Proceedings of Machine Learning Research vol XX:1–49, 2023

Multi-Agent Reinforcement Learning with Reward Delays

Yuyang Zhang YUYANGZHANG@G.HARVARD.EDU
Runyu Zhang RUNYUZHANG@FAS.HARVARD.EDU
Harvard University, School of Engineering and Applied Science

Gen Li LIGEN@WHARTON.UPENN.EDU
University of Pennsylvania, Wharton Department of Statistics and Data Science

Yuantao Gu GYT@TSINGHUA.EDU.CN
Tsinghua University, Department of Electronic Engineering

Na Li NALI@SEAS.HARVARD.EDU
Harvard University, School of Engineering and Applied Science

Abstract

This paper considers multi-agent reinforcement learning (MARL) where the rewards are received after delays and the delay time varies among agents. Based on the V-learning framework, this paper proposes MARL algorithms that efficiently deal with reward delays. When the delays are finite, our algorithm reaches a coarse correlated equilibrium (CCE) with rate $\tilde{\mathcal{O}}(\frac{H^3\sqrt{S\mathcal{T}_K}}{K} + \frac{H^3\sqrt{SA}}{\sqrt{K}})$ where K is the number of episodes, H is the planning horizon, S is the size of the state space, A is the size of the largest action space, and $\mathcal{T}_K$ is the measure of the total delay defined in the paper. Moreover, our algorithm can be extended to cases with infinite delays through a reward skipping scheme. It achieves convergence rate similar to the finite delay case.

Keywords: Reward Delays, Markov Games, Multi-Agent Reinforcement Learning

1. Introduction

Multi-agent reinforcement learning (MARL) finds extensive applications such as recommendation systems (Zhao et al., 2020), medical treatments (Li et al., 2022; Martinho et al., 2021), multi-agent robotics systems (Brambilla et al., 2013; Malus et al., 2020; Choi and Ahn, 2010), autonomous driving (Kiran et al., 2021), etc. In these multi-agent problems, individuals aim to learn to interact with the environment under the influence of other agents.

Motivated by the empirical success of MARL, there is a surge of recent studies focusing on designing MARL algorithms with theoretical convergence guarantees such as V-learning, V-learning OMD, SPoCMAR, etc (Jin et al., 2021; Song et al., 2021; Daskalakis et al., 2022; Mao and Başar, 2022). In these algorithms, agents rely on real-time observations of the reward values to update their policies or value functions. However, in reality, rewards generally come with delays. For example, in multi-agent robotics systems (Duan et al., 2022), it takes time for the robots to communicate and gather the information of other agents, leading to delays in rewards. Another example is the medical treatment process (Li et al., 2022), where the effectiveness of a treatment strategy cannot be observed immediately. It generally takes a long time for a patient to respond and recover. In addition, the delays are typically heterogeneous, depending on factors such as the status of communication channels or thepatient's physiological state. All these real-life examples suggest that it is crucial to understand how delays affect the learning process and how to design RL algorithms that could accommodate reward delays efficiently.

There has been extensive empirical work on MARL with reward delays. Different approaches are proposed to handle delays, including but not restricted to evaluating curiosity (Shao et al., 2019), learning temporal structures (Hauwere et al., 2011), predicting strategic interactions between agents (Tang et al., 2018), and predicting the environment (Firoiu et al., 2018) with neural networks. However, from the theoretical perspective, few results are known for MARL. Other related settings include single-agent reinforcement learning (SARL) and multi-arm bandit (MAB). For SARL, two recent work (Lancewicki et al., 2022; Jin et al., 2022) studies adversarial reward delays. Unfortunately, their methods suffer from the curse of dimensionality when directly extended to MARL. Other previous work (Walsh et al., 2009; Katsikopoulos and Engelbrecht, 2003) only focuses on constant reward delays. For MAB, Gyorgy and Joulani (2021); Zimmert and Seldin (2020); Gael et al. (2020) tackle adversarial reward delays while Cesa-Bianchi et al. (2016); Neu et al. (2010) focus on constant reward delays. In the MARL setting, there also exist lines of work studying state or action delays in multi-agent settings (Agarwal and Aggarwal, 2021; Bouteiller et al., 2020b; Chen et al., 2020), but the settings studied in these papers are different from the reward delay considered in our paper and thus are out of scope of this paper.

Our Contributions. In this paper, we focus on a specific type of MARL model, the general-sum Markov games (Shapley, 1953; Littman et al., 2001). We propose the delay-adaptive multi-agent V-learning (DA-MAVL) to learn coarse-correlated equilibria (CCEs) under reward delays, in which the learning of the agents can be finished in a fully decentralized manner. Namely, every agent runs its own learning algorithm without communicating with others. Note that this is nontrivial because different agents may receive reward for the same visit at different episodes due to the heterogeneous delays among agents which might lead to misalignment and divergent behavior. Our DA-MAVL algorithm circumvents this problem by carefully selecting proper reward information for learning and aligning the behaviour of the agents, leading the algorithm to converge to CCEs.

In the case of finite delays, our algorithm achieves the CCE-gap as small as $\tilde{\mathcal{O}}(\frac{H^3\sqrt{S\mathcal{T}_K}}{K} + \frac{H^3\sqrt{SA}}{\sqrt{K}})$ with K episode-samples (Theorem 1). Here H is the planning horizon, S is the size of the state space, $A = \max_m |\mathcal{A}_m|$ is the largest size of one agent' action space, and $\mathcal{T}_K$ can be seen as a measure of the total delay. In the worst case, $\sqrt{\mathcal{T}_K}$ is the order of $\mathcal{O}(\sqrt{K})$, which implies the CCE-gap is as small as $\tilde{\mathcal{O}}(\frac{1}{\sqrt{K}})$. This dependence of K matches the original result of V-learning (Jin et al., 2021; Song et al., 2021), indicating that DA-MAVL successfully aligns the behaviour of the agents. Moreover, both terms are independent of the number of agents, meaning that DA-MAVL scales nicely with the system size. Our proposed DA-MAVL algorithm can be easily extended to cases with infinite delays. With a novel skipping metric inspired by Zimmert and Seldin (2020), our algorithm can skip the infinite delays without previous knowledge of the delay sequence and achieve the CCE-gap similarly to the finite delay case (Theorem 1). To the best of our knowledge, our results give the first convergence rate guarantee for general-sum MGs under adversarial reward delays.

Due to the space limit, we defer "related work", detailed algorithms, proofs, and simulation settings in the appendix of the online report (Zhang et al., 2022) of this submission.

2. Problem Setup & Preliminary

2.1. Markov Games with Reward Delays

We study general-sum Markov games (MGs, also called stochastic games in Shapley (1953)) with reward delays. In its episodic and tabular form, an MG can be defined by the following tuple:

$$\mathcal{MG}\Big(H, \mathcal{S}, \{\mathcal{A}_m\}_{m\in[M]}, \{\mathbb{P}_h\}_{h\in[H]}, \{r_{m,h}\}_{m\in[M],h\in[H]}, \{d^n_{m,h}(s)\}_{m\in[M],h\in[H],s\in\mathcal{S},n\in[K]}\Big). \quad (1)$$

Here we use $[m]$ to denote the set of $\{1,\ldots,m\}$, and in the subscripts, $m \in [M]$ stands for the agents, $k \in [K]$ stands for the episode, $h \in [H]$ stands for the time step over the finite horizon. $\mathcal{S}$ is a global state space with cardinality $S = |\mathcal{S}|$. $\mathcal{A}_m$ is the action space of agent m. Define $A = \max_{m\in[M]} |\mathcal{A}_m|$. The joint action space is given by $\boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_M$, and the joint action is given as $\boldsymbol{a} = (a_1,\ldots,a_M)$. $\mathbb{P}_h(s'|s,\boldsymbol{a})$ with $s, s' \in \mathcal{S}$, $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$ is the transition function for step h. $r_{m,h}(s,\boldsymbol{a})$ is a deterministic reward for agent m at step h when the current state and joint action are s and $\boldsymbol{a}$. The sequence $\{d^n_{m,h}(s)\}_{m\in[M],h\in[H],s\in\mathcal{S},n\in\mathbb{N}}$ represents the delays of the rewards which will be introduced in detail in later paragraphs. Without loss of generality, we assume that every episode k starts from a fixed initial state s_1.[1] At every step h, every agent observes the current state s^k_h, takes action $a^k_{m,h}$. The environment then transits to the next state s^k_{h+1} according to $\mathbb{P}_h$ until step $H+1$ is reached.

Visits and happening order: When the agents visit state s at step h for the n-th time, we say that the n-th visit of (h,s) happens, and n is the happening order of this visit.

Delays: In this paper, we allow the delays of the rewards to be heterogeneous among different agents m, different visits (h,s) and different happening orders n. In specific, for the n-th visit of (h,s) which happens at episode k, agent m will receive its reward $r_{m,h}(s,\boldsymbol{a})$ by the end of episode $k + d^n_{m,h}(s)$. When $d^n_{m,h}(s) = 0$, our model is the classical MG where the reward $r_{m,h}(s,\boldsymbol{a})$ is received by the end of k.

(Un)received visits, (un)usable visits: When the reward of a visit has been received, we call the visit a received visit; otherwise, we call it an unreceived visit. It is worth mentioning that visits that happen early are not necessarily received early.

We denote the episode in which the n-th visit of (h,s) happens as $k^n_h(s)$. At the beginning of episode $k^n_h(s)$,[2] for agent m, because of the reward delays, some of the first $n-1$ visits of (h,s) may not be received. In this case, we define index $e^n_{m,h}(s)$ as the earliest unreceived visit:

$$e^n_{m,h}(s) := \min\Big\{j : d^j_{m,h}(s) + k^j_h(s) > k^n_h(s) - 1, j \in [n-1]\Big\}. \tag{2}$$

If all of the first $n-1$ visits have been received, we define

$$e^n_{m,h}(s) = n. \tag{3}$$

It means that all the visits of (h,s) that happen earlier than the $e^n_{m,h}(s)$-th visit have been received before episode $k^n_h(s)$ starts; but the $e^n_{m,h}(s)$-th visit has not been received yet.

