_{k=1}^K\left(\sum_{m\in\mathcal{I}_k\setminus\{m_1^k\}}\sum_{h=1}^{H_m} c_h^m - V_k^\star(s_1^{m_2^k})\right) - B_\star K_f$$
$$\le R^2(K_f) - B_\star K_f,$$

which is a lower order term since $R^2(K_f)$ is sub-linear w.r.t $K_f$. Putting everything together completes the proof. □

We are now ready to prove Theorem 5.

*Proof.* We simply apply Theorem 11 with $R^1$ determined by Theorem 4 and $R^2$ determined by Theorem 3. □

**Algorithm 6** MVP-Base

**Parameters:** failure probability $\delta$.
**Initialize:** $\widehat{\chi} \leftarrow 0$, and for all $(s, a, s')$, $\mathbf{C}(s, a) \leftarrow 0$, $\mathbf{M}(s, a) \leftarrow 0$, $\mathbf{N}(s, a) \leftarrow 0$, $\mathbf{N}(s, a, s') \leftarrow 0$.
**Initialize:** `Update(1)`.
**for** $m = 1, \ldots, M$ **do**
  **for** $h = 1, \ldots, H$ **do**
    Play action $a_h^m = \mathrm{argmin}_a\, \check{Q}_h(s_h^m, a)$, receive cost $c_h^m$ and next state $s_{h+1}^m$.
    $\mathbf{C}(s_h^m, a_h^m) \leftarrow c_h^m$, $\mathbf{M}(s_h^m, a_h^m) \xleftarrow{+} 1$, $\mathbf{N}(s_h^m, a_h^m) \xleftarrow{+} 1$, $\mathbf{N}(s_h^m, a_h^m, s_{h+1}^m) \xleftarrow{+} 1$.
    **if** $s_{h+1}^m = g$ *or* $\mathbf{M}(s_h^m, a_h^m) = 2^l$ *or* $\mathbf{N}(s_h^m, a_h^m) = 2^l$ *for some integer* $l \geq 0$ **then**
      **break** (which starts a new interval).
  $\widehat{\chi} \xleftarrow{+} C^m - \check{V}_1(s_1^m)$.
1 **if** $\widehat{\chi} > \chi_m$ *(defined in Lemma 31)* **then** terminate. **(Test 1)**
  `Update(`$m+1$`)`.
2 **if** $\|\check{V}_h\| > B/2$ *for some* $h$ ***(Test 2)*** **then** terminate.

**Procedure** `Update(`$m$`)`
  $\check{V}_{H+1}(s) \leftarrow 2B_\star \mathbb{I}\{s \neq g\}$, $\check{V}_h(g) \leftarrow 0$ for all $h \leq H$, and $\iota \leftarrow 2^{11} \cdot \ln\left(\frac{2SAHKm}{\delta}\right)$.
3 $\eta \leftarrow \min\{\frac{B_\star S\sqrt{A}}{T_\star\sqrt{m}}, \frac{1}{2^8 H}\}$.
  **for** *all* $(s, a)$ **do**
    $\mathbf{N}^+(s, a) \leftarrow \max\{1, \mathbf{N}(s, a)\}$, $\mathbf{M}^+(s, a) \leftarrow \max\{1, \mathbf{M}(s, a)\}$, $\bar{c}(s, a) \leftarrow \frac{\mathbf{C}(s,a)}{\mathbf{M}^+(s,a)}$,
    $\bar{P}_{s,a}(\cdot) \leftarrow \frac{\mathbf{N}(s,a,\cdot)}{\mathbf{N}^+(s,a)}$, $\widehat{c}(s, a) \leftarrow \max\left\{0, \bar{c}(s, a) - \sqrt{\frac{\bar{c}(s,a)\iota}{\mathbf{M}^+(s,a)}} - \frac{\iota}{\mathbf{M}^+(s,a)}\right\}$,
4   $\check{c}(s, a) \leftarrow \widehat{c}(s, a) + 8\eta$.
  **for** $h = H, \ldots, 1$ **do**
    $b_h(s, a) \leftarrow \max\left\{7\sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, \check{V}_{h+1})\iota}{\mathbf{N}^+(s,a)}}, \frac{49B\sqrt{S}\iota}{\mathbf{N}^+(s,a)}\right\}$ for all $(s, a)$.
    $\check{Q}_h(s, a) = \max\{0, \check{c}(s, a) + \bar{P}_{s,a}\check{V}_{h+1} - b_h(s, a)\}$ all $(s, a)$.
    $\check{V}_h(s) = \mathrm{argmin}_a\, \check{Q}_h(s, a)$ for all $s$.

# F Omitted Details in Section 7

In this section, we present all proofs and details of learning without the knowledge of non-stationarity. We first provide a base algorithm in Appendix F.1. The rest of this section then discusses the meta algorithm MASTER adopted from [Wei and Luo, 2021], and its regret guarantee combining with the base algorithm.

## F.1 Base Algorithm

We first present the base algorithm used in MASTER (Algorithm 6). The main idea is again incorporating a correction term to penalize long horizon policy and has the effect of cancelling the non-stationarity along the learner's trajectory when it is not too large (Line 3). When the non-stationarity is large, on the other hand, we detect it through two non-stationary tests (Line 1 and Line 2), and reset the knowledge of the environment (more details to follow).

**Test 1** is a combination of the first two tests of Algorithm 4, which directly checks whether the estimated regret is too large. This is also similar to the second test of the MASTER algorithm [Wei and Luo, 2021]. **Test 2** is the same as the third test of Algorithm 4, which guards the magnitude of the estimated value function. When tests fail, the algorithm directly terminate instead of resetting some accumulators. Note that the status of $\mathbf{M}$ and $\mathbf{N}$ are completely identical in this algorithm, but we still maintain them separately so that the auxiliary lemmas in Appendix A are still applicable. The rest of the algorithm largely follows the design of Algorithm 2.

**Notations** Note that here $\mathbf{M}$ and $\mathbf{N}$ are only reset at the initialization step. Thus, $i_m^c = i_m^P = 1$, $L_{c,m} = L_{P,m} = 1$, $\Delta_{c,m} = \Delta_{c,[1,m]}$ and $\Delta_{P,m} = \Delta_{P,[1,m]}$. Let $\Delta'_m = (\Delta_{c,m} + B\Delta_{P,m})$ and denote by $\eta_m$, $\check{Q}_h^m$, $\check{V}_h^m$ the value of $\eta$, $\check{Q}_h$, and $\check{V}_h$ at the beginning of interval $m$. Denote by $\check{c}^m$ the value of $\check{c}$ at the beginning of interval $m$ and define $\check{c}_h^m = \check{c}(s_h^m, a_h^m)$. Also define $\check{Q}_h^{\pi^\star,m}$ and $\check{V}_h^{\pi^\star,m}$ as the action-value function and value function w.r.t cost $c^m(s,a) + 8\eta_m$, transition $P^m$, and policy $\pi^\star_{k(m)}$.

**Lemma 29.** *With probability at least $1-2\delta$, if Algorithm 6 does not terminate up to interval $m \leq K$, then $\check{Q}_h^m(s,a) \leq \check{Q}_h^{\pi^\star,m}(s,a) + \Delta'_m T_h^{\pi^\star,m}(s,a)$.*

*Proof.* We prove this by induction on $h$. The base case of $h = H+1$ is clearly true. For $h \leq H$, by **Test 2** and the induction step, we have $\check{V}_{h+1}^m(s) \leq \min\{B/2, \check{V}_{h+1}^{\pi^\star,m}(s) + \Delta'_m T_{h+1}^{\pi^\star,m}(s)\} \leq \check{V}_{h+1}^{\pi^\star,m}(s) + x_{h+1}^m(s) \leq B$ where $x_h^m(s) = \min\{B/2, \Delta'_m T_h^{\pi^\star,m}(s)\}$. Thus,

$$\begin{aligned}
&\check{c}^m(s,a) + \bar{P}_{s,a}^m \check{V}_{h+1}^m - b^m(s,a,\check{V}_{h+1}^m)\\
&\leq \check{c}^m(s,a) + \bar{P}_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + x_{h+1}^m) - b^m(s,a,\check{V}_{h+1}^{\pi^\star,m} + x_{h+1}^m) && \text{(Lemma 48)}\\
&\overset{\text{(i)}}{\leq} \check{c}^m(s,a) + \widetilde{P}_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + x_{h+1}^m) && \text{(Lemma 8)}\\
&\leq c^m(s,a) + 8\eta_m + \Delta_{c,m} + P_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + x_{h+1}^m) + \Delta_{P,m}B && \text{(Lemma 5)}\\
&\leq \check{Q}_h^{\pi^\star,m}(s,a) + \Delta'_m T_h^{\pi^\star,m}(s,a).
\end{aligned}$$

Note that in (i) we use the fact that $|\{\check{V}_h^{\pi^\star,m} + x_h^m\}_{m,h}| \leq (HK+1)^6$ since $|\{V_h^{\pi^\star,m}\}_{m,h}| \leq HK+1$, $|\{\eta_m\}_m| \leq K+1$, $|\{\Delta'_m\}_m| \leq K+1$, and $|\{T_h^{\pi^\star,m}\}_{m,h}| \leq HK+1$. $\square$

**Lemma 30.** *With probability at least $1-2\delta$, for all $m \leq K$, if $\Delta'_m \leq \eta_m$, then $\check{Q}_h^m(s,a) \leq \check{Q}_h^{\pi^\star,m}(s,a) + \eta_m T_h^{\pi^\star,m}(s,a) \leq B/2$. Moreover, if **Test 2** fails in interval $m$, then $\Delta'_{m+1} > \eta_{m+1}$.*