We call a received visit as *usable* if and only if all visits happening earlier have all been received. Before episode $k^n_h(s)$ starts, the usable visits of (h,s) would be the visits that happen before the $e^n_{m,h}(s)$-th visit, i.e., visits with happening order $1, 2, \ldots, e^n_{m,h}(s) - 1$.

In our algorithm, to ensure that the agents are aligned, we only use some of the usable visits at every episode. Consequently, the performance of our algorithm strongly relates to the number of unusable and unreceived visits. We define a counting sequence $\{\mathcal{T}^n_{m,h}(s)\}_{m\in[M],h\in[H],s\in\mathcal{S},n\in[K]}$ as:

$$\mathcal{T}^n_{m,h}(s) := \sum_{i=1}^{n}\left(i - e^i_{m,h}(s)\right) = \sum_{i=1}^{n}\left(i - \min\Big\{j : d^j_{m,h}(s) + k^j_h(s) > k^i_h(s) - 1\Big\}\right), \tag{4}$$

where $\mathcal{T}^n_{m,h}(s)$ counts the accumulated number of unusable and unreceived visits of (h,s) till the n-th visit. Note that in the classical MG setting without reward delays, we have $\mathcal{T}^n_{m,h}(s) = 0$.

1. For any MG with initial distribution μ, one can always add a step with only one state ahead and let the transition function be μ for all actions, leading to an equivalent MG with a fixed initial state.
2. Without causing any confusion, we will use "at the beginning of episode k" and "by the end of episode $k-1$" interchangeably.

2.2. Objectives - Coarse Correlated Equilibrium

Agent m's policy is denoted as $\pi_m = \{\pi_{m,h}\}_{h\in[H]}$. The policy at step h is $\pi_{m,h} : \Omega \times (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S} \to \Delta_{\mathcal{A}_m}$, where $\pi_{m,h}$ maps a random sample ω_h from probability space Ω and a trajectory $(s_1, \boldsymbol{a}_1, \ldots, s_h)$ to a point in probability simplex $\Delta_{\mathcal{A}_m}$. An important subclass of policy is the *independent Markov policy*, with $\pi_{m,h} : \mathcal{S} \to \Delta_{\mathcal{A}_m}$ maps the current state to a point in probability simplex $\Delta_{\mathcal{A}_m}$.

A *joint policy* π is a set of policies $\{\pi_m\}_{m\in[M]}$ of all agents. If the random samples $\{\omega_h \in \Omega\}_{h\in[H]}$ are shared among all agents, policies of all agents are correlated. In this case, we denote the joint policy π as $\pi = \pi_1 \odot \pi_2 \odot \ldots \odot \pi_M$, and call π as a correlated policy. We also use $\pi_{-m} = \pi_1 \odot \ldots \pi_{m-1} \odot \pi_{m+1} \ldots \odot \pi_M$ to denote the policy excluding agent m. If the randomness of π_m is independent of other policies π_{-m}, i.e., the random samples $\{\omega_h \in \Omega\}_{h\in[H]}$ are shared among agents except agent m, we denote the joint policy as $\pi = \pi_m \times \pi_{-m}$.

For a joint policy π, we define its value function for agent m as:

$$V^{\pi}_{m,h}(s_h) := \mathbb{E}_{\pi}\Big[\sum_{h'=h}^{H} r_{m,h'}(s_{h'}, \boldsymbol{a}_{h'})|s_h\Big], \ \ \forall m \in [M]. \tag{5}$$

Given policy π_{-m}, the best response for agent m is defined as the best policy that maximizes the value function for agent m, i.e., $\pi^{\dagger}_m = \arg\max_{\pi_m} V^{\pi_m \times \pi_{-m}}_{m,1}$. For notation simplicity, we denote the value function of the best response as $V^{\dagger,\pi_{-m}}_{m,h} = V^{\pi^{\dagger}_m,\pi_{-m}}_{m,h}$. Our objective is to find a joint policy π that is an ϵ-coarse correlated equilibrium (CCE) defined as follows:

Definition 1 (Coarse Correlated Equilibrium (CCE (Young, 2004))) *We define the CCE-gap of a joint policy π as:*

$$\textit{CCE-gap}(\pi) := \max_{m\in[M]} (V^{\dagger,\pi_{-m}}_{m,1} - V^{\pi}_{m,1})(s_1). \tag{6}$$

A joint policy π is a CCE if the CCE-gap is zero:

$$\textit{CCE-gap}(\pi) = 0. \tag{7}$$

A joint policy π is an ϵ-CCE if the CCE-gap satisfies:

$$\textit{CCE-gap}(\pi) \leq \epsilon. \tag{8}$$

When agents reach a CCE, they have no incentive to deviate to any independent policies.

3. Delay-Adaptive Multi-Agent V-Learning

In this section, we present our main algorithm: Delay-Adaptive Multi-Agent V-Learning (DA-MAVL). Similar to V-learning in Jin et al. (2021); Song et al. (2021), DA-MAVL contains two consecutive algorithms - i) the training algorithm (Algorithm 1), where agents learn and store a set of independent Markov policies $\{\hat{\pi}^k_{m,h}\}_{m\in[M],h\in[H],k\in[K]}$, and ii) the output algorithm (Algorithm 2), which constructs the final output policies (which can be correlated and non-Markov) $\{\pi_m\}_{m\in[M]}$ from the set of independent Markov policies $\{\hat{\pi}^k_{m,h}\}_{m\in[M],h\in[H],k\in[K]}$. The training algorithm is *decentralized*, i.e., agents update their own policies with their own delayed rewards and without communication between agents. The algorithm framework resembles the V-learning algorithm but comes with a mechanism that carefully chooses *usable* visits for learning. This mechanism enables agents to align their behaviour under the influence of heterogeneous reward delays and leads the algorithm toward convergence (see more discussions at the end of next subsection).

Recall that $k_h^n(s)$ is the episode when the agents visit (h, s) for the n-th time. For agent m, we also define $\overline{n}_{m,h}^k(s)$ as the count of happened visits of (h, s) and define $\underline{n}_{m,h}^k(s)$ as the count of usable visits of (h, s) at the beginning of episode k.

3.1. The Training Algorithm

We now present the training algorithm for DA-MAVL (Algorithm 1). The algorithm contains three major processes, which we name as 'Preparation', 'Learning' and 'Sampling'. At each episode k, for every agent m and (h, s), the three processes are carried out iteratively through episodes:

- In the 'Preparation' process, we keep track of three important sets, namely the set of *visits to be used* $\mathcal{F}_{m,h}(s)$ (including all usable visits that have not been used previously), the set of *unusable visits* $\mathcal{M}_{m,h}^+(s)$ and the set of *unreceived visits* $\mathcal{M}_{m,h}^-(s)$. Usable Visits in $\mathcal{F}_{m,h}(s)$ will be fed into later processes and will no longer be used again in future episodes. Unusable and unreceived visits in $\mathcal{M}_{m,h}(s) = \mathcal{M}_{m,h}^+(s) \cup \mathcal{M}_{m,h}^-(s)$ are stored in memory until they become usable.

  Note that for set $\mathcal{M} = \mathcal{M}_{m,h}(s)$ (or $\mathcal{M} = \mathcal{M}_{m,h}^-(s)$), whose entries are tuples $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r)$ (or $(i, a, \hat{\pi}, \overline{V}', \underline{V}')$) indexed by the first element i, we define $\arg\{\mathcal{M}\} := \{i\}$ as the set of indices.

- In the 'Learning' process, visits in $\mathcal{F}_{m,h}(s)$ are fed into subroutines 'VALUE_UPDATE' and 'POLICY_OPT' (Algorithm 3 and Algorithm 4 in Appendix C.1 in Zhang et al. (2022)) consecutively in their happening orders. Subroutine 'VALUE_UPDATE' keeps track of an "optimistic" value estimate $\overline{V}_{m,h}(s)$ by using all visits in $\mathcal{F}_{m,h}(s)$ with parameters α_i and $\overline{\beta}_{m,h}^i(s)$. Note that α_i can be viewed as the learning rate and $\overline{\beta}_{m,h}^i(s)$ is a bonus term. Subroutine 'POLICY_OPT' runs an adversarial-bandit-type algorithm (similar to the algorithm in Zimmert and Seldin (2020)) to update the policy, where the bandit loss is calculated using the optimistic value estimates $\overline{V}_{m,h}(s)$.

  Note that in Algorithm 1 and Subroutine 'VALUE_UPDATE', we also introduce a pessimistic value estimate, $\underline{V}_{m,h}(s)$. This pessimistic estimate is an auxiliary variable that is not needed for running the algorithm but is used in the proof.

- In the 'Sampling' process, every agent chooses its action based on the updated policy, and the next state is sampled. Finally, every agent stores related information, receives delayed rewards and moves on to the next step $h + 1$.

Discussions - The role of usable visits. As previously mentioned, one key challenge for the decentralized learning algorithm is to avoid misalignment under heterogeneous reward delays of different agents. Our algorithm addresses this challenge by only using usable visits for learning in subroutines 'VALUE_UPDATE' and 'POLICY_OPT'. The main intuition is to ensure that the *happening order* of the visits is also *the order in which they are used in the subroutines*. Consequently, although the reward of a visit might be received and used in different episodes for different agents, the orders in which the visits are used remain the same among agents. This enables agents to keep aligned without any communication during the training algorithm and leads to better performance.

To better understand the role of usable visits, we also compare our algorithm numerically with the naive algorithm, where visits are immediately fed into subroutines once they are received (see Appendix C.2 in Zhang et al. (2022) for details). Notice that in the naive algorithm, the reward of the same visit may be used in different orders among agents, which causes misalignment of the agents. The numerical results are discussed in Section 6, where we indeed observe that with the notion of usable visits, our algorithm outperforms the naive algorithm. Additionally, we also tried to analyze the performance of the naive method, yet failed to give a rigorous bound due to the misalignment caused by the heterogeneity of the delay. However, it remains an open question to prove or to disapprove whether the naive method would converge to a CCE.

Algorithm 1: DA-MAVL Training for Agent m

Init: $\forall(h,s), \overline{n}^0_{m,h}(s) \leftarrow 0, \underline{n}^0_{m,h}(s) \leftarrow 0, \mathcal{T}^0_{m,h}(s) \leftarrow 0, \mathcal{F}_{m,h}(s) \leftarrow \emptyset, \mathcal{M}_{m,h}(s) \leftarrow \emptyset$;

for *Episode* $k = 1, \ldots, K$ **do**
  Receive initial state s^k_1;
  for *Step* $h = 1, \ldots, H$ **do**
    // Preparation
    $s \leftarrow s^k_h$;
    for $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r) \in \mathcal{M}^+_{m,h}(s)$ **do**
      if $\forall j < i, j \notin \arg\{\mathcal{M}^-_{m,h}(s)\}$ **then**
        Save $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r)$ to $\mathcal{F}_{m,h}(s)$; Remove $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r)$ from $\mathcal{M}^+_{m,h}(s)$;
    $\overline{n} \leftarrow \overline{n}^k_{m,h}(s) = \overline{n}^{k-1}_{m,h}(s) + 1$; $\underline{n} \leftarrow \underline{n}^k_{m,h}(s) = \underline{n}^{k-1}_{m,h}(s) + |\mathcal{F}_{m,h}(s)|$;
    $\mathcal{T}^{\overline{n}}_{m,h}(s) \leftarrow \mathcal{T}^{\overline{n}-1}_{m,h}(s) + |\mathcal{M}_{m,h}(s)|$;
    // Learning
    $\overline{V}^k_{m,h}(s), \underline{V}^k_{m,h}(s) \leftarrow \text{VALUE\_UPDATE}_{m,h,s}\big(\mathcal{F}_{m,h}(s), \underline{n}\big)$;
    $\hat{\pi}^k_{m,h}(\cdot|s) \leftarrow \text{POLICY\_OPT}_{m,h,s}\big(\mathcal{F}_{m,h}(s), \overline{n}\big)$;
    // Sampling
    Take action $a^k_{m,h} \sim \hat{\pi}^k_{m,h}(\cdot|s)$; Observe next state s^k_{h+1};
    for $s' \in \mathcal{S}\backslash s$ **do**
      $\overline{n}^k_{m,h}(s') \leftarrow \overline{n}^{k-1}_{m,h}(s')$; $\underline{n}^k_{m,h}(s') \leftarrow \underline{n}^{k-1}_{m,h}(s')$;
      $\overline{V}^k_{m,h}(s') \leftarrow \overline{V}^{k-1}_{m,h}(s')$; $\underline{V}^k_{m,h}(s') \leftarrow \underline{V}^{k-1}_{m,h}(s')$; $\hat{\pi}^k_{m,h}(\cdot|s') \leftarrow \hat{\pi}^{k-1}_{m,h}(\cdot|s')$;
  for *Step* $h = 1, \ldots, H$ **do**
    Save $\big(\overline{n}^k_{m,h}(s^k_h), a^k_{m,h}, \hat{\pi}^k_h(a^k_{m,h}|s^k_h), \overline{V}^k_{m,h+1}(s^k_{h+1}), \underline{V}^k_{m,h+1}(s^k_{h+1})\big)$ to $\mathcal{M}^-_{m,h}(s^k_h)$;
  Receive delayed rewards for all states s;
  for *Delayed Reward* (m, h, s, i, r) **do**
    Extract and remove $\big(i, a, \hat{\pi}, \overline{V}', \underline{V}'\big)$ from $\mathcal{M}^-_{m,h}(s)$;
    Save $\big(i, a, \hat{\pi}, \overline{V}', \underline{V}', r\big)$ to $\mathcal{M}^+_{m,h}(s)$;

3.2. Execution of the Output Policy

Algorithm 1 outputs a set of independent Markov policies $\{\hat{\pi}^k_{m,h}\}_{m\in[M],h\in[H],k\in[K]}$. Based on this policy set, we now construct joint policy $\pi = \{\pi_m\}_{m\in[M]}$ as the output of DA-MAVL. The policy is defined by its execution in Algorithm 2. Notice that all random samples (line 1 and line 4) are shared across all agents. This algorithm mainly follows V-learning in Jin et al. (2021); Song et al. (2021) except for the modification in line 3. It ensures the estimated value functions in Algorithm 1 upper bound the policy performance. We refer readers to the original work for more intuitions.

4. Performance Guarantee and Proof Sketches

Recall that the counting sequence $\{\mathcal{T}^n_{m,h}(s)\}_{m\in[M],h\in[H],s\in\mathcal{S},n\in[K]}$ (Equation (4)) is agent m's accumulated count of unusable and unreceived visits till the n-th visit of (h,s). Also, recall that $\underline{n}^k_{m,h}(s)$

Algorithm 2: DA-MAVL Output for π_m

Sample $k \sim \text{Uniform}([K])$;
for *step* $h = 1, \ldots, H$ **do**
  Observe current state s_h; $n \leftarrow \max_m \underline{n}^k_{m,h}(s_h)$;
  Sample i from $[n]$ with probability α^i_n; $k \leftarrow k^i_h(s_h)$;
  Take action $a_{m,h} \sim \hat{\pi}^k_{m,h}(\cdot|s_h)$;

is the count of usable visits of (h, s) at the beginning of episode k. Using the two notations, we define $\mathcal{T}_K := \max_{m,h} \sum_{s\in\mathcal{S}} \mathcal{T}_{m,h}^{\underline{n}^K_{m,h}(s)}(s)$ which will be used in bounding the CCE-gap after K episodes. We will assume that the reward delays of the MG are upper bounded by some constant.

Assumption 1 *The delays are bounded by d_{max}, that is,* $\max_{m\in[M],h\in[H],s\in\mathcal{S},n\in[K]} d^n_{m,h}(s) \le d_{max}$.

Now we are ready to present the performance guarantee for DA-MAVL:

Theorem 1 *Under Assumption 1, for any $\delta \in (0, 1)$, $K \ge d^2_{max} S\iota^3$ where $\iota = \log(4MHSAK/\delta)$, suppose Algorithm 1 is run for K episodes, then the following equation holds for the output policy π of Algorithm 2 with probability at least $1 - \delta$*

$$\textit{CCE-gap}(\pi) = \max_{m\in[M]} \left(V^{\dagger,\pi_{-m}}_{m,1} - V^{\pi}_{m,1}\right)(s_1) \lesssim H^3\sqrt{S\mathcal{T}_K/K^2}\iota^2 + H^3\sqrt{SA\iota/K}. \tag{9}$$

Under Assumption 1, it can be shown (with Lemma 38 in Appendix E.1 in Zhang et al. (2022)):

$$\mathcal{T}_K = \max_{m,h} \sum_{s\in\mathcal{S}} \mathcal{T}_{m,h}^{\underline{n}^K_{m,h}(s)}(s) \le \max_{m,h} \sum_{s\in\mathcal{S}} \sum_{n=1}^{\underline{n}^K_{m,h}(s)} \left(n - e^n_{m,h}(s)\right) \le d_{max} \sum_{s\in\mathcal{S}} \overline{n}^K_{m,h}(s) \le K d_{max}.$$

Substituting it into Theorem 1 gives the CCE-gap of order $\tilde{\mathcal{O}}(\frac{H^3\sqrt{Sd_{max}}+H^3\sqrt{SA}}{\sqrt{K}})$. In other words, in the worst case where the delays are always d_{max} and every (h, s) is visited for K times, at most $K = \tilde{\mathcal{O}}(\frac{H^6S(d_{max}+A)}{\epsilon^2})$ episodes are needed for an ϵ-CCE. The influence of the reward delays is linearly bounded by term $\tilde{\mathcal{O}}(\frac{H^6Sd_{max}}{\epsilon^2})$ and tends to 0 when d_{max} goes to 0. Note that our result bears an extra factor H compared with V-learning (Jin et al., 2021; Song et al., 2021), even when all delays are zero. This is because we have to choose the parameters generously so that our algorithm is adaptive to potential delays.

4.1. Proof Sketch of Theorem 1

The proof can be broken down into the following three steps.

STEP 1: Bound the 'Policy Optimization Regret'. For notation simplicity, we let k_n denote $k^n_h(s)$. For every pair (m, h, s, n), we first define the policy optimization regret $R^n_{m,h}(s)$:

$$R^n_{m,h}(s) = \max_{a_m\in\mathcal{A}_m} \sum_{i=1}^{n} \alpha^i_n \left[\mathbb{E}\left(r_{m,h}(s, \boldsymbol{a}) + \overline{V}^{k_i}_{m,h+1}(s')\right) - \left(r^{k_i}_{m,h} + \overline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1})\right)\right], \tag{10}$$

where $\boldsymbol{a} = (a_m, \boldsymbol{a}_{-m})$, α_n^i is the weight which we define in Equation (22) in Appendix D in Zhang et al. (2022), and the expectation is taken over $\boldsymbol{a}_{-m} \sim \hat{\pi}_{-m,h}^{k_i}(\cdot|s)$ and $s' \sim \mathbb{P}_h(\cdot|s, \boldsymbol{a})$. Intuitively, it measures the performance of the first n outputs of subroutine 'POLICY_OPT$_{m,h,s}$' in Algorithm 1, i.e. Markov policies $\{\hat{\pi}_{m,h}^{k_i}(s)\}_{i\in[n]}$. Under Assumption 1, we give the following upper bound:

Lemma 1 *Suppose Assumption 1 holds. For* $\forall(m, h, s, k) \in [M] \times [H] \times \mathcal{S} \times [K]$*, the following inequality holds with probability at least* $1 - \delta/2$

$$R_{m,h}^n(s) \leq 12H^2\sqrt{\frac{nA + \mathcal{T}_{m,h}^n(s)}{n^2}\iota} + 2H^2\frac{d_{max}}{n}\iota. \tag{11}$$

In this lemma, the key difference from V-learning is that the subroutine needs to learn the n-th output, i.e. $\hat{\pi}_{m,h}^{k_n}(s)$, without access to all reward information of the first $n-1$ visits of (h, s) due to the reward delays. We have to measure the influence of the delays on outputs. By comparing it with the no-delay versions, we can show that the influence of the delays can be reflected by term $\sqrt{\mathcal{T}_{m,h}^n(s)/n^2}$ and d_{max}/n in Equation (11).