*Proof.* First note that $\check{Q}_h^{\pi^\star,m}(s,a) \leq \frac{B}{4} + 8\eta_m T_h^{\pi^\star,m}(s,a) \leq \frac{B}{4} + 8H\eta_m \leq \frac{B}{3}$. We prove the first statement by induction on $h$. The base case of $h = H+1$ is clearly true. For $h \leq H$, note that:

$$\begin{aligned}
&\check{c}^m(s,a) + \bar{P}_{s,a}^m \check{V}_{h+1}^m - b^m(s,a,\check{V}_{h+1}^m)\\
&\leq \check{c}^m(s,a) + \bar{P}_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + \eta_m T_{h+1}^{\pi^\star,m}) - b^m(s,a,\check{V}_{h+1}^{\pi^\star,m} + \eta_m T_{h+1}^{\pi^\star,m})\\
& && \text{(induction step and Lemma 48)}\\
&\overset{\text{(i)}}{\leq} \check{c}^m(s,a) + \widetilde{P}_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + \eta_m T_{h+1}^{\pi^\star,m}) && \text{(Lemma 8)}\\
&\leq c^m(s,a) + 8\eta_m + \Delta_{c,m} + P_{s,a}^m(\check{V}_{h+1}^{\pi^\star,m} + \eta_m T_{h+1}^{\pi^\star,m}) + \Delta_{P,m}(B/3 + H\eta_m) && \text{(Lemma 5)}\\
&\leq \check{Q}_h^{\pi^\star,m}(s,a) + \eta_m T_h^{\pi^\star,m}(s,a). && (H\eta_m \leq B/12 \text{ and } \Delta'_m \leq \eta_m)
\end{aligned}$$

Note that in (i) we use the fact that $|\{\check{V}_h^{\pi^\star,m} + \eta_m T_h^{\pi^\star,m}\}_{m,h}| \leq (HK+1)^6$ since $|\{V_h^{\pi^\star,m}\}_{m,h}| \leq HK+1$, $|\{\eta_m\}_m| \leq K+1$, and $|\{T_h^{\pi^\star,m}\}_{m,h}| \leq HK+1$. The second statement is simply by the contraposition of the first statement. $\square$

**Lemma 31.** *With probability at least $1-12\delta$, for any $M' \leq K$, if $\Delta'_{M'} \leq \eta_{M'}$, then*

$$\sum_{m=1}^{M'}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - \check{V}_1^m(s_1^m)\right) = \tilde{\mathcal{O}}\left(B_\star S\sqrt{AM'} + B_\star S^2 A\right) \triangleq \chi_{M'}.$$

*Moreover, if **Test 1** fails in interval $m$, then $\Delta'_m > \eta_m$.*

*Proof.* By $\Delta'_{M'} \leq \eta_{M'}$ and Lemma 30, the algorithm will not terminate by **Test 2** before interval $M'$ with probability at least $1-2\delta$. Then with probability at least $1-4\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - \check{V}_1^m(s_1^m)\right)\\
&\leq \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(c_h^m + \check{V}_{h+1}^m(s_{h+1}^m) - \check{V}_h^m(s_h^m)\right) + \tilde{\mathcal{O}}\left(B_\star SA\right) \qquad (\text{Lemma 12 and } L_{M'}=\mathcal{O}(1))\\
&\leq \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(c_h^m - \widehat{c}_h^m + \check{V}_{h+1}^m(s_{h+1}^m) - P_h^m\check{V}_{h+1}^m + (P_h^m - \bar{P}_h^m)\check{V}_{h+1}^m + b_h^m - 8\eta_m\right) + \tilde{\mathcal{O}}\left(B_\star SA\right)\\
&\qquad\qquad (\text{definition of } \check{V}_h^m(s_h^m))\\
&\leq \tilde{\mathcal{O}}\left(\sqrt{SAC_{M'}} + \sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^m) + B_\star SA}\right)\\
&\qquad + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left((\widetilde{P}_h^m - \bar{P}_h^m)\check{V}_{h+1}^m + B\mathbf{n}_h^m + b_h^m - 5\eta_m\right),
\end{aligned}$$

where in the last inequality we apply Lemma 3, $i^c_{M'} = i^P_{M'} = 1$, $\Delta'_{M'} \leq \eta_{M'}$, $P_h^m\check{V}_{h+1}^m \leq \widetilde{P}_h^m\check{V}_{h+1}^m + B(\mathbf{n}_h^m + \Delta_{P,m})$, Lemma 49 and Lemma 50 on both $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c_h^m - c^m(s_h^m, a_h^m))$, and Lemma 49 on $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\check{V}_{h+1}^m(s_{h+1}^m) - P_h^m\check{V}_{h+1}^m)$. Now note that with probability at least $1-6\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}((\widetilde{P}_h^m - \bar{P}_h^m)\check{V}_{h+1}^m + b_h^m + B\mathbf{n}_h^m) + \tilde{\mathcal{O}}\left(\sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^m)}\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^m) + B_\star S^2A}\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m} b_h^m + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B\Delta_{P,m}}{64}\\
&\qquad (\mathbf{n}_h^m \leq \tfrac{1}{\mathbf{N}_h^m}, \text{Lemma 6, Cauchy-Schwarz inequality, Lemma 11, and } L_{P,M'}=1)\\
&= \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, \check{V}_{h+1}^m) + B_\star S^2A}\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{B\Delta_{P,m}}{32}.\\
&\qquad (\text{Lemma 10, } L_{P,M'}=1, \text{and AM-GM inequality})\\
&= \tilde{\mathcal{O}}\left(\sqrt{B_\star S^2A(C_{M'} + M')} + B_\star S^2A\right) + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}\frac{\Delta'_m}{16}. \quad (\text{Lemma 32 and AM-GM inequality})
\end{aligned}$$

Plugging this back and by $\Delta'_{M'} \leq \eta_{M'}$, we have

$$C_{M'} - \sum_{m=1}^{M'}\check{V}_1^m(s_1^m) = \tilde{\mathcal{O}}\left(\sqrt{B_\star S^2A(C_{M'} + M')} + B_\star S^2A\right).$$

Solving a quadratic inequality w.r.t $C_{M'}$ (Lemma 45), we have $C_{M'} = \tilde{\mathcal{O}}(B_\star M' + \sqrt{B_\star S^2AM'} + B_\star S^2A)$. Plugging this back completes the proof of the first statement. The second statement is simply by the contraposition of the first statement. $\square$

**Theorem 12.** *Suppose Algorithm 4 does not terminate up to interval $M' \leq K$ (including $M'$) and $s_1^m = s_{\text{init}}$ for $m \leq M'$. Then with probability at least $1-2\delta$, $\mathring{R}_{M'} = \tilde{\mathcal{O}}(B_\star S\sqrt{AM'} + B_\star S^2A + \sum_{m=1}^{M'}\Delta'_m T_\star)$.*

*Proof.* We decompose the regret as follows:

$$\begin{aligned}
\mathring{R}_{M'} &= \sum_{m=1}^{M'}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - V_1^{\pi^\star,m}(s_1^m)\right)\\
&= \sum_{m=1}^{M'}\left(\sum_{h=1}^{H_m} c_h^m + c_{H_m+1}^m - \check{V}_1^m(s_1^m)\right) + \sum_{m=1}^{M'}\left(\check{V}_1^m(s_1^m) - \check{V}_1^{\pi^\star,m}(s_1^m)\right) + 8T_\star\sum_{m=1}^{M'}\eta_m\\
&\leq \chi_{M'} + \sum_{m=1}^{M'}\Delta'_m T_\star + 8T_\star\sum_{m=1}^{M'}\eta_m. \qquad \text{(**Test 2** and Lemma 29)}
\end{aligned}$$

Plugging in the definition of $\chi_{M'}$ and $\eta_m$ completes the proof. $\square$

**Lemma 32.** *With probability at least $1-4\delta$, $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,\check{V}_{h+1}^m) = \tilde{\mathcal{O}}(B_\star(C_{M'}+M') + B_\star^2S^2A + B_\star\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta'_m)$ for any $M'\leq K$.*

*Proof.* Applying Lemma 9 with $\left\|\check{V}_h^m\right\|_\infty \leq B$ (**Test 2**), with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,\check{V}_{h+1}^m)\\
&= \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\check{V}_{H_m+1}^m(s_{H_m+1}^m)^2 + \sum_{m=1}^{M'}\sum_{h=1}^{H_m}B_\star(\check{V}_h^m(s_h^m) - P_h^m\check{V}_{h+1}^m)_+ + B_\star^2\right)\\
&= \tilde{\mathcal{O}}\left(B_\star(C_{M'}+M') + B_\star\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,\check{V}_{h+1}^m)} + B_\star^2S^2A + B_\star\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta'_m\right),
\end{aligned}$$

where in the last step we apply

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\check{V}_h^m(s_h^m) - P_h^m\check{V}_{h+1}^m)_+ = \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\check{Q}_h^m(s_h^m,a_h^m) - P_h^m\check{V}_{h+1}^m)_+\\
&\leq \sum_{m=1}^{M'}\sum_{h=1}^{H_m}(\check{c}_h^m + (\bar{P}_h^m - \widetilde{P}_h^m)\check{V}_{h+1}^m + B\Delta_{P,m})_+\\
&\qquad\qquad ((a)_+ - (b)_+ \leq (a-b)_+\text{, definition of } \check{Q}_h^m\text{, and } b_h^m \geq 0)\\
&\leq \sum_{m=1}^{M'}\sum_{h=1}^{H_m}c^m(s_h^m,a_h^m) + M' + \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,\check{V}_{h+1}^m)} + B_\star S^2A\right) + 2\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta'_m\\
&\qquad\qquad \text{(Lemma 5, } 8\eta_m \leq \tfrac{1}{H}\text{, Lemma 6, Cauchy-Schwarz inequality, and Lemma 11)}\\
&\leq \tilde{\mathcal{O}}\left(\sum_{m=1}^{M'}\sum_{h=1}^{H_m}c_h^m + M' + \sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,\check{V}_{h+1}^m)} + B_\star S^2A\right) + 2\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\Delta'_m.\\
&\qquad\qquad \text{(Lemma 50)}
\end{aligned}$$

Solving a quadratic inequality w.r.t $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,\check{V}_{h+1}^m)$ (Lemma 45) completes the proof. $\square$

### F.2 Preliminaries

Here we adopt the MASTER algorithm in [Wei and Luo, 2021] to our finite-horizon approximation scheme. There are several issues we need to address: 1) under the protocol of Algorithm 1, the total number of intervals and the non-stationarity in each interval are not fixed before learning start; besides, we need to prove an anytime regret guarantee, so that it can translate back to a regret guarantee on the original SSP (see Lemma 16); 2) when the base algorithm has a regret guarantee $\mathring{R}_m \leq \min\{c_1\sqrt{m}+c_2, c_3 m\}$ without non-stationarity, the original MASTER algorithm ensures a dynamic regret whose dominating term scale with $c_1 + c_2c_3/c_1$; this is undesirable as $c_3 = \tilde{\mathcal{O}}(T_{\max})$ in our case, and ideally we want $c_3 = \tilde{\mathcal{O}}(B_\star)$; 3) when base algorithms incorporate correction term, the original analysis of the non-stationarity tests breaks as discussed in Section 7; 4) The analysis in [Wei and Luo, 2021, Lemma 17] that bounds the cost of non-stationary detection only works for oblivious adversary. Our modified MASTER algorithm (Algorithm 8) manages to address all these issues.

**Setup** To give a general result, we define the dynamic regret for the first $M'$ intervals as $\widetilde{R}_{M'} = \sum_{m=1}^{M'}(C^m - f^\star_m)$, where the choice of benchmark $\{f^\star_m\}_{m=1}^{M'}$ is flexible depending on the problem and the algorithm.

**Notations** For any interval $\mathcal{I}=[s,e]$, define $\Delta_{\mathcal{I}} = \sum_{m=s}^{e-1}\Delta(m)$ and $L_{\mathcal{I}} = 1+\sum_{m=s}^{e-1}\mathbb{I}\{\Delta(m)\neq 0\}$, where $\Delta(m)\in\mathbb{R}_+^{\mathbb{N}_+}$ is some non-stationarity measure satisfying $|f^\star_{m+1}-f^\star_m|\leq\Delta(m)$.

We make the following assumption on the base algorithm used in the MASTER algorithm, and then show two algorithms satisfying the assumption.

**Assumption 1.** *Base algorithm $\mathfrak{A}$ with failure probability $\delta$ on intervals $[1,M']$ outputs an estimate $\widetilde{f}_m$ at the beginning of interval $m\leq M'$ if it does not terminate before interval $m$. Moreover, there exists a non-decreasing function $R(m)=\min\{c_1\sqrt{m}+c_2, c_3m\}$ with $c_3\geq 1$ and non-stationarity measure $\Delta$ such that $r(m)=R(m)/m$ is non-increasing, $r(m)\geq\frac{1}{\sqrt{m}}$, $\widetilde{f}_m\leq c_4\leq c_3$ for all $m$, and with probability at least $1-\delta$, for any $m\leq M'$, as long as $\Delta_{[1,m]}\leq r(m)$ and $\mathfrak{A}$ does not terminate up to interval $m$ (including $m$), without knowing $\Delta_{[1,m]}$ we have:*

$$\widetilde{f}_m \leq f^\star_m + r(m),\quad \sum_{\tau=1}^m\left(C^\tau - \widetilde{f}_\tau\right)\leq R(m),\ \text{and}\quad \sum_{\tau=1}^m(f^\star_\tau - C^\tau)\leq R(m).$$

**Lemma 33.** *Algorithm 2 with arbitrary initial state for each interval satisfies Assumption 1 with $f^\star_m = V_1^{\star,m}(s_1^m)$, $\widetilde{f}_m = V_1^m(s_1^m)$, $\Delta(m)=\tilde{\mathcal{O}}((\Delta_{c,[m,m+1]}+B\Delta_{P,[m,m+1]})H)$, $R(m) = \tilde{\mathcal{O}}(\min\{B_\star S\sqrt{Am}+B_\star S^2A, Hm\})$, and $c_4=\tilde{\mathcal{O}}(B_\star)$.*

*Proof.* The first two properties are simply by Lemma 18 and Theorem 9 with $L_{c,m}=L_{P,m}=1$ and $\Delta_{[1,m]}\leq r(m)$ with a large enough constant hidden in $\tilde{\mathcal{O}}(\cdot)$ in the definition of $\Delta(m)$. For the third property, with high probability,

$$\sum_{\tau=1}^m(f^\star_\tau - C^\tau) = \sum_{\tau=1}^m(V_1^{\star,\tau}(s_1^m)-C^\tau) = \tilde{\mathcal{O}}\left(\sqrt{B_\star\sum_{\tau=1}^m C^\tau}+B_\star\right). \qquad \text{(Lemma 35)}$$

$$= \tilde{\mathcal{O}}\left(\sqrt{B_\star\left(\sum_{\tau=1}^m\widetilde{f}_\tau + R(m)\right)}+B_\star\right)\leq\tilde{\mathcal{O}}\left(B_\star\sqrt{m}+B_\star\right)+\frac{1}{2}R(m).$$

(the second property, $V_1^m(s_1^m)=\tilde{\mathcal{O}}(B_\star)$, and AM-GM inequality)

Plugging in the definition of $R(m)$ completes the proof (again with a large enough constant hidden in $\tilde{\mathcal{O}}(\cdot)$ in the definition of $R(m)$). □

**Lemma 34.** *Algorithm 6 with $m\leq K$ and $s_1^m = s_{init}$ satisfies Assumption 1 with $f^\star_m = V_1^{\pi^\star,m}(s_{init})$, $\widetilde{f}_m = \check{V}_1^m(s_1^m)$, $\Delta(m)=\tilde{\mathcal{O}}((\Delta_{c,[m,m+1]}+B\Delta_{P,[m,m+1]})T_\star)$, $R(m)=\tilde{\mathcal{O}}(\min\{B_\star S\sqrt{Am}+B_\star S^2A, Hm\})$, and $c_4=\tilde{\mathcal{O}}(B_\star)$.*

*Proof.* For the first property, by Lemma 29, $\Delta_{[1,m]}\leq r(m)$ and a large enough constant hidden in $\tilde{\mathcal{O}}(\cdot)$ in the definition of $\Delta(m)$, we have

$$\widetilde{f}_m = \check{V}_1^m(s_1^m)\leq\check{V}_1^{\pi^\star,m}(s_1^m)+\Delta'_mT_\star\leq f^\star_m + 8T_\star\eta_m+\tilde{\mathcal{O}}\left(\Delta_{[1,m]}\right)\leq f^\star_m + r(m).$$

**Algorithm 7** MALG

**Input:** order $n$, regret density function $r$.

**for** $l = 0, \ldots, n$ **do**

  **for** $m \in \{0, 2^l, 2 \cdot 2^l, \ldots, 2^n - 2^l\}$ **do**

    With probability $\frac{r(2^n)}{r(2^m)}$, assigns a new base algorithm on intervals $[m+1, m+2^l]$.

**for** *each interval* $m$ **do**

  Let $\mathfrak{A}$ be the algorithm that covers interval $m$ with shortest scheduled length, output $\widetilde{g}_m = \widetilde{f}_m^{\mathfrak{A}}$ (which is the $\widetilde{f}_m$ output by $\mathfrak{A}$), follow $\mathfrak{A}$'s decision, and update $\mathfrak{A}$ with environment's feedback.

  **if** $\mathfrak{A}$ *terminates* **then** terminate.

The second property is simply by **Test 2** (Lemma 31) of Algorithm 6 (again with a large enough constant hidden in $\tilde{\mathcal{O}}(\cdot)$ in the definition of $R(m)$). For the third property,

$$\sum_{\tau=1}^{m}(f_\tau^\star - C^\tau) = \sum_{\tau=1}^{m}(V_1^{\pi^\star,\tau}(s_{\text{init}}) - C^\tau) \leq \frac{B_\star m}{K} + \sum_{\tau=1}^{m}(V_1^{\star,\tau}(s_{\text{init}}) - C^\tau) \leq R(m),$$

where the first inequality is by Lemma 46 and the last step follows similar arguments as in Lemma 33. $\square$

**Lemma 35.** *With probability at least* $1 - 3\delta$, *for any* $m \leq M$, $\sum_{\tau=1}^{m}(V_1^{\star,\tau}(s_1^\tau) - C^\tau) = \tilde{\mathcal{O}}(\sqrt{B_\star \sum_{\tau=1}^{m} C^\tau} + B_\star)$.

*Proof.* With probability at least $1 - 3\delta$,

$$\begin{aligned}
\sum_{\tau=1}^{m}(V_1^{\star,\tau}(s_1^m) - C^\tau) &\leq \sum_{\tau=1}^{m}\sum_{h=1}^{H_\tau}(V_h^{\star,\tau}(s_h^\tau) - V_{h+1}^{\star,\tau}(s_{h+1}^\tau) - c_h^\tau) && (V_{H_\tau+1}^{\star,\tau}(s_{H_\tau+1}^\tau) \leq c_{H_\tau+1}^\tau) \\
&\leq \sum_{\tau=1}^{m}\sum_{h=1}^{H_\tau}(P_h^\tau V_{h+1}^{\star,\tau} - V_{h+1}^{\star,\tau}(s_{h+1}^\tau)) = \tilde{\mathcal{O}}\left(\sqrt{\sum_{\tau=1}^{m}\sum_{h=1}^{H_\tau}\mathbb{V}(P_h^\tau, V_{h+1}^{\star,\tau})} + B_\star\right) \\
& && (V_h^{\star,\tau}(s_h^\tau) \leq Q_h^{\star,\tau}(s_h^\tau, a_h^\tau) \text{ and Lemma 49}) \\
&= \tilde{\mathcal{O}}\left(\sqrt{B_\star \sum_{\tau=1}^{m} C^\tau} + B_\star\right). && (\text{Lemma 19})
\end{aligned}$$

$\square$

### F.3 MALG: Multi-Scale Learning with Base Algorithm

Following [Wei and Luo, 2021, Section 3], we first introduce MALG (Algorithm 7), which runs multiple instances of base algorithms in a multi-scale manner. We then combine MALG with non-stationarity detection to obtain the MASTER algorithm in Appendix F.4. We always run MALG on a segment (an interval of intervals) of length $2^n$ for some integer $n$, which we call a *block*. Since we want to obtain an anytime regret guarantee, the failure probability of base algorithms and MALG need to be adjusted adaptively. Specifically, if an MALG instance is scheduled on intervals $[M_\dagger - 2^n + 1, M_\dagger]$, then the regret guarantee of this MALG instance and the failure probability of base algorithms it maintains depends on $M_\dagger$. However, we ignore the dependency on $M_\dagger$ in algorithms and analysis since the regret bound only has logarithmic dependency on $M_\dagger$.