STEP 2: Optimism and Pessimism. Utilizing the regret defined above, we carefully design bonuses $\overline{\beta}_{m,h}^n(s)$ and $\underline{\beta}_n$ in subroutine 'VALUE_UPDATE' as follows:

$$\overline{\beta}_{m,h}^n(s) = R_{m,h}^n(s) + 2H^2\frac{d_{max}}{n}\iota, \quad \underline{\beta}_n = 2\sqrt{\frac{H^3}{n}\iota} + 2H^2\frac{d_{max}}{n}\iota. \tag{12}$$

With the bonuses, we can show that the value estimates $\overline{V}_{m,h}^k(s)$ and $\underline{V}_{m,h}^k(s)$ in Algorithm 1 upper and lower bound the performance of policy $\pi_{m,h}^k$.

Lemma 2 *Suppose Assumption 1 holds. For* $\forall(m, h, s, k) \in [M] \times [H] \times \mathcal{S} \times [K]$*, the following inequality holds with probability at least* $1 - \delta$

$$\overline{V}_{m,h}^k(s) \geq V_{m,h}^{\dagger,\pi_{-m,h}^k}(s), \quad \underline{V}_{m,h}^k(s) \leq V_{m,h}^{\pi_h^k}(s). \tag{13}$$

In this lemma, policy $\pi_h^k(s)$ can be seen as part of the output policy π in Algorithm 2, which is used at steps from h to H. It is formally defined in Algorithm 12 in Appendix D in Zhang et al. (2022).

We note that it is nontrivial to ensure optimism and pessimism under the influence of heterogeneous reward delays among agents. Notice that $\overline{V}_{m,h}^k(s)$ and $\underline{V}_{m,h}^k(s)$ are calculated only with information of agent m. However, the output policy π_h^k, as in $V_{m,h}^{\dagger,\pi_{-m,h}^k}(s)$ and $V_{m,h}^{\pi_h^k}(s)$, is a *correlated* policy that takes information of all agents into consideration. Therefore, under heterogeneous delays, it is challenging for $\overline{V}_{m,h}^k(s)$ and $\underline{V}_{m,h}^k(s)$ to upper or lower bound $V_{m,h}^{\dagger,\pi_{-m,h}^k}(s)$ and $V_{m,h}^{\pi_h^k}(s)$, breaking the original optimism and pessimism results in V-learning (Jin et al., 2021; Song et al., 2021). We carefully design bonuses as in Equation 13 to solve this difficulty and ensure optimism and pessimism.

STEP 3: Bound the CCE-gap. Finally, given Lemma 2, it suffices to bound the gap between the optimistic and pessimistic value estimates $\sum_{k=1}^{K}(\overline{V}_{m,1}^k - \underline{V}_{m,1}^k)(s_h^k)$.

As is mentioned in Step 2, the value estimates $\overline{V}_{m,h}^k(s)$ and $\underline{V}_{m,h}^k(s)$ are calculated without access to all information due to the reward delays. This fact increases the variance of the value estimates. In the proof of this lemma, we carefully consider the number of unreceived and unusable visits for every episode and analyze its cumulative influence across all episodes.

5. Extension to Infinite Delays

5.1. The Skipping Scheme

The performance of the DA-MAVL algorithm in Section 3 heavily relies on the assumption that delays are finite. One single infinite delay could prevent the algorithm from convergence because all visits that happen later are unusable. In this case, it is worth skipping some of the rewards for better performance. Following the intuitions of Zimmert and Seldin (2020), we extend DA-MAVL and design a new skipping metric to deal with infinite delays in MARL. Details for the extended algorithm (DA-MAVL with Reward Skipping) are presented in Appendix C.3 in Zhang et al. (2022).

The critical part of the 'Skipping' process is to determine when to skip a visit. When the n-th visit of (h, s) happens, we maintain the skipping metric $\phi_{m,h}^{i,n}(s) = \sum_{j=i+1}^{n}(j-i)$ if the i-th visit of (h, s) is unreceived. Intuitively speaking, $\phi_{m,h}^{i,n}(s)$ upper bounds the contribution of the i-th visit to $\mathcal{T}_{m,h}^{n}(s)$. It is beneficial to skip the i-th visit if $\phi_{m,h}^{i,n}(s)$ becomes large enough. Following the intuition of previous reward skipping method in Zimmert and Seldin (2020) in the adversarial bandit setting, we skip the i-th visit if $\phi_{m,h}^{i,n}(s)$ exceeds threshold $\sqrt{\mathcal{T}_{m,h}^{n}(s)}$. However, we would like to point out that our design of the skipping metric $\phi_{m,h}^{i,n}(s)$ is not a direct generalization of previous skipping method. Unlike the multi-agent setting considered in this paper, the adversarial bandit setting does not need to consider the heterogeneity of reward delays among agents, thus their algorithm update does not need to wait for visits to become usable. Correspondingly, the skipping metric $n-i$ in previous method would fail in our setting, because it no longer upper-bounds the contribution of the i-th visit to $\mathcal{T}_{m,h}^{n}(s)$.

5.2. Performance Guarantee for DA-MAVL with Reward Skipping

Recall the notation $k_h^n(s)$ stands for the episode when n-th visit of (h, s) happens. With the skipping scheme, we can also relax Assumption 1 to the following:

Assumption 2 *For $\forall(m,h,s,n) \in [M]\times[H]\times\mathcal{S}\times[K]$, there exists a constant C satisfying:*

$$|\{i \le n : d_{m,h}^{i}(s) + k_h^i(s) \ge k_h^n(s)\}| \le C. \tag{14}$$

Intuitively, Assumption 2 requires that for every pair (m, h, s, n), there are at most C unreceived visits before the n-th visit of $(h, s)$for agent m. This implies that either large delays do not appear too many times or delays are not large enough to influence performance. It is worth noting that the finite delay Assumption 1 implies Assumption 2 with $C = d_{max}$. But Assumption 2 is more general than Assumption 1 because Assumption 2 holds even if there are less than C infinite delays.

Given a subset of visit indices $\mathcal{L} \subset [K]$, at episode $k_h^n(s)$ when the n-th visit of (h, s) happens, we define variable $e_{m,h}^{n,\mathcal{L}}(s)$ as the earliest unreceived visit outside of $\mathcal{L}$:

$$e_{m,h}^{n,\mathcal{L}}(s) := \min\left\{j : d_{m,h}^{j}(s) + k_h^j(s) > k_h^n(s) - 1, j \in [n-1]\backslash\mathcal{L}\right\}. \tag{15}$$

If all of the first $n-1$ visits are received, we let $e_{m,h}^{n,\mathcal{L}}(s) = n$. Now we define $\mathcal{T}_{m,h}^{n,\mathcal{L}}(s)$ as follows:

$$\mathcal{T}_{m,h}^{n,\mathcal{L}}(s) := \sum_{i=1}^{n} i - e_{m,h}^{i,\mathcal{L}}(s). \tag{16}$$

It counts the accumulated number of unusable and unreceived visits outside of $\mathcal{L}$ for the first n visits of (h, s). Finally, we also define $\mathcal{T}_{m,h}^{K,\mathcal{L}} := \sum_{s\in\mathcal{S}} \mathcal{T}_{m,h}^{\overline{n}_{m,h}^{K}(s),\mathcal{L}}(s)$. Intuitively, it counts the accumulated number of unusable and unreceived visits outside of $\mathcal{L}$ during the K episodes.

Now we are ready to present the performance guarantee for DA-MAVL with Reward Skipping:

Theorem 2 *Under Assumption 2, for any $\delta \in (0,1)$, $K \geq C^6 S^3 \iota^3$ where $\iota = \log(4MHSAK/\delta)$, suppose Algorithm 9 is run for K episodes, then the following equation holds for the output policy π of Algorithm 2 with probability at least $1-\delta$*

$$\textit{CCE-gap}(\pi) \lesssim CH^3 \max_{m,h} \min_{\mathcal{L}} \left\{ \frac{S|\mathcal{L}|}{K} + \sqrt{\frac{S\mathcal{T}_{m,h}^{K,\mathcal{L}}}{K^2}} \right\} \iota^2 + H^3 \sqrt{\frac{SA}{K}} \iota. \tag{17}$$

Theorem 2 implies that DA-MAVL with Reward Skipping can still obtain convergence to CCE when there are infinite delays. Consider the case where all delays are upper bounded by constant d_{max}, except for C infinite delays for every (h, s). Let $\mathcal{L}_{m,h} = \{n : \exists s, d_{m,h}^n(s) = \infty\}$ denote all visit indices where the delay is infinite for some state s and fixed pair (m, h). We then have $|\mathcal{L}_{m,h}| \leq CS$ and $\mathcal{T}_{m,h}^{K,\mathcal{L}_{m,h}} \leq K d_{max}$. Substituting into Theorem 2 gives CCE-gap of order $\tilde{\mathcal{O}}(\frac{H^3\sqrt{Sd_{max}}+H^3\sqrt{SA}}{\sqrt{K}})$, which is exactly the same as the result of Theorem 1.

6. Simulations

We simulate our algorithms in a simple single-state MG with $M = 3, S = 1, A = 2, H = 1$. Due to the space limit, the simulation settings are introduced in Appendix B in Zhang et al. (2022). We only present the simulation results in Figure 1. We can see that our algorithm outperforms the naive algorithm (mentioned in Section 3) when delays are finite. Moreover, our novel skipping metric outperforms previous skipping method (mentioned in Section 5.1) when delays are infinitely large.