We show that MALG ensures a multi-scale regret guarantee in the following lemma. Below we say an algorithm is of order $l$ if it is scheduled on a segment of length $2^l$. Also denote by $\widetilde{f}_m^{\mathfrak{A}}$ the $\widetilde{f}_m$ output by $\mathfrak{A}$.

**Lemma 36.** *For a given* $M_\dagger \geq 1$, *let* $\widehat{n} = \log_2 M_\dagger + 1$ *and* $\widehat{R}(m) = 2^{10}\widehat{n}\ln(2M_\dagger/\delta)R(m)$. *Algorithm 7 scheduled on* $[M_\dagger - 2^n + 1, M_\dagger]$ *with input* $n \leq \log_2 M_\dagger$ *guarantees for any* $\mathfrak{A}$ *it maintains and any* $m \in [\mathfrak{A}.s, \mathfrak{A}.e]$, *as long as* $\Delta_{[\mathfrak{A}.s,m]} \leq r(m')$ *where* $m' = m - \mathfrak{A}.s + 1$ *and all* *base algorithms it maintains do not terminate up to interval $m$ (including $m$), we have with high probability:*

$$\widetilde{g}_m \leq f_m^\star + r(m''), \quad \sum_{\tau=\mathfrak{A}.s}^{m} (C^\tau - \widetilde{g}_\tau) \leq \widehat{R}(m'), \quad \textit{and} \quad \sum_{\tau=\mathfrak{A}.s}^{m} (f_\tau^\star - C^\tau) \leq \widehat{R}(m'),$$

*where $m''$ is the number of intervals that $\mathfrak{A}'$ is active up to interval $m$, and $\mathfrak{A}'$ is the active algorithm in interval $m$.*

*Proof.* Fix a base algorithm $\mathfrak{A}$ and $m \in [\mathfrak{A}.s, \mathfrak{A}.e]$. Suppose $\mathfrak{A}'$ is active in interval $m$, which implies $[\mathfrak{A}'.s, \mathfrak{A}'.e] \subseteq [\mathfrak{A}.s, \mathfrak{A}.e]$. For the first statement, note that $\Delta_{[\mathfrak{A}'.s,m]} \leq \Delta_{[\mathfrak{A}.s,m]} \leq r(m') \leq r(m'')$ since $r$ is non-increasing. Thus, by the guarantee of $\mathfrak{A}'$ (Assumption 1), we have

$$\widetilde{g}_m \leq f_m^\star + r(m'').$$

For the second statement, first note that:

$$\sum_{\tau=\mathfrak{A}.s}^{m} (C^\tau - \widetilde{g}_\tau) = \sum_{l=0}^{n} \sum_{\mathfrak{A}' \in \mathcal{S}_l} \sum_{\tau=\mathfrak{A}.s}^{m} (C^\tau - \widetilde{f}_\tau^{\mathfrak{A}'}) \mathbb{I}\{\mathfrak{A}' \text{ is active at } \tau\},$$

where $\mathcal{S}_l$ is the set of base algorithms of order $l$ which starts within $[\mathfrak{A}.s, m]$. For a fix $l$, suppose $\mathcal{S}_l = \{\mathfrak{A}'_1, \ldots, \mathfrak{A}'_N\}$, and define $\mathcal{I}_i = [\mathfrak{A}.s, m] \cap [\mathfrak{A}'_i.s, \mathfrak{A}'_i.e]$. Note that $\{\mathcal{I}_i\}_{i=1}^N$ are disjoint, and $\Delta_{\mathcal{I}_i} \leq \Delta_{[\mathfrak{A}.s,m]} \leq r(m') \leq r(|\mathcal{I}_i|)$. Moreover, $[\mathfrak{A}'_i.s, \mathfrak{A}'_i.e] \subseteq [\mathfrak{A}.s, \mathfrak{A}.e]$ if $\mathfrak{A}'_i$ is active at some interval within $[\mathfrak{A}.s, m]$. Therefore, by the the guarantee of $\mathfrak{A}'_i$ (Assumption 1) we have:

$$\sum_{i=1}^{N} \sum_{\tau=\mathfrak{A}.s}^{m} (C^\tau - \widetilde{f}_\tau^{\mathfrak{A}'_i}) \mathbb{I}\{\mathfrak{A}'_i \text{ is active at } \tau\} \leq \sum_{i=1}^{N} R(|\mathcal{I}_i|) \leq N \cdot R(\min\{2^l, m'\}).$$

Now we need to bound $N$. Note that $\mathbb{E}[N] \leq \frac{r(2^n)}{r(2^l)}(\frac{m'}{2^l} + 1)$ by the scheduling rule. By Lemma 50, with probability at least $1 - \frac{\delta}{(2M_\dagger)^6}$ (simply choose a small enough failure probability such that the failure probability over all $M_\dagger \geq 1$ and all base algorithms is bounded), $N \leq 2\mathbb{E}[N] + 2^8 \ln(2M_\dagger/\delta) \leq \frac{2r(2^n)}{r(2^l)} \frac{m'}{2^l} + 258 \ln(2M_\dagger/\delta)$ and

$$\begin{aligned} N \cdot R(\min\{2^l, m'\}) &\leq \left(\frac{2r(2^n)}{r(2^l)} \frac{m'}{2^l} + 258 \ln(2M_\dagger/\delta)\right) R(\min\{2^l, m'\}) \\ &\leq \left(\frac{2R(m')}{R(2^l)} + 258 \ln(2M_\dagger/\delta)\right) R(\min\{2^l, m'\}) \leq 2^9 \ln(2M_\dagger/\delta) R(m'). \\ &\hspace{20em} (r(2^n) \leq r(m')) \end{aligned}$$

Summing over $l$ and by $n+1 \leq \widehat{n}$ proves the second statement. For the third statement, by Lemma 35,

$$\begin{aligned} \sum_{\tau=\mathfrak{A}.s}^{m} (f_\tau^\star - C^\tau) &= \tilde{\mathcal{O}}\left(\sqrt{B_\star \sum_{\tau=\mathfrak{A}.s}^{m} C^\tau} + B_\star\right) = \tilde{\mathcal{O}}\left(\sqrt{B_\star \left(\sum_{\tau=\mathfrak{A}.s}^{m} \widetilde{g}_\tau + \widehat{R}(m')\right)} + B_\star\right) \\ &\hspace{20em} \text{(the second statement)} \\ &\leq \tilde{\mathcal{O}}\left(B_\star \sqrt{m'}\right) + \frac{1}{2}\widehat{R}(m') \leq \widehat{R}(m'). \\ &\hspace{14em} (\widetilde{g}_\tau \leq c_4 = \tilde{\mathcal{O}}(B_\star) \text{ and AM-GM inequality}) \end{aligned}$$

This completes the proof. □

### F.4 Non-stationarity Detection: Single Block Regret Analysis

Now we introduce the MASTER algorithm (Algorithm 8) that performs non-stationarity tests and restarts. We first show the regret bound on a single block of order $n$ (of length $2^n$) that starts from $m_n$ and ends on $E_n$. Clearly $E_n \leq m_n + 2^n - 1$ since it may terminate earlier than planned. Also let $M_\dagger = m_n + 2^n - 1$ be the planned last interval. Define $\widehat{r}(m) = \widehat{R}(m)/m$, $\alpha_l = r(2^l)$, $\widehat{\alpha}_l = \widehat{r}(2^l)$, and $l_0 = \max_l\{12\widehat{\alpha}_{l-1} > c_4\}$. We divide the whole block $[m_n, E_n]$ into near-stationary segments $\mathcal{I}_1, \ldots, \mathcal{I}_\ell$ with $\mathcal{I}_i = [s_i, e_i]$, such that $\Delta_{\mathcal{I}_i} \leq r(|\mathcal{I}_i|)$ and $\Delta_{[s_i, e_i+1]} > r(|\mathcal{I}_i| + 1)$ for $i < \ell$. Note that the partition depends on the learner's behavior, but whether $m \in \mathcal{I}_i$ is determined at the beginning of interval $m$ before interaction starts. In the following lemma we give a bound on $\ell$.

**Algorithm 8** MASTER

**Input:** $\widehat{r}(\cdot)$ (defined in Appendix F.4).
**Initialize:** $m \leftarrow 1$.

1 **for** $n = 0, 1, \ldots$ **do**
  Set $m_n \leftarrow m$, and initialize a MALG (Algorithm 7) instance on $[m_n, m_n + 2^n - 1]$.
  **while** $m < m_n + 2^n$ **do**
    Receive $\widetilde{g}_m$ from MALG, follow MALG's decision, and suffer $C^m$.
2   Update MALG and set $U^l_m = \max_{\tau \in [m_n + 2^l - 1, m]} \widetilde{g}^l_\tau$ for all $0 \le l \le n$, where $\widetilde{g}^l_\tau = \frac{1}{2^l} \sum_{\tau' = \tau - 2^l + 1}^{\tau} \widetilde{g}_{\tau'}$ and $U^l_m = 0$ if $m < m_n + 2^l - 1$.
    Perform **Test 1** and **Test 2**, and increment $m \leftarrow m + 1$.
    **if** *either test fails or MALG terminates* **then** restart from Line 1

3 **Test 1**: If $m = \mathfrak{A}.e$ for some order-$l$ $\mathfrak{A}$ and $\frac{1}{2^l} \sum_{\tau = \mathfrak{A}.s}^{\mathfrak{A}.e} C^\tau \le U^l_m - 9\widehat{r}(2^l)$, return *fail*.
**Test 2**: If $\frac{1}{m - m_n + 1} \sum_{\tau = m_n}^{m} (C^\tau - \widetilde{g}_\tau) \ge 3\widehat{r}(m - m_n + 1)$. return *fail*.

**Lemma 37.** *Let $\mathcal{J} = [m_n, E_n]$. We have $\ell \le L_{\mathcal{J}}$ and $\ell \le 1 + (2c_1^{-1}\Delta_{\mathcal{J}})^{2/3}|\mathcal{J}|^{1/3} + c_3^{-1}\Delta_{\mathcal{J}}$.*

*Proof.* The first statement is clearly true. For the second statement follows from Lemma 13. $\square$

We also define $\widetilde{g}^l_\tau = \frac{1}{2^l} \sum_{\tau' = \tau - 2^l + 1}^{\tau} \widetilde{g}_{\tau'}$ and $f^{\star,l}_\tau = \frac{1}{2^l} \sum_{\tau' = \tau - 2^l + 1}^{\tau} f^\star_{\tau'}$ for $\tau \ge m_n + 2^l - 1$. We first show a running average version of the first statement in Lemma 36.

**Lemma 38.** *For any $\tau \ge m_n + 2^l - 1$, if for any $m \in [\tau - 2^l + 1, \tau]$, $\Delta_{[\mathfrak{A}.s, m]} \le r(m - \mathfrak{A}.s + 1)$ where $\mathfrak{A}$ is the base algorithm of MALG active in interval $m$, then $\widetilde{g}^l_\tau \le f^{\star,l}_\tau + \widehat{\alpha}_l$ with high probability.*

*Proof.* The case of $l = 0$ is clearly true by Lemma 36. For $l > 0$, we have

$$
\begin{aligned}
\widetilde{g}^l_\tau &= \frac{1}{2^l} \sum_{\tau' = \tau - 2^l + 1}^{\tau} \widetilde{g}_{\tau'} = \frac{1}{2^l} \sum_{\tau' = \tau - 2^l + 1}^{\tau} \sum_{l'=0}^{n} \sum_{\mathfrak{A}' \in \mathcal{S}_{l'}} \widetilde{f}^{\mathfrak{A}'}_{\tau'} \mathbb{I}\{\mathfrak{A}' \text{ is active at } \tau'\} \\
&\le \frac{1}{2^l} \sum_{\tau' = \tau - 2^l + 1}^{\tau} \sum_{l'=0}^{n} \sum_{\mathfrak{A}' \in \mathcal{S}_{l'}} (f^\star_{\tau'} + r(m^{\mathfrak{A}'}_{\tau'})) \mathbb{I}\{\mathfrak{A}' \text{ is active at } \tau'\} \\
&\qquad\qquad (\Delta_{[\mathfrak{A}'.s, m]} \le r(m - \mathfrak{A}'.s + 1) \le r(m^{\mathfrak{A}'}_{\tau'}) \text{ and Assumption 1}) \\
&\le f^{\star,l}_\tau + \frac{1}{2^l} \sum_{l'=0}^{n} \sum_{\mathfrak{A}' \in \mathcal{S}_{l'}} \sum_{\tau' = \tau - 2^l + 1}^{\tau} r(m^{\mathfrak{A}'}_{\tau'}) \mathbb{I}\{\mathfrak{A}' \text{ is active at } \tau'\} \\
&\le f^{\star,l}_\tau + \frac{2}{2^l} \sum_{l'=0}^{n} |\mathcal{S}_{l'}| R(\min\{2^l, 2^{l'}\}),
\end{aligned}
$$

where $m^{\mathfrak{A}'}_{\tau'}$ is the number of intervals that $\mathfrak{A}'$ is active up to $\tau'$, $\mathcal{S}_{l'}$ is the set of order $l'$ base algorithms that intersect with $[\tau - 2^l + 1, \tau]$, and in the last inequality we use the fact that for any $m \ge 1$,

$$
\sum_{\tau=1}^{m} r(\tau) = \sum_{\tau=1}^{m} \min\left\{\frac{c_1}{\sqrt{\tau}} + \frac{c_2}{\tau}, c_3\right\} \le \min\left\{\sum_{\tau=1}^{m} \left(\frac{c_1}{\sqrt{\tau}} + \frac{c_2}{\tau}\right), c_3 m\right\} \le 2R(m).
$$

For $l' \ge l$, we have $|\mathcal{S}_{l'}| \le 2$. For $l' < l$, note that $\mathbb{E}[|\mathcal{S}_{l'}|] \le \frac{r(2^n)}{r(2^{l'})}(2^{l-l'} + 1)$. By Lemma 50, with high probability, $|\mathcal{S}_{l'}| \le 2\mathbb{E}[|\mathcal{S}_{l'}|] + 2^8 \ln(2M_\dagger/\delta) \le \frac{2R(2^l)}{R(2^{l'})} + 258 \ln(2M_\dagger/\delta)$. Plugging these back, we obtain

$$
\frac{2}{2^l} \sum_{l'=0}^{n} |\mathcal{S}_{l'}| R(\min\{2^l, 2^{l'}\}) \le \frac{2}{2^l} \sum_{l'=0}^{l-1} \left(\frac{2R(2^l)}{R(2^{l'})} + 258 \ln(2M_\dagger/\delta)\right) R(2^{l'}) + 4 \sum_{l'=l}^{n} \alpha_l \le \widehat{\alpha}_l.
$$

This completes the proof. $\square$

Now we show the guarantee of non-stationarity detection on a single block $[m_n, E_n]$. Define $\tau_i(l)$ as the smallest interval $\tau \in \mathcal{I}_i^l \triangleq [s_i + 2^l - 1, e_i]$ ($\tau_i(l) = e_i + 1$ if such an interval does not exist) such that $\widetilde{g}_\tau^l - f_\tau^{\star,l} > 12\widehat{\alpha}_l$, and $\xi_i(l) = e_i - \tau_i(l) + 1$.

**Lemma 39.** *Let the event in Lemma 36 hold. Then with high probability,*

$$\sum_{\tau=m_n}^{E_n} (C^\tau - \widetilde{g}_\tau) \leq 3\widehat{R}(E_n - m_n + 1) + c_3,$$
$$\sum_{\tau=m_n}^{E_n} (\widetilde{g}_\tau - f_\tau^\star) \leq \tilde{\mathcal{O}}\left(\sum_{i=1}^{\ell} \widehat{R}(|\mathcal{I}_i|)\right) + 2^{10}\sum_{l=l_0}^{n} \frac{\alpha_l}{\alpha_n}\widehat{R}(2^l)\ln(2M_\dagger/\delta).$$

*Proof.* The first statement trivially holds by **Test 2** and the estimated regret in a single interval is at most $c_3$ (Assumption 1). For the second statement, define $d_\tau^l = \widetilde{g}_\tau^l - f_\tau^{\star,l}$. For a particular $\mathcal{I}_i$ and any $l \geq 0$, let $\mathcal{I}'_i = \mathcal{I}_i \cap [\tau_i(l) - 1]$. If $|\mathcal{I}'_i| \leq 2 \cdot 2^{l+1}$, then clearly $\sum_{\tau \in \mathcal{I}_i^l, \tau < \tau_i(l)} d_\tau^l \leq |\mathcal{I}'_i| \cdot 12\widehat{\alpha}_l \leq \min\{|\mathcal{I}_i|, 2 \cdot 2^{l+1}\} \cdot 12\widehat{\alpha}_l$. If $|\mathcal{I}'_i| > 2 \cdot 2^{l+1}$, then $\mathcal{I}'_i$ can be partitioned into three segments $\mathcal{H}_i^0 = [s_i, s_i + 2^{l+1} - 1]$, $\mathcal{H}_i^1 = [\tau_i(l) - 2^{l+1}, \tau_i(l) - 1]$, and $\mathcal{H}_i^2 = [s_i + 2^{l+1}, \tau_i(l) - 2^{l+1} - 1]$. Note that for $\tau \in \mathcal{H}_i^2$, the weight of $d_\tau^0$ within the sum $\sum_{\tau \in \mathcal{I}_i^{l+1}, \tau < \tau_i(l)} d_\tau^{l+1}$ is 1. Therefore, $\sum_{\tau \in \mathcal{H}_i^2} d_\tau^0 \leq \sum_{\tau \in \mathcal{I}_i^{l+1}, \tau < \tau_i(l)} d_\tau^{l+1}$. Moreover, $\sum_{\tau \in \mathcal{H}_i^0 \cup \mathcal{H}_i^1} d_\tau^0 = 2^l(d_{s_i + 2^l - 1}^l + d_{s_i + 2^{l+1} - 1}^l + d_{\tau_i(l) - 1}^l + d_{\tau_i(l) - 2^l - 1}^l) \leq 2 \cdot 2^{l+1} \cdot 12\widehat{\alpha}_l = \min\{|\mathcal{I}_i|, 2 \cdot 2^{l+1}\} \cdot 12\widehat{\alpha}_l$. This gives

$$\begin{aligned}
\sum_{\tau \in \mathcal{I}_i^l, \tau < \tau_i(l)} d_\tau^l \leq \sum_{\tau \in \mathcal{I}'_i} d_\tau^0 &= \left(\sum_{\tau \in \mathcal{H}_i^2} + \sum_{\tau \in \mathcal{H}_i^0 \cup \mathcal{H}_i^1}\right) d_\tau^0 \\
&\leq \sum_{\tau \in \mathcal{I}_i^{l+1}, \tau < \tau_i(l)} d_\tau^{l+1} + \min\{|\mathcal{I}_i|, 2 \cdot 2^{l+1}\} \cdot 12\widehat{\alpha}_l \\
&\leq \sum_{\tau \in \mathcal{I}_i^{l+1}, \tau < \tau_i(l+1)} d_\tau^{l+1} + 12\widehat{\alpha}_l\xi_i(l+1) + \min\{|\mathcal{I}_i|, 2^l\} \cdot 48\widehat{\alpha}_l. \\
&\hspace{6cm} (d_\tau^{l+1} = \tfrac{1}{2}(d_\tau^l + d_{\tau - 2^l}^l)) \\
&\leq \sum_{\tau \in \mathcal{I}_i^{l+1}, \tau < \tau_i(l+1)} d_\tau^{l+1} + 24\widehat{\alpha}_{l+1}\xi_i(l+1) + \min\{|\mathcal{I}_i|, 2^l\} \cdot 48\widehat{\alpha}_l. \\
&\hspace{6cm} \left(\widehat{\alpha}_l = \tfrac{\widehat{R}(2^l)}{2^l} \leq \tfrac{2\widehat{R}(2^{l+1})}{2^{l+1}} \leq 2\widehat{\alpha}_{l+1}\right)
\end{aligned}$$