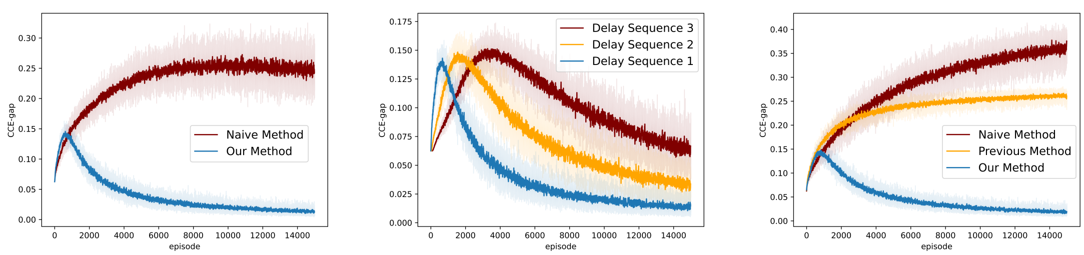

Figure 1: **Left**: CCE-gap for Output Policy of DA-MAVL (Our Method) and the Naive Algorithm (Naive Method); **Center**: CCE-gap for Output Policy of DA-MAVL under Different Delay Sequences; **Right**: CCE-gap for Skipping Metrics in DA-MAVL with Reward Skipping (Our Method), in Previous Work (Zimmert and Seldin, 2020) (Previous Method) and No Skipping (Naive Method).

7. Conclusion

This paper studies MARL with reward delays. For finite delays, we propose MARL algorithms with a novel mechanism to choose proper visits for learning, so that agents can reach a CCE even when facing heterogeneous delays. We also adapt our algorithm to cases with infinite delays using a novel reward skipping metric. High probability bounds are given on the CCE-gap of our algorithms. There are many interesting future directions, such as proving or disproving the convergence of the naive algorithm (Appendix C.2 in Zhang et al. (2022)), providing lower bounds on the CCE-gap for MARL with reward delays, relaxing Assumption 2 for infinite delays, extending current results to MGs with function approximation, etc.

References

Mridul Agarwal and Vaneet Aggarwal. Blind decision making: Reinforcement learning with delayed observations. *Pattern Recognition Letters*, 150:176–182, 2021.

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Yann Bouteiller, Simon Ramstedt, Giovanni Beltrame, Christopher Pal, and Jonathan Binas. Reinforcement learning with random delays. In *International conference on learning representations*, 2020a.

Yann Bouteiller, Simon Ramstedt, Giovanni Beltrame, Christopher Pal, and Jonathan Binas. Reinforcement learning with random delays. In *International conference on learning representations*, 2020b.

Manuele Brambilla, Eliseo Ferrante, Mauro Birattari, and Marco Dorigo. Swarm robotics: a review from the swarm engineering perspective. *Swarm Intelligence*, 7(1):1–41, 2013.

Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR, 2016.

Baiming Chen, Mengdi Xu, Zuxin Liu, Liang Li, and Ding Zhao. Delay-aware multi-agent reinforcement learning for cooperative and competitive environments. *arXiv preprint arXiv:2005.05441*, 2020.

Young-Cheol Choi and Hyo-Sung Ahn. A survey on multi-agent reinforcement learning: Coordination problems. In *Proceedings of 2010 IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications*, pages 81–86. IEEE, 2010.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

Peihu Duan, Lidong He, Zhisheng Duan, and Ling Shi. Distributed cooperative lqr design for multi-input linear systems. *IEEE Transactions on Control of Network Systems*, 2022.

Vlad Firoiu, Tina Ju, and Josh Tenenbaum. At human speed: Deep reinforcement learning with action delay. *arXiv preprint arXiv:1810.07286*, 2018.

Manegueu Anne Gael, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356. PMLR, 2020.

Andras Gyorgy and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*, pages 3988–3997. PMLR, 2021.

Yann-Michaël De Hauwere, Peter Vrancx, and Ann Nowé. Solving sparse delayed coordination problems in multi-agent reinforcement learning. In *International Workshop on Adaptive and Learning Agents*, pages 114–133. Springer, 2011.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret for adversarial mdp with delayed bandit feedback. *arXiv preprint arXiv:2201.13172*, 2022.

Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-) convex optimization: Optimism, composite objectives, and variational bounds. In *International Conference on Algorithmic Learning Theory*, pages 681–720. PMLR, 2017.

Konstantinos V Katsikopoulos and Sascha E Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE transactions on automatic control*, 48(4):568–574, 2003.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Learning adversarial markov decision processes with delayed feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7281–7289, 2022.

Tianhao Li, Zhishun Wang, Wei Lu, Qian Zhang, and Dengfeng Li. Electronic health records based reinforcement learning for treatment optimizing. *Information Systems*, 104:101878, 2022.

Michael L Littman et al. Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pages 322–328, 2001.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

Andreja Malus, Dominik Kozjek, et al. Real-time order dispatching for a fleet of autonomous mobile robots using multi-agent reinforcement learning. *CIRP annals*, 69(1):397–400, 2020.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, pages 1–22, 2022.

Diogo Martinho, João Carneiro, José Neves, Paulo Novais, Juan Corchado, and Goreti Marreiros. A reinforcement learning approach to improve user achievement of health-related goals. In *EPIA Conference on Artificial Intelligence*, pages 266–277. Springer, 2021.

Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. *Advances in Neural Information Processing Systems*, 23, 2010.

Kun Shao, Zhentao Tang, Yuanheng Zhu, Nannan Li, and Dongbin Zhao. A survey of deep reinforcement learning in video games. *arXiv preprint arXiv:1912.10944*, 2019.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

Jayakumar Subramanian, Amit Sinha, and Aditya Mahajan. Robustness and sample complexity of model-based marl for general-sum markov games. *arXiv preprint arXiv:2110.02355*, 2021.

Hongyao Tang, Jianye Hao, Tangjie Lv, Yingfeng Chen, Zongzhang Zhang, Hangtian Jia, Chunxu Ren, Yan Zheng, Zhaopeng Meng, Changjie Fan, et al. Hierarchical deep multiagent reinforcement learning with temporal abstraction. *arXiv preprint arXiv:1809.09332*, 2018.

Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems*, 18(1):83–105, 2009.

Hang Yin, Yu Wang, Xukai Zhang, and Peng Li. Feedback delay impaired reinforcement learning: Principal components analysis of reward positivity. *Neuroscience letters*, 685:179–184, 2018.

H Peyton Young. *Strategic learning and its limits*. OUP Oxford, 2004.

Yuyang Zhang, Runyu Zhang, Gen Li, Yuantao Gu, and Na Li. Multi-agent reinforcement learning with reward delays. In *Online report,* `https://nali.seas.harvard.edu/files/nali/files/damavl.pdf`, 2022.

Xiangyu Zhao, Long Xia, Lixin Zou, Hui Liu, Dawei Yin, and Jiliang Tang. Whole-chain recommendations. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1883–1891, 2020.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pages 3285–3294. PMLR, 2020.

Appendix A. Related Work

MARL algorithms with theoretical guarantees. In MARL, it is a standard objective to find CCEs of the underlying MG (Jin et al., 2021). There is a recent line of work providing non-asymptotic guarantees for learning CCEs of general-sum MGs. Generally speaking, existing model-based algorithms (Subramanian et al., 2021; Liu et al., 2021) suffer from the curse of dimensionality in finding CCEs (Jin et al., 2021; Song et al., 2021). They require samples exponentially related to the number of agents to achieve the learning objective. Recent model-free algorithms (Jin et al., 2021; Song et al., 2021; Daskalakis et al., 2022; Mao and Başar, 2022) have successfully broken the curse, providing sample complexity not directly related to the number of agents. V-Learning in Jin et al. (2021); Song et al. (2021) is one of the algorithms that achieve this breakthrough.

Another common objective in this setting is the Nash Equilibrium. However, it is proven PPAD-hard by previous work (Daskalakis et al., 2009).

Delays in MARL. Different kinds of delays may occur in MARL, including but not restricted to state delays (or observation delays), action delays and reward delays (or feedback delays) (Agarwal and Aggarwal, 2021). We acknowledge that there exist lines of work on state and action delays (Agarwal and Aggarwal, 2021; Bouteiller et al., 2020b; Chen et al., 2020), but they are beyond the scope of this paper. In MARL with state delays, the major challenge is how to predict the current state from previous information. In MARL with action delays, the challenge is to predict when the actions will take effect. However, the focus of our paper is to better evaluate the current state and action with available information instead of predicting what they are. Most paper concerning reward delays in MARL is empirical (Shao et al., 2019; Tang et al., 2018; Hauwere et al., 2011). As discussed previously, they resort to alternative mechanisms to guide the agents instead of using rewards directly. Currently, the mechanisms are formed using deep neural networks that are hard to explain theoretically.

Reward delays in MDP. Recently, empirical single-agent RL algorithms have achieved great progress in handling reward delays (Bouteiller et al., 2020a; Arjona-Medina et al., 2019; Yin et al., 2018). In contrast, theoretical aspect of the problem is relatively unexplored. The two available previous work (Lancewicki et al., 2022; Jin et al., 2022) studies adversarial reward delays. Take the latter as an example. For an adversarial MDP with state space size S, action space size A, planning horizon H and total delay D, their algorithm achieves the optimal gap $\tilde{\mathcal{O}}(H^2 S\sqrt{A/K} + H^{5/4}(SA)^{1/4}\sqrt{D/K^2})$ with high probability after K episodes. This result matches ours (Theorem 1) in terms of episode number K and total delay D. However, as mentioned before, these model-based methods maintain exponentially many parameters that prohibit them from being tractable in the multi-agent setting. Our algorithm, based on V-Learning, is a completely different model-free algorithm that successfully breaks the curse.

Reward delays in multi-arm bandit (MAB). There exist extensive theoretic work in MAB that deals with delays. Gyorgy and Joulani (2021); Zimmert and Seldin (2020); Joulani et al. (2017); Gael et al. (2020) tackle delays chosen by adversarial while Cesa-Bianchi et al. (2016); Neu et al. (2010) concern constant delays. Typically, their algorithms gives optimal gap $\tilde{\mathcal{O}}\sqrt{A/K + D/K^2}$, where A is the number of actions, D is the total delay, and K is the number of episodes. All the above literature provides precious insights for this paper. But the MAB setting is completely different from MARL, since there are no state transitions and cooperation between agents.

Appendix B. Simulation Settings

We simulate a simple single-state MG with $M = 3, S = 1, A = 2, H = 1$. The agents visit state s_1. Reward $r = 1$ is given if all agents choose action one, and $r = 0.5$ is given if all agents choose action two. Otherwise, reward $r = 0$.

Finite Delays. We first simulate DA-MAVL (Section 3) with respect to the naive algorithm (Appendix C.2) under delay sequence 1 as follows:

$$d^n_{1,1}(s_1) = 20 - 2 \cdot (i \bmod 10), \quad d^n_{2,1}(s_1) = 5, \quad d^n_{3,1}(s_1) = 5. \tag{18}$$

We plot the CCE-gap $\max_{m\in[M]}(V^{\dagger,\pi^k_{-m,1}}_{m,1} - V^{\pi^k_1}_{m,1})(s_1)$ of every episode k in Figure 1 (left). DA-MAVL (Our Method) achieves satisfying convergence results, which aligns with our intuition in section 3.1. Contrarily, the naive algorithm (Naive Method) fails to converge in limited episodes.

We then show the influence of the delays on DA-MAVL in Figure 1 (center). Delay sequence 2 and delay sequence 3 are four and nine times the value of delay sequence 1. The numeric result matches Theorem 1 in two ways: i). The CCE-gap of our output policy converges to 0; ii). The CCE-gap of our output policy is positively related to d_{max}.

Infinite Delays. For the simulations of infinite delays, we set the delays as follows:

$$d^n_{1,1}(s_1) = \begin{cases} \infty & , n \bmod 10 \leq 5 \\ 0 & , else \end{cases}, \quad d^n_{2,1}(s_1) = 5, \quad d^n_{3,1}(s_1) = 5. \tag{19}$$

Namely, for agent 1, there will be five infinite delays every ten visits of s_1. The numerical results for skipping metric in DA-MAVL with Reward Skipping (Algorithm 9) (Our Method), skipping metric in previous work (Zimmert and Seldin, 2020) (Previous Method) and no skipping scheme in DA-MAVL (Algorithm 1)(Naive Method) is shown in Figure 1 (right). As is suggested in the figure, the algorithm without delay skipping will not converge because of the infinite delays. However, our algorithm with method skips all infinite delays and behaves as if they do not exist, and enjoys superiority over previous skipping methods in the setting of this paper.

Appendix C. Algorithms and Subroutines

C.1. Subroutines for DA-MAVL Training (Algorithm 1)

Algorithm 3: Subroutine VALUE_UPDATE$_{m,h,s}$ for agent m for Algorithm 1

Init: $n \leftarrow 0$, $\tilde{V}_{m,h}(s) \leftarrow H + 1 - h$, $\underset{\sim}{V}_{m,h}(s) \leftarrow 0$; $\overline{V}_{m,h}(s) \leftarrow H + 1 - h$; $\underline{V}_{m,h}(s) \leftarrow 0$;

Receive $\mathcal{F}$, $\underline{n}$;

$\tilde{V}_{m,h}(s) \leftarrow \tilde{V}_{m,h}(s) - \overline{\beta}^n_{m,h}(s)$; $\underset{\sim}{V}_{m,h}(s) \leftarrow \underset{\sim}{V}_{m,h}(s) + \underline{\beta}_n$; $n \leftarrow \underline{n}$;

for $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r) \in \mathcal{F}$ **do**

    $\tilde{V}_{m,h}(s) \leftarrow (1 - \alpha_i)\tilde{V}_{m,h}(s) + \alpha_i(r + \overline{V}')$;

    $\underset{\sim}{V}_{m,h}(s) \leftarrow (1 - \alpha_i)\underset{\sim}{V}_{m,h}(s) + \alpha_i(r + \underline{V}')$;

$\tilde{V}_{m,h}(s) \leftarrow \tilde{V}_{m,h}(s) + \overline{\beta}^n_{m,h}(s)$; $\underset{\sim}{V}_{m,h}(s) \leftarrow \underset{\sim}{V}_{m,h}(s) - \underline{\beta}_n$;

$\overline{V}_{m,h}(s) \leftarrow \min\left\{H + 1 - h, \tilde{V}_{m,h}(s), \overline{V}_{m,h}(s)\right\}$;$\underline{V}_{m,h}(s) \leftarrow \max\left\{0, \underset{\sim}{V}_{m,h}(s), \underline{V}_{m,h}(s)\right\}$;

Algorithm 4: Subroutine POLICY_OPT$_{m,h,s}$ for agent m for Algorithm 1

Init: $\forall a \in \mathcal{A}, \hat{L}_{m,h}(s,a) \leftarrow 0$;
Receive $\mathcal{F}, \overline{n}$;
for $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r) \in \mathcal{F}$ **do**
  for $a' \in \mathcal{A}$ **do**
    $\hat{l}^i_{m,h}(s,a') = \mathbb{I}(a'=a)\left[\frac{H-r-\overline{V}'}{H}\right] \Big/ \left[\hat{\pi} + \gamma^i_{m,h}(s)\right]$;
    $\hat{L}_{m,h}(s,a') = \hat{L}_{m,h}(s,a') + w_i \hat{l}^i_{m,h}(s,a')$;
  $\hat{\pi}_h(\cdot|s) \propto \exp\big(-(\eta^{\overline{n}}_{m,h}(s)/w_{\overline{n}})\hat{L}_{m,h}(s,\cdot)\big)$;
Empty $\mathcal{F}$;

Parameters for the above subroutines are defined as follows:

$$
\alpha_n = \frac{H+1}{H+n},\ n \ge 1, \quad \gamma^n_{m,h}(s) = \eta^n_{m,h}(s) = \sqrt{\frac{\iota}{nA + \mathcal{T}^n_{m,h}(s)}},\ n \ge 1,
$$

$$
\overline{\beta}^n_{m,h}(s) = \begin{cases} 12H^2\sqrt{\frac{nA+\mathcal{T}^n_{m,h}(s)}{n^2}\iota} + 4H^2\frac{d_{max}}{n}\iota, & n \ge 1 \\ 0, & n = 0 \end{cases}, \tag{20}
$$

$$
\underline{\beta}_n = \begin{cases} 2\sqrt{\frac{H^3}{n}\iota} + 2H^2\frac{d_{max}}{n}\iota, & n \ge 1 \\ 0, & n = 0 \end{cases}, \quad w_n = \begin{cases} \alpha_n \prod_{i=2}^n (1-\alpha_i)^{-1}, & n \ge 2 \\ 1, & n = 1 \end{cases},
$$

where $\log(4MHSAK/\delta)$, and $\mathcal{T}^n_{m,h}(s)$ is a parameter maintained by Algorithm 1. Note that here $\underline{\beta}_n$ does not depend on (m,h,s).

C.2. Naive Multi-Agent V-Learning (Naive-MAVL)

Algorithm 5: Naive-MAVL Training for Agent m

Init: $\forall (s,h), \overline{n}^0_{m,h}(s) \leftarrow 0; \underline{n}^0_{m,h}(s) \leftarrow 0; \mathcal{T}^0_{m,h}(s) \leftarrow 0; \mathcal{F}_{m,h}(s) \leftarrow \emptyset; \mathcal{M}^-_{m,h}(s) \leftarrow \emptyset$;
for *Episode* $k = 1, \ldots, K$ **do**
  Receive initial state s^k_1;
  for *Step* $h = 1, \ldots, H$ **do**
    //Preparation
    $s \leftarrow s^k_h; \overline{n} \leftarrow \overline{n}^k_{m,h}(s) = \overline{n}^{k-1}_{m,h}(s) + 1; \underline{n} \leftarrow \underline{n}^k_{m,h}(s) = \underline{n}^{k-1}_{m,h}(s) + |\mathcal{F}_{m,h}(s)|$;
    $\mathcal{T}^{\overline{n}}_{m,h}(s) \leftarrow \mathcal{T}^{\overline{n}-1}_{m,h}(s) + |\mathcal{M}_{m,h}(s)|$;
    // Learning
    $\overline{V}^k_{m,h}(s), \underline{V}^k_{m,h}(s) \leftarrow$ VALUE_UPDATE$_{m,h,s}\big(\mathcal{F}_{m,h}(s), \underline{n}\big)$;
    $\hat{\pi}^k_h(\cdot|s) \leftarrow$ POLICY_OPT$_{m,h,s}\big(\mathcal{F}_{m,h}(s), \underline{n}, \overline{n}\big)$;
    // Sampling
    Take action $a^k_h \sim \hat{\pi}^k_h(\cdot|s)$; Observe next state s^k_{h+1};
    for $s' \in \mathcal{S} \backslash s$ **do**
      $\overline{n}^k_{m,h}(s') \leftarrow \overline{n}^{k-1}_{m,h}(s'); \ \underline{n}^k_{m,h}(s') \leftarrow \underline{n}^{k-1}_{m,h}(s')$;
      $\overline{V}^k_{m,h}(s) \leftarrow \overline{V}^{k-1}_{m,h}(s); \ \underline{V}^k_{m,h}(s) \leftarrow \underline{V}^{k-1}_{m,h}(s); \ \hat{\pi}^k_{m,h}(\cdot|s) \leftarrow \hat{\pi}^{k-1}_{m,h}(\cdot|s)$;

---

for *Step* $h = 1, \ldots, H$ **do**
  Save $\left(\overline{n}^k_{m,h}(s^k_h), a^k_h, \hat{\pi}^k_h(a^k_h|s^k_h), \overline{V}^k_{m,h+1}(s^k_{h+1}), \underline{V}^k_{m,h+1}(s^k_{h+1})\right)$ to $\mathcal{M}^-_{m,h}(s^k_h)$;
Receive delayed reward for all states s;
for *Delayed Reward* (m, h, s, i, r) **do**
  Extract and remove $\left(i, a, \hat{\pi}, \overline{V}', \underline{V}'\right)$ from $\mathcal{M}^-_{m,h}(s)$;
  Save $\left(i, a, \hat{\pi}, \overline{V}', \underline{V}', r\right)$ to $\mathcal{F}_{m,h}(s)$;

Algorithm 6: Execution of Output Policy π_m for Naive-MAVL

Sample $k \sim$ Uniform$([K])$;
for *step* $h = 1, \ldots, H$ **do**
  Observe current state s_h; $n \leftarrow \underline{n}^k_{m,h}(s_h)$;
  Sample i from $[n]$ with probability α^i_n; $k \leftarrow k^i_{m,h}(s_h)$;
  Take action $a_{m,h} \sim \hat{\pi}^k_{m,h}(\cdot|s_h)$;

Notice that the random samples can not be shared across all agents, because the visits are used in different orders for different agents in the subroutines.