Combining the two cases above, we have

$$\sum_{\tau \in \mathcal{I}_i^l, \tau < \tau_i(l)} d_\tau^l \leq \sum_{\tau \in \mathcal{I}_i^{l+1}, \tau < \tau_i(l+1)} d_\tau^{l+1} + 24\widehat{\alpha}_{l+1}\xi_i(l+1) + \min\{|\mathcal{I}_i|, 2^l\} \cdot 48\widehat{\alpha}_l.$$

Applying this recursively, we have for a given $\mathcal{I}_i$,

$$\begin{aligned}
\sum_{\tau \in \mathcal{I}_i} (\widetilde{g}_\tau - f_\tau^\star) = \sum_{\tau \in \mathcal{I}_i^0, \tau < \tau_i(0)} d_\tau^0 &\leq \sum_{\tau \in \mathcal{I}_i^n, \tau < \tau_i(n)} d_\tau^n + 24\sum_{l=0}^{n-1} \widehat{\alpha}_{l+1}\xi_i(l+1) + 48\sum_{l=0}^{n-1} \widehat{R}(\min\{|\mathcal{I}_i|, 2^l\}) \\
&\hspace{3cm} (\min\{|\mathcal{I}_i|, 2^l\}\widehat{\alpha}_l \leq \widehat{r}(\min\{|\mathcal{I}_i|, 2^l\}) \min\{|\mathcal{I}_i|, 2^l\} = \widehat{R}(\min\{|\mathcal{I}_i|, 2^l\})) \\
&\leq 12|\mathcal{I}_i|\widehat{\alpha}_n + 24\sum_{l=1}^{n} \widehat{\alpha}_l\xi_i(l) + \tilde{\mathcal{O}}\left(\widehat{R}(|\mathcal{I}_i|)\right) \leq 24\sum_{l=1}^{n} \widehat{\alpha}_l\xi_i(l) + \tilde{\mathcal{O}}\left(\widehat{R}(|\mathcal{I}_i|)\right).
\end{aligned}$$

Summing over all $i$ and by $l < l_0 \implies 12\widehat{\alpha}_l > c_4 \implies \xi_i(l) = 0$, we have:

$$\sum_{\tau=m_n}^{E_n} (\widetilde{g}_\tau - f_\tau^\star) \leq \tilde{\mathcal{O}}\left(\sum_{i=1}^{\ell} R(|\mathcal{I}_i|)\right) + 24\sum_{l=l_0}^{n}\sum_{i=1}^{\ell} \widehat{\alpha}_l\xi_i(l).$$

Now note that for any fixed $l$, we have:

$$\begin{aligned}\sum_{i=1}^{\ell} \widehat{\alpha}_l \xi_i(l) &= \widehat{\alpha}_l \sum_{i=1}^{\ell} \min\{\xi_i(l), 4\cdot 2^l\} + \widehat{\alpha}_l \sum_{i=1}^{\ell} (\xi_i(l) - 4\cdot 2^l)_+ \\ &\le 4\sum_{i=1}^{\ell}\left(\widehat{R}(|\mathcal{I}_i|) + \frac{4\alpha_l}{\alpha_n}\widehat{R}(2^l)\ln(2M_\dagger/\delta)\right).\end{aligned}$$

(Lemma 40 and $\widehat{\alpha}_l \min\{\xi_i(l), 4\cdot 2^l\} \le 4\widehat{r}(\min\{\xi_i(l), 2^l\}) \min\{\xi_i(l), 2^l\} = 4\widehat{R}(\min\{\xi_i(l), 2^l\})$)

Putting everything together completes the proof. □

**Lemma 40.** *For any $l \le n$, $\sum_{i=1}^{\ell} \widehat{\alpha}_l(\xi_i(l) - 4\cdot 2^l)_+ \le \frac{4\alpha_l}{\alpha_n}\widehat{R}(2^l)\ln(2M_\dagger/\delta)$ with high probability.*

*Proof.* Denote by $A_l$ the number of candidate starting points of an order-$l$ algorithm in $[\tau_i(l), e_i - 2\cdot 2^l]$ for some $i$. Note that this quantity is lower bounded by $\sum_{i=1}^{\ell}(\xi_i(l) - 4\cdot 2^l)_+/2^l$. Moreover, if in interval $m \in [\tau_i(l), e_i - 2\cdot 2^l]$, an order-$l$ algorithm $\mathfrak{A}$ starts, then **Test 1** is performed at $m + 2^l - 1 \le e_i$, and **Test 1** returns *fail* with high probability because

$$\begin{aligned}\frac{1}{2^l}\sum_{\tau=\mathfrak{A}.s}^{\mathfrak{A}.e} C^\tau &\le \frac{1}{2^l}\sum_{\tau=\mathfrak{A}.s}^{\mathfrak{A}.e} \widetilde{g}_\tau + \widehat{\alpha}_l \qquad (\Delta_{[\mathfrak{A}.s,\mathfrak{A}.e]} \le \Delta_{\mathcal{I}_i} \le r(|\mathcal{I}_i|) \le r(2^l) \text{ and Lemma 36}) \\ &\le \frac{1}{2^l}\sum_{\tau=\mathfrak{A}.s}^{\mathfrak{A}.e} f^\star_\tau + 2\widehat{\alpha}_l \\ &\qquad (\text{Lemma 38, } \Delta_{\mathcal{I}_i} \le r(|\mathcal{I}_i|), \text{ and } [\mathfrak{A}'.s, \mathfrak{A}'.e] \subseteq [\mathfrak{A}.s, \mathfrak{A}.e] \text{ if } \mathfrak{A}' \text{ is active within } [\mathfrak{A}.s, \mathfrak{A}.e]) \\ &\le f^{\star,l}_{\tau_i(l)} + 2\widehat{\alpha}_l + \Delta_{\mathcal{I}_i} \\ &\le \widetilde{g}^l_{\tau_i(l)} - 12\widehat{\alpha}_l + 3\widehat{\alpha}_l \le \widetilde{g}^l_{\tau_i(l)} - 9\widehat{\alpha}_l. \qquad (\Delta_{\mathcal{I}_i} \le r(|\mathcal{I}_i|) \le r(2^l) \le \widehat{\alpha}_l)\end{aligned}$$

This is a contradiction by the definition of $E_n$. Therefore, all candidate starting points of order-$l$ algorithm in $[\tau_i(l), e_i - 2\cdot 2^l]$ does not instantiate an order-$l$ algorithm. Let $X_m = \{m \in [\tau_i(l), e_i - 2\cdot 2^l] \text{ for some } i\}$, $X'_m = \{m \in [\tau_i(l), e_i] \text{ for some } i\}$, $Y_m = \{(m - m_n) \mod 2^l = 0\}$ and $Z_m = \{\nexists \text{ order-}l\ \mathfrak{A}' \text{ such that } \mathfrak{A}'.s = m\}$, we have

$$A_l = \sum_{m=m_n}^{m_n+2^m-1} \mathbb{I}\{X_m, Y_m\} = \sum_{m=m_n}^{m_n+2^m-1} \mathbb{I}\{X_m, Y_m, Z_m\} \le \sum_{m=m_n}^{m_n+2^m-1} \mathbb{I}\{X'_m, Y_m, Z_m\}.$$

Note that conditioned on $X'_m \cap Y_m$, the event $Z_m$ happens with a constant probability $1 - \frac{\alpha_n}{\alpha_l}$. Moreover, $Z_m = 0$ implies $X'_{m'} = 0$ for $m' > m$. Therefore, $\sum_{m=m_n}^{m_n+2^m-1} \mathbb{I}\{X'_m, Y_m, Z_m\}$ counts the number of trials up to the first success with success probability $\frac{\alpha_n}{\alpha_l}$ of each trial. Then with probability at least $1 - \delta/(2M_\dagger^2)$, we have $A_l \le \frac{4\alpha_l}{\alpha_n}\ln(2M_\dagger/\delta)$. Thus,

$$\sum_{i=1}^{\ell} \widehat{\alpha}_l(\xi_i(l) - 4\cdot 2^l)_+ \le \widehat{\alpha}_l 2^l \cdot \frac{4\alpha_l}{\alpha_n}\ln(2M_\dagger/\delta) \le \frac{4\alpha_l}{\alpha_n}\widehat{R}(2^l)\ln(2M_\dagger/\delta).$$

This completes the proof. □

Now we present the regret guarantee in a single block.

**Lemma 41.** *Within a single block $\mathcal{J} = [m_n, E_n]$, we have*

$$\sum_{m\in\mathcal{J}} (C^m - f^\star_m) = \tilde{\mathcal{O}}\left(c_1\sqrt{\ell|\mathcal{J}|} + c_2\ell + \left(c_1 + \frac{c_2c_4}{c_1}\right)2^{n/2} + \frac{c_2^2}{c_3} + c_3\right).$$

*Proof.* By Lemma 39, we have

$$\begin{aligned}\sum_{m\in\mathcal{J}} (C^m - f^\star_m) &= \sum_{m\in\mathcal{J}} (C^m - \widetilde{g}_m) + \sum_{m\in\mathcal{J}} (\widetilde{g}_m - f^\star_m) \\ &= \tilde{\mathcal{O}}\left(\widehat{R}(|\mathcal{J}|) + \sum_{i=1}^{\ell}\widehat{R}(|\mathcal{I}_i|) + \sum_{l=l_0}^{n}\frac{\alpha_l}{\alpha_n}\widehat{R}(2^l) + c_3\right).\end{aligned}$$

Note that by Cauchy-Schwarz inequality:

$$\widehat{R}(|\mathcal{J}|)+\sum_{i=1}^{\ell}\widehat{R}(|\mathcal{I}_i|)=\tilde{\mathcal{O}}\left(\left(c_1\sqrt{|\mathcal{J}|}+c_2\right)+\sum_{i=1}^{\ell}\left(c_1\sqrt{|\mathcal{I}_i|}+c_2\right)\right)=\tilde{\mathcal{O}}\left(c_1\sqrt{\ell|\mathcal{J}|}+c_2\ell\right).$$

Moreover, by the definition of $l_0$, we have $12\widehat{\alpha}_{l_0}=2^{10}\widehat{n}\ln(2M_{\dagger}/\delta)\min\{\frac{c_1}{\sqrt{2^{l_0}}}+\frac{c_2}{2^{l_0}},c_3\}\le c_4$, which implies $c_2\le c_4 2^{l_0}$ by $c_4\le c_3$. Now for any $l\ge l_0$,

$$\begin{aligned}\frac{\alpha_l}{\alpha_n}\widehat{R}(2^l)&=\tilde{\mathcal{O}}\left(\frac{R(2^l)^2}{R(2^n)}2^{n-l}\right)=\tilde{\mathcal{O}}\left(\frac{c_1^2 2^l+c_2^2}{c_1 2^{n/2}+c_2}2^{n-l}+\frac{c_1^2 2^l+c_2^2}{c_3 2^n}2^{n-l}\right)\\&=\tilde{\mathcal{O}}\left(c_1 2^{n/2}+\frac{c_2c_4 2^{l_0}}{c_1}2^{n/2-l}+\frac{c_1^2}{c_3}+\frac{c_2^2}{c_3}2^{-l}\right)\\&=\tilde{\mathcal{O}}\left(\left(c_1+\frac{c_2c_4}{c_1}\right)2^{n/2}+\frac{c_1^2}{c_3}+\frac{c_2^2}{c_3}2^{-l}\right)=\tilde{\mathcal{O}}\left(\left(c_1+\frac{c_2c_4}{c_1}\right)2^{n/2}+\frac{c_2^2}{c_3}2^{-l}\right),\end{aligned}$$

where in the last inequality we assume $c_1\le c_3 2^{n/2}$ without loss of generality and have $\frac{c_1^2}{c_3}\le c_1 2^{n/2}$ (note that if $c_1>c_3 2^{n/2}$, then $c_1 2^{n/2}>c_3 2^n$ and the regret bound is vacuous). Summing over $l$ and putting everything together, we obtain:

$$\sum_{m\in\mathcal{J}}(C^m-f_m^\star)=\tilde{\mathcal{O}}\left(c_1\sqrt{\ell|\mathcal{J}|}+c_2\ell+\left(c_1+\frac{c_2c_4}{c_1}\right)2^{n/2}+\frac{c_2^2}{c_3}+c_3\right).$$

□

### F.5 Single Epoch Regret Analysis

We call $[m_0,E]$ an *epoch* if $m_0$ is the first interval after restart from Line 1 or $m_0=1$, and $E$ is the first interval where a restart after interval $m$ is triggered. The regret guarantee in a single epoch is shown in the following lemma.

**Lemma 42.** *Let $\mathcal{E}$ be an epoch, then $\sum_{m\in\mathcal{E}}(C^m-f_m^\star)=\tilde{\mathcal{O}}(c_1\sqrt{\ell_{\mathcal{E}}|\mathcal{E}|}+c_2\ell_{\mathcal{E}}+(c_1+\frac{c_2c_4}{c_1})\sqrt{|\mathcal{E}|}+\frac{c_2^2}{c_3}+c_3)$, where $\ell_{\mathcal{E}}=\tilde{\mathcal{O}}(1+(c_1^{-1}\Delta_{\mathcal{E}})^{2/3}|\mathcal{E}|^{1/3}+c_3^{-1}\Delta_{\mathcal{E}})$ and $\ell_{\mathcal{E}}=\tilde{\mathcal{O}}(L_{\mathcal{E}})$.*

*Proof.* Suppose $\mathcal{E}$ consists of blocks $\mathcal{J}_1,\ldots,\mathcal{J}_n$ and the number of near stationary segments (as discussed in Appendix F.4) in $\mathcal{J}_i$ is $\ell_i$. Then, $|\mathcal{E}|=\Theta(2^n)$, and by Lemma 41 and Cauchy-Schwarz inequality,

$$\begin{aligned}\sum_{m\in\mathcal{E}}(C^m-f_m^\star)&=\tilde{\mathcal{O}}\left(c_1\sum_{i=1}^{n}\sqrt{\ell_i|\mathcal{J}_i|}+c_2\sum_{i=1}^{n}\ell_i+\left(c_1+\frac{c_2c_4}{c_1}\right)2^{n/2}+\frac{c_2^2}{c_3}+c_3\right)\\&=\tilde{\mathcal{O}}\left(c_1\sqrt{\sum_{i=1}^{n}\ell_i|\mathcal{E}|}+c_2\sum_{i=1}^{n}\ell_i+\left(c_1+\frac{c_2c_4}{c_1}\right)\sqrt{|\mathcal{E}|}+\frac{c_2^2}{c_3}+c_3\right).\end{aligned}$$

Finally by Lemma 37 and Hölder's inequality, $\sum_{i=1}^n\ell_i=\tilde{\mathcal{O}}(1+(c_1^{-1}\Delta_{\mathcal{E}})^{2/3}|\mathcal{E}|^{1/3}+c_3^{-1}\Delta_{\mathcal{E}})$ and $\sum_{i=1}^n\ell_i=\tilde{\mathcal{O}}(L_{\mathcal{E}})$. □

### F.6 Full Regret Guarantee

To derive the full regret guarantee of the MASTER algorithm (Algorithm 8), we first bound the number of epochs by the following two lemmas. Define $\mathfrak{N}_{[1,M']}$ as the number of times MALG terminates within $[1,M']$

**Lemma 43.** *Let $m$ be in an epoch starting from interval $m_0$. If $\Delta_{[m_0,m]}\le r(m-m_0+1)$, then no restart would be triggered by **Test 1** or **Test 2** in interval $m$ with high probability.*

*Proof.* We first show that **Test 1** would not fail. Let $m = \mathfrak{A}.e$ where $\mathfrak{A}$ is any order-$l$ base algorithm in a block of order $n$ starting from $m_n$. Then with high probability,

$$U_m^l = \max_{\tau\in[m_n+2^l-1,m]} \widetilde{g}_\tau^l \le \max_{\tau\in[m_n+2^l-1,m]} f_\tau^{\star,l} + \widehat{r}(2^l) \qquad \text{(Lemma 38)}$$
$$\le \frac{1}{2^l}\sum_{\tau=\mathfrak{A}.s}^{m} f_\tau^\star + \widehat{r}(2^l) + \Delta_{[m_n,m]}$$
$$\le \frac{1}{2^l}\sum_{\tau=\mathfrak{A}.s}^{m} C^\tau + 2\widehat{r}(2^l) + \Delta_{[m_n,m]} \le \frac{1}{2^l}\sum_{\tau=\mathfrak{A}.s}^{m} C^\tau + 3\widehat{r}(2^l).$$
$$\text{(Lemma 36 and } \Delta_{[m_n,m]} \le \Delta_{[m_0,m]} \le r(m-m_0+1) \le r(2^l))$$

Thus, **Test 1** would not fail. For **Test 2**, by Lemma 36 and $\Delta_{[m_n,m]} \le \Delta_{[m_0,m]} \le r(m-m_0+1) \le r(m-m_n+1)$:

$$\sum_{\tau=m_n}^{m}(C^\tau - \widetilde{g}_\tau) \le \widehat{R}(m-m_n+1),$$

Thus, **Test 2** also would not fail. ☐

**Lemma 44.** *Assuming that MALG does not terminate without non-stationarity, with high probability, the number of epochs within* $[1, M']$ *is upper bounded by* $L_{[1,M']}$ *and* $1 + (2c_1^{-1}\Delta_{[1,M']})^{2/3}{M'}^{1/3} + c_3^{-1}\Delta_{[1,M']} + \mathfrak{N}_{[1,M']}$.

*Proof.* The first upper bound is clearly true by partitioning $[1, M']$ into segments without non-stationarity. For the second upper bound, by Lemma 43, if an epoch $[m_0, E]$ is not the last epoch, then $\Delta_{[m_0,E]} > r(E - m_0 + 1)$ or MALG terminates with high probability. Applying Lemma 13 completes the proof. ☐

**Theorem 13.** *If Assumption 1 holds, then MASTER (Algorithm 8) ensures with high probability (ignoring lower order terms), for any* $M' \ge 1$*:*

$$\widetilde{R}_{M'} = \tilde{\mathcal{O}}\left(\left(c_1 + \frac{c_2c_4}{c_1}\right)\sqrt{L_{[1,M']}M'}\right) \text{ and}$$
$$\widetilde{R}_{M'} = \tilde{\mathcal{O}}\left(\left(c_1 + \frac{c_2c_4}{c_1}\right)\sqrt{(\mathfrak{N}_{[1,M']}+1)M'} + \left(c_1^{2/3} + \frac{c_2c_4}{c_1^{4/3}}\right)\Delta_{[1,M']}^{1/3}{M'}^{2/3}\right).$$

*Proof.* Let $\mathcal{E}_1, \ldots, \mathcal{E}_N$ be epochs in $[1, M']$ and $\mathcal{E} = \bigcup_{i=1}^N \mathcal{E}_i$. Then by Lemma 42 and Cauchy-Schwarz inequality, we have:

$$\widetilde{R}_{M'} = \tilde{\mathcal{O}}\left(\sum_{i=1}^{N}\left(c_1\sqrt{\ell_{\mathcal{E}_i}|\mathcal{E}_i|} + c_2\ell_{\mathcal{E}_i} + \left(c_1 + \frac{c_2c_4}{c_1}\right)\sqrt{|\mathcal{E}_i|} + \frac{c_2^2}{c_3} + c_3\right)\right)$$
$$= \tilde{\mathcal{O}}\left(c_1\sqrt{\ell_{\mathcal{E}}M'} + c_2\ell_{\mathcal{E}} + \left(c_1 + \frac{c_2c_4}{c_1}\right)\sqrt{NM'} + \left(\frac{c_2^2}{c_3} + c_3\right)N\right),$$

where $\ell_{\mathcal{E}} = \sum_{i=1}^N \ell_{\mathcal{E}_i}$. Below we assume sub-linear $L_{[1,M']}, \Delta_{[1,M']}$ and only write down dominating terms. For $L$-dependent bound, note that $N \le L_{[1,M']}$ by Lemma 44 and $\ell_{\mathcal{E}} \le N + L_{[1,M']} = \tilde{\mathcal{O}}(L_{[1,M']})$ by Lemma 42. Thus, $c_2\ell_{\mathcal{E}} + (\frac{c_2^2}{c_3} + c_3)N$ is a lower order term, and

$$\widetilde{R}_{M'} = \tilde{\mathcal{O}}\left(\left(c_1 + \frac{c_2c_4}{c_1}\right)\sqrt{L_{[1,M']}M'}\right).$$