Algorithm 7: Subroutine VALUE_UPDATE$_{m,h,s}$ for agent m for Algorithm 5

Init: $n \leftarrow 0$; $\tilde{V}_{m,h}(s) \leftarrow H + 1 - h$; $\underset{\sim}{V}_{m,h}(s) \leftarrow 0$; $\overline{V}_{m,h}(s) \leftarrow H + 1 - h$; $\underline{V}_{m,h}(s) \leftarrow 0$;
Receive $\mathcal{F}, \underline{n}, \overline{n}$;
$\tilde{V}_{m,h}(s) \leftarrow \tilde{V}_{m,h}(s) - \overline{\beta}^n_{m,h}(s)$; $\underset{\sim}{V}_{m,h}(s) \leftarrow \underset{\sim}{V}_{m,h}(s) + \underline{\beta}_n$; $n \leftarrow \underline{n} - |\mathcal{F}|$;
for $(j, a, \hat{\pi}, \overline{V}', \underline{V}', r) \in \mathcal{F}$ **do**
  $n \leftarrow n + 1$;
  $\tilde{V}_{m,h}(s) \leftarrow (1 - \alpha_n)\tilde{V}_{m,h}(s) + \alpha_n\left(r + \overline{V}'\right)$;
  $\underset{\sim}{V}_{m,h}(s) \leftarrow (1 - \alpha_n)\underset{\sim}{V}_{m,h}(s) + \alpha_n\left(r + \underline{V}'\right)$;
$\tilde{V}_{m,h}(s) \leftarrow \tilde{V}_{m,h}(s) + \overline{\beta}^n_{m,h}(s)$; $\underset{\sim}{V}_{m,h}(s) \leftarrow \underset{\sim}{V}_{m,h}(s) - \underline{\beta}_n$;
$\overline{V}_{m,h}(s) \leftarrow \min\left\{H + 1 - h, \tilde{V}_{m,h}(s), \overline{V}_{m,h}(s)\right\}$;$\underline{V}_{m,h}(s) \leftarrow \max\left\{0, \underset{\sim}{V}_{m,h}(s), \underline{V}_{m,h}(s)\right\}$;

Algorithm 8: Subroutine POLICY_OPT$_{m,h,s}$ for agent m for Algorithm 5

Init: $\forall a \in \mathcal{A}, \hat{L}_{m,h}(s, a) \leftarrow 0$;
Receive $\mathcal{F}, \underline{n}, \overline{n}$; $n \leftarrow \underline{n} - |\mathcal{F}|$;
for $(j, a, \hat{\pi}, \overline{V}', \underline{V}', r) \in \mathcal{F}$ **do**
  $n \leftarrow n + 1$;
  for $a' \in \mathcal{A}$ **do**
    $\hat{l}^n_{m,h}(s, a') = \mathbb{I}(a' = a)\left[\frac{H - r - \overline{V}'}{H}\right] \Big/ \left[\hat{\pi} + \gamma^j_{m,h}(s)\right]$;
    $\hat{L}_{m,h}(s, a') = \hat{L}_{m,h}(s, a') + w_n \hat{l}^n_{m,h}(s, a')$;
$\hat{\pi}_h(\cdot|s) \propto \exp\left(-(\eta^{\overline{n}}_{m,h}(s)/w_{\overline{n}})\hat{L}_{m,h}(s, \cdot)\right)$;
Empty $\mathcal{F}$;

All parameters share the same definition with those in previous subsection.

C.3. DA-MAVL with Reward Skipping

Algorithm 9: DA-MAVL Training with Reward Skipping for Agent m

Init: $\forall(s,h)$, $\overline{n}^0_{m,h}(s) \leftarrow 0$, $\underline{n}^0_{m,h}(s) \leftarrow 0$, $\mathcal{T}^0_{m,h}(s) \leftarrow 0$, $\mathcal{F}_{m,h}(s) \leftarrow \emptyset$, $\mathcal{M}_{m,h}(s) \leftarrow \emptyset$;

for *Episode* $k = 1, \dots, K$ **do**
  Receive initial state s^k_1;
  for *Step* $h = 1, \dots, H$ **do**
    $s \leftarrow s^k_h$;
    // Skipping
    $\overline{n} \leftarrow \overline{n}^k_{m,h}(s) = \overline{n}^{k-1}_{m,h}(s) + 1$; $\underline{n} \leftarrow \underline{n}^k_{m,h}(s) = \underline{n}^{k-1}_{m,h}(s)$;
    for $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r) \in \mathcal{M}^-_{m,h}(s)$ **do**
      $\phi^{i,\overline{n}}_{m,h}(s) \leftarrow \phi^{i,\overline{n}}_{m,h}(s) + \overline{n} - i$;
      if $\phi^{i,\overline{n}}_{m,h}(s) > \sqrt{\mathcal{T}^{\overline{n}}_{m,h}(s)}$ **then**
        Save $(i, a, \hat{\pi}, H, 0, 0)$ to $\mathcal{F}_{m,h}(s)$; Remove $(i, a, \hat{\pi}, \overline{V}', \underline{V}')$ from $\mathcal{M}^-_{m,h}(s)$;
        $\underline{n} \leftarrow \underline{n} + 1$;
    // Preparation
    for $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r) \in \mathcal{M}^+_{m,h}(s)$ **do**
      if $\forall j < i, j \notin \arg\{\mathcal{M}^-_{m,h}(s)\}$ **then**
        Save $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r)$ to $\mathcal{F}_{m,h}(s)$; Remove $(i, a, \hat{\pi}, \overline{V}', \underline{V}')$ from $\mathcal{M}^+_{m,h}(s)$;
    $\mathcal{T}^{\overline{n}}_{m,h}(s) \leftarrow \mathcal{T}^{\overline{n}-1}_{m,h}(s) + |\mathcal{M}_{m,h}(s)|$;
    // Learning
    $\overline{V}^k_{m,h}(s), \underline{V}^k_{m,h}(s) \leftarrow$ VALUE_UPDATE$_{m,h,s}\big(\mathcal{F}_{m,h}(s), \underline{n}, \overline{n}\big)$;
    $\hat{\pi}^k_h(\cdot|s) \leftarrow$ POLICY_OPT$_{m,h,s}\big(\mathcal{F}_{m,h}(s), \underline{n}, \overline{n}\big)$;
    // Execution
    Take action $a^k_h \sim \hat{\pi}^k_h(\cdot|s)$; Observe next state s^k_{h+1};
    for $s' \in \mathcal{S} \backslash s$ **do**
      $\overline{n}^k_{m,h}(s') \leftarrow \overline{n}^{k-1}_{m,h}(s')$; $\underline{n}^k_{m,h}(s') \leftarrow \underline{n}^{k-1}_{m,h}(s')$;
      $\overline{V}^k_{m,h}(s') \leftarrow \overline{V}^{k-1}_{m,h}(s')$; $\underline{V}^k_{m,h}(s') \leftarrow \underline{V}^{k-1}_{m,h}(s')$; $\hat{\pi}^k_{m,h}(\cdot|s') \leftarrow \hat{\pi}^{k-1}_{m,h}(\cdot|s')$;
  for *Step* $h = 1, \dots, H$ **do**
    Save $\big(\overline{n}^k_{m,h}(s^k_h), a^k_h, \hat{\pi}^k_h(a^k_h|s^k_h), \overline{V}^k_{m,h+1}(s^k_{h+1}), \underline{V}^k_{m,h+1}(s^k_{h+1})\big)$ to $\mathcal{M}^-_{m,h}(s^k_h)$;
  Receive delayed reward for all states s;
  for *Delayed Reward* (m, h, s, i, r) **do**
    if $(i, a, \hat{\pi}, \overline{V}', \underline{V}') \notin \mathcal{M}^-_{m,h}(s)$ **then**
      continue;
    Extract and remove $\big(i, a, \hat{\pi}, \overline{V}', \underline{V}'\big)$ from $\mathcal{M}^-_{m,h}(s)$;
    Save $\big(i, a, \hat{\pi}, \overline{V}', \underline{V}', r\big)$ to $\mathcal{M}^+_{m,h}(s)$;

Algorithm 10: Subroutine VALUE_UPDATE$_{m,h,s}$ for agent m for Algorithm 9

Init: $n \leftarrow 0$, $n' \leftarrow 0$, $\tilde{V}_{m,h}(s) \leftarrow 0$, $\underset{\sim}{V}_{m,h}(s) \leftarrow 0$; $\overline{V}_{m,h}(s) \leftarrow 0$; $\underline{V}_{m,h}(s) \leftarrow 0$;

for *Episode* $k = 1, \ldots, K$ **do**

  Receive $\mathcal{F}$, $\underline{n}$, $\overline{n}$;

  $\tilde{V}_{m,h}(s) \leftarrow \tilde{V}_{m,h}(s) - \overline{\beta}^{n,n'}_{m,h}(s)$; $\underset{\sim}{V}_{m,h}(s) \leftarrow \underset{\sim}{V}_{m,h}(s) + \underline{\beta}^{n,n'}_{m,h}(s)$; $n \leftarrow \underline{n}$; $n' \leftarrow \overline{n}$;

  for $(i, a, \hat{\pi}, \overline{V}', \underline{V}', r) \in \mathcal{F}$ **do**

    $\tilde{V}_{m,h}(s) \leftarrow (1-\alpha_i)\tilde{V}_{m,h}(s) + \alpha_i\left(r + \overline{V}'\right)$;

    $\underset{\sim}{V}_{m,h}(s) \leftarrow (1-\alpha_i)\underset{\sim}{V}_{m,h}(s) + \alpha_i\left(r + \underline{V}'\right)$;

  $\tilde{V}_{m,h}(s) \leftarrow \tilde{V}_{m,h}(s) + \overline{\beta}^{n,n'}_{m,h}(s)$;$\underset{\sim}{V}_{m,h}(s) \leftarrow \underset{\sim}{V}_{m,h}(s) - \underline{\beta}^{n,n'}_{m,h}(s)$;

  $\overline{V}_{m,h}(s) \leftarrow \min\left\{H{+}1{-}h, \tilde{V}_{m,h}(s), \overline{V}_{m,h}(s)\right\}$;$\underline{V}_{m,h}(s) \leftarrow \max\left\{0, \underset{\sim}{V}_{m,h}(s), \underline{V}_{m,h}(s)\right\}$;

Algorithm 9 shares the same subroutine 'POLICY_OPT$_{m,h,s}$' with Algorithm 1. The definition of 'VALUE_UPDATE$_{m,h,s}$' is presented in Subroutine 10. Parameters for Subroutine 10 share the same definitions with those in Subroutine 3 except:

$$\overline{\beta}^{n,n'}_{m,h}(s) = \begin{cases} 24H^2C\sqrt{\frac{\mathcal{T}^{n'}_{m,h}(s)}{n^2}}\iota + 18H^2\sqrt{\frac{A}{n}}\iota, & n \geq 1 \\ 0, & n = 0 \end{cases},$$
$$\underline{\beta}^{n,n'}_{m,h}(s) = \begin{cases} 2H^2\frac{\sqrt[4]{4\mathcal{T}^{n'}_{m,h}(s)}+2}{n} + 2\sqrt{\frac{H^3}{n}}\iota, & n \geq 1 \\ 0, & n = 0 \end{cases}, \tag{21}$$
$$\gamma^n_{m,h}(s) = \eta^n_{m,h}(s) = \sqrt{\frac{\iota}{nA + \mathcal{T}^n_{m,h}(s)}}, \quad n \geq 1.$$

where $\mathcal{T}^n_{m,h}(s)$ is a parameter maintained by Algorithm 9.

Appendix D. Notations

In this section, we summarize and introduce the important notations.

Recall $\log(4MHSAK/\delta)$. Recall $\{\alpha_n\}_{n\in\mathbb{N}}$ and $\{w_n\}_{n\in\mathbb{N}}$ as defined in Equation (20). We also define an auxiliary sequence $\{\alpha^i_n\}_{i\in[n],n\in\mathbb{N}}$ as follows:

$$\alpha^i_n = \begin{cases} \alpha_i\prod_{j=i+1}^n(1-\alpha_j), & n > i \geq 1 \\ \alpha_i, & n = i \geq 1 \end{cases}, \quad \alpha^0_n = \begin{cases} 0, & n > 0 \\ 1, & n = 0 \end{cases}. \tag{22}$$

We summarize the important properties of this sequence in Lemma 6.

Consider agent m and (h,s). $\underline{n}^k_{m,h}(s)$ denotes the count of usable visits, $\overline{n}^k_{m,h}(s)$ denotes the count of happened visits at episode k. $\pi^k_{m,h}(\cdot|s)$, $a^k_{m,h}$ and $r^k_{m,h}$ denote the policy, action and reward at episode k. $\mathcal{M}^{-,k}_{m,h}(s)$, $\mathcal{M}^{+,k}_{m,h}(s)$ and $\mathcal{F}^k_{m,h}(s)$ denote the set of **unreceived visits**, **unusable visits** and **visits to be used** at the beginning of episode k. In other words, they refer to sets $\mathcal{M}^-_{m,h}(s)$, $\mathcal{M}^+_{m,h}(s)$ and $\mathcal{F}_{m,h}(s)$ before the 'Learning' process of episode k. We also let $\mathcal{M}^k_{m,h}(s) = \mathcal{M}^{-,k}_{m,h}(s) \cup \mathcal{M}^{+,k}_{m,h}(s)$. When (m,h,s) is fixed in the context, the above notations will be abbreviated as $\underline{n}_k$, $\overline{n}_k$, $\pi_k(\cdot)$, a_k, r_k, $\mathcal{M}_k = \mathcal{M}^-_k \cup \mathcal{M}^+_k$ and $\mathcal{F}_k$.

Consider agent m and the n-th visit of (h, s). $k_h^n(s)$ denotes the episode when it happens. $e_{m,h}^n(s)$ denotes the earliest unreceived visit when it happens. For this visit, parameters $\eta_{m,h}^n(s)$, $\gamma_{m,h}^n(s)$, $\overline{\beta}_{m,h}^n(s)$, $\underline{\beta}_n$ are maintained by Algorithm 1 and its subroutines (Algorithm 3 and 4). Parameters $\eta_{m,h}^n(s)$, $\gamma_{m,h}^n(s)$, $\{\overline{\beta}_{m,h}^{i,n}(s)\}_{i\in[n]}$, $\{\underline{\beta}_{m,h}^{i,n}(s)\}_{i\in[n]}$ and $\{\phi_{m,h}^{i,n}(s)\}_{i\in[n]}$ are maintained by the Algorithm 9 and its subroutines (Algorithm 10 and 4). It is worth notice that the skipping metric $\phi_{m,h}^{i,n}(s)$ in Algorithm 9 can be written as:

$$\phi_{m,h}^{i,n}(s) = \sum_{j=i+1}^{n} (j-i) \cdot \mathbb{I}\{i \in \arg \mathcal{M}_{m,h}^{k_j}(s)\} \tag{23}$$

When (m, h, s) is fixed in the context, the above notations are abbreviated as k_n, e_n η_n, γ_n, $\{\overline{\beta}_{i,n}\}_{i\in[n]}$, $\{\underline{\beta}_{i,n}\}_{i\in[n]}$ and $\{\phi_{i,n}\}_{i\in[n]}$.

For pair (m, h, s, n), we also define three bandit losses $\hat{l}_{m,h}^n(s,a)$, $l_{m,h}^n(s,a)$ and $\bar{l}_{m,h}^n(s,a)$:

$$\begin{aligned}
\hat{l}_{m,h}^n(s,a) &= \frac{1}{H}\mathbb{I}(a = a_{m,h}^{k_n})\Big[H - r_{m,h}^{k_n} - \overline{V}_{m,h+1}^{k_n}(s_{h+1}^{k_n})\Big] \Big/ \Big[\hat{\pi}_{m,h}^{k_n}(a|s) + \gamma_{m,h}^n(s)\Big], \\
l_{m,h}^n(s,a) &= \frac{1}{H}\Big[H - r_{m,h}(s,\boldsymbol{a}) - \overline{V}_{m,h+1}^{k_n}(s_{h+1}^{k_n})\Big], \text{where} \quad \boldsymbol{a} = (a, a_{-m,h}^{k_n}), \\
\bar{l}_{m,h}^n(s,a) &= \frac{1}{H} \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a,a_{-m,h}) \\ a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_n},\ s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \Big[H - r_{m,h}(s,\boldsymbol{a}) - \overline{V}_{m,h+1}^{k_n}(s')\Big].
\end{aligned} \tag{24}$$

Here the first loss $\hat{l}_{m,h}^n(s,a)$ is the bandit loss used in Subroutine 4, while the other two are its variants. When (m, h, s) is fixed in the context, they can be abbreviated $\hat{l}_n(a)$, $l_n(a)$ and $\bar{l}_n(a)$.

Consider agent m and the first n visits of (h, s). Recall $R_{m,h}^n(s)$ denotes the policy optimization regret. Recall $\mathcal{T}_{m,h}^n(s)$ denotes the count of unusable and unreceived visits. In Algorithm 1, it can be written as:

$$\mathcal{T}_{m,h}^n(s) = \sum_{i=1}^{n} i - \arg\min_j \Big[d_{m,h}^j(s) + k_j \geq k_i\Big] = \sum_{i=1}^{n} i - e_{m,h}^n(s) = \sum_{i=1}^{n} |\mathcal{M}_{m,h}^{k_i}(s)|, \tag{25}$$

where $e_{m,h}^n(s)$ and $\mathcal{M}_{m,h}^{k_i}(s)$ refer to the variable and set related Algorithm 1. In Algorithm 9 with reward skipping, we define the set of skipped visits of (h, s) during the first n visits of (h, s) as $\mathcal{O}_{m,h}^n(s)$. Then $\mathcal{T}_{m,h}^n(s)$ can be written as:

$$\mathcal{T}_{m,h}^n(s) = \sum_{i=1}^{n} i - \mathop{\arg\min}_{j\notin\mathcal{O}_{m,h}^i(s)} \Big[d_{m,h}^j(s) + k_j \geq k_i\Big] = \sum_{i=1}^{n} i - e_{m,h}^n(s) = \sum_{i=1}^{n} |\mathcal{M}_{m,h}^{k_i}(s)|, \tag{26}$$

where $e_{m,h}^n(s)$ and $\mathcal{M}_{m,h}^{k_i}(s)$ refer to the variable and set related Algorithm 9. We also define $\mathcal{T}_{m,h}^{n,\mathcal{L}}(s)$ to denote the count of unreceived and unusable visits during the first n visits of (h, s), if the delays of visits in $\mathcal{L}$ are set to 0:

$$\mathcal{T}_{m,h}^n(s) = \sum_{i=1}^{n} i - \mathop{\arg\min}_{j\notin\mathcal{L}} \Big[d_{m,h}^j(s) + k_j \geq k_i\Big]. \tag{27}$$

When (m, h, s) is fixed in the context, the above notations are abbreviated as R_n, $\mathcal{O}_n$, $\mathcal{T}_n$, $\mathcal{T}_{n,\mathcal{L}}$.

Finally, we introduce policies $\{\pi^k_{m,h}\}_{m\in[M],h\in[H],k\in[K]}$ defined by their execution procedures:

Algorithm 11: Policy Certification for $\pi^k_{m,h}$

for *Episode* $h' = h, \ldots, H$ **do**
- Observe current state $s_{h'}$; $n \leftarrow \max_m \underline{n}^k_{m,h'}(s_{h'})$;
- Sample i from $[n]$ with probability α^i_n; $k \leftarrow k^i_{h'}(s_{h'})$;
- Take action $a_{m,h'} \sim \hat{\pi}^k_{m,h'}(\cdot|s_{h'})$;

While definitions of $\{\pi^k_{m,h}\}_{m\in[M],h\in[H],k\in[K]}$ and $\{\pi_m\}_{m\in[M]}$ are similar, their differences are two-fold: (1) $\pi^k_{m,h}$ begins from a given k while π_m begins by sampling a k from $[K]$; (