For $\Delta$-dependent bound, note that by Lemma 42, Hölder's inequality, and Lemma 44,

$$\ell_{\mathcal{E}} = \tilde{\mathcal{O}}\left(N + (c_1^{-1}\Delta_{[1,M']})^{2/3}{M'}^{1/3} + c_3^{-1}\Delta_{[1,M']}\right)$$
$$= \tilde{\mathcal{O}}\left(\mathfrak{N}_{[1,M']} + 1 + (c_1^{-1}\Delta_{[1,M']})^{2/3}{M'}^{1/3} + c_3^{-1}\Delta_{[1,M']}\right).$$

Ignoring lower order term of the form $\sqrt{\Delta_{[1,M']}M'}$, we have

$$
\begin{aligned}
&c_1\sqrt{\ell_{\mathcal{E}}M'}+\left(c_1+\frac{c_2c_4}{c_1}\right)\sqrt{NM'}\\
&=\tilde{\mathcal{O}}\left(\left(c_1+\frac{c_2c_4}{c_1}\right)\sqrt{\left(\mathfrak{N}_{[1,M']}+1+(c_1^{-1}\Delta_{[1,M']})^{2/3}M'^{1/3}+c_3^{-1}\Delta_{[1,M']}\right)M'}\right)\\
&=\tilde{\mathcal{O}}\left(\left(c_1+\frac{c_2c_4}{c_1}\right)\sqrt{(\mathfrak{N}_{[1,M']}+1)M'}+\left(c_1^{2/3}+\frac{c_2c_4}{c_1^{4/3}}\right)\Delta_{[1,M']}^{1/3}M'^{2/3}\right).
\end{aligned}
$$

The remaining $c_2\ell_{\mathcal{E}}+(\frac{c_2^2}{c_3}+c_3)N$ is again a lower order term. □

### F.7 Proof of Theorem 6

We are ready to present the regret guarantee of the MASTER algorithm combining with different base algorithms. Recall $L=1+\sum_{k=1}^{K-1}\mathbb{I}\{P_{k+1}\neq P_k \text{ or } c_{k+1}\neq c_k\}$.

**Theorem 14.** *Let $\mathfrak{A}$ be Algorithm 8 with Algorithm 2 as base algorithm. Then Algorithm 1 with $\mathfrak{A}$ ensures with high probability, for any $K'\in[K]$,*

$$R_{K'}=\tilde{\mathcal{O}}\left(\min\left\{B_\star S\sqrt{ALK'},B_\star S\sqrt{AK'}+(B_\star^2S^2A(\Delta_c+B_\star\Delta_P)T_{\max})^{1/3}K'^{2/3}\right\}\right).$$

*Proof.* By Lemma 33 and Theorem 13 with $\mathfrak{N}_{[1,M]}=0$, we have for any $M'\leq M$,

$$\mathring{R}_{M'}\leq\widetilde{R}_{M'}=\tilde{\mathcal{O}}\left(\min\left\{B_\star S\sqrt{AL_{[1,M]}M'},B_\star S\sqrt{AM'}+(B_\star^2S^2A\Delta_{[1,M]})^{1/3}M'^{2/3}\right\}\right),$$

where $L_{[1,M]}=L$ and $\Delta_{[1,M]}=\tilde{\mathcal{O}}((\Delta_c+B_\star\Delta_P)T_{\max})$. Applying Lemma 16, we have for any $K'\in[K]$ (ignoring lower order terms),

$$\mathring{R}_{M_{K'}}=\tilde{\mathcal{O}}\left(\min\left\{B_\star S\sqrt{ALK'},B_\star S\sqrt{AK'}+(B_\star^2S^2A(\Delta_c+B_\star\Delta_P)T_{\max})^{1/3}K'^{2/3}\right\}\right).$$

Applying Lemma 15 completes the proof. □

We are now ready to prove Theorem 6.

*Proof of Theorem 6.* By Lemma 30 and Lemma 31, when Algorithm 6 terminates in interval $E$ where $[m_0,E]$ is an epoch, we have $\Delta'_{[m_0,E+1]}>\eta_{E-m_0+2}$. Therefore, $\mathfrak{N}_{[1,K]}=\tilde{\mathcal{O}}(1+(B_\star^2S^2A)^{-1/3}(T_\star\Delta'_{[1,K]})^{2/3}K^{1/3}+H\Delta'_{[1,K]})$ by Lemma 13 and the definition of $\eta_m$. Then by Lemma 34 and Theorem 13, we have $\mathfrak{A}_1$ ensures when $s_1^m=s_{\text{init}}$ for $m\leq K$,

$$\mathring{R}_K=\widetilde{R}_K=\tilde{\mathcal{O}}\left(\min\left\{B_\star S\sqrt{ALK},B_\star S\sqrt{AK}+(B_\star^2S^2A(\Delta_c+B_\star\Delta_P)T_\star)^{1/3}K^{2/3}\right\}\right),$$

where we apply $\Delta'_{[1,K]}=\tilde{\mathcal{O}}(\Delta_c+B_\star\Delta_P)$, $L_{[1,K]}=L$, and $\Delta_{[1,K]}=\tilde{\mathcal{O}}((\Delta_c+B_\star\Delta_P)T_\star)$. Moreover, by Theorem 14, $\mathfrak{A}_2$ ensures $R_{K'}$ being sub-linear w.r.t $K'$ for any $K'\in[K]$. Applying Theorem 11 completes the proof. □

## G Auxiliary Lemmas

**Lemma 45.** *[Chen et al., 2022b, Lemma 48] $x\leq a\sqrt{x}+b$ implies $x\leq(a+\sqrt{b})^2\leq 2a^2+2b$.*

**Lemma 46.** *[Rosenberg and Mansour, 2021, Lemma 6] Let $\pi$ be a policy whose expected hitting time starting from any state is at most $\tau$. Then for any $\delta\in(0,1)$, with probability at least $1-\delta$, it takes no more than $4\tau\ln\frac{2}{\delta}$ steps to reach the goal state following $\pi$.*

**Lemma 47.** *[Chen et al., 2021a, Lemma 30] For any random variable $X$ with $\|X\|_\infty\leq C$, we have* $\textsc{Var}[X^2]\leq 4C^2\textsc{Var}[X]$.

**Lemma 48.** *([Chen et al., 2021a, Lemma 31]) Define* $\Upsilon = \{v \in [0, B]^{\mathcal{S}_+} : v(g) = 0\}$. *Let* $f : \Delta_{\mathcal{S}_+} \times \Upsilon \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ *with* $f(p, v, n, B, \iota) = pv - \max\left\{c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}}, c_2\frac{B\iota}{n}\right\}$ *with* $c_1^2 \leq c_2$. *Then* $f$ *satisfies for all* $p \in \Delta_{\mathcal{S}_+}, v \in \Upsilon$ *and* $n, \iota > 0$,

1. $f(p, v, n, B, \iota)$ *is non-decreasing in* $v(s)$, *that is,*

$$\forall v, v' \in \Upsilon, v(s) \leq v'(s), \forall s \in \mathcal{S}^+ \implies f(p, v, n, B, \iota) \leq f(p, v', n, B, \iota);$$

2. $f(p, v, n, B, \iota) \leq pv - \frac{c_1}{2}\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} - \frac{c_2}{2}\frac{B\iota}{n}$.

**Lemma 49** (Any interval Freedman's inequality)**.** *Let* $\{X_i\}_{i=1}^\infty$ *be a martingale difference sequence w.r.t the filtration* $\{\mathcal{F}_i\}_{i=0}^\infty$ *and* $|X_i| \leq B$ *for some* $B > 0$. *Then with probability at least* $1 - \delta$, *for all* $1 \leq l \leq n$ *simultaneously,*

$$\left|\sum_{i=l}^n X_i\right| \leq 3\sqrt{\sum_{i=l}^n \mathbb{E}[X_i^2|\mathcal{F}_{i-1}] \ln\frac{16B^2n^5}{\delta}} + 2B\ln\frac{16B^2n^5}{\delta} \tag{5}$$

$$\leq 3\sqrt{2\sum_{i=l}^n X_i^2 \ln\frac{16B^2n^5}{\delta} + 18B\ln\frac{16B^2n^5}{\delta}}. \tag{6}$$

*Proof.* For each $l \geq 1$, by [Chen et al., 2022a, Lemma 38], with probability at least $1 - \frac{\delta}{4l^2}$, Eq. (5) holds for all $n \geq l$. Then by Lemma 50, with probability at least $1 - \frac{\delta}{4l^2}$, Eq. (6) holds for all $n \geq l$. Applying a union bound over $l$ completes the proof. $\square$

**Lemma 50.** *Suppose* $\{X_i\}_{i=1}^\infty$ *is a sequence of random variables w.r.t the filtration* $\{\mathcal{F}_i\}_{i=0}^\infty$ *and satisfies* $X_i \in [0, B]$ *for some* $B > 0$. *Then with probability at least* $1 - \delta$, *for all* $1 \leq l \leq n$ *simultaneously,*

$$\sum_{i=l}^n \mathbb{E}[X_i|\mathcal{F}_{i-1}] \leq 2\sum_{i=l}^n X_i + 12B\ln\frac{2n}{\delta},$$

$$\sum_{i=l}^n X_i \leq 2\sum_{i=l}^n \mathbb{E}[X_i|\mathcal{F}_{i-1}] + 24B\ln\frac{2n}{\delta}.$$

*Proof.* For each $l \geq 1$, by [Chen et al., 2022a, Lemma 39], with probability at least $1 - \frac{\delta}{2l^2}$, the two inequalities above hold for all $n \geq l$. Taking a union bound over $l$ completes the proof. $\square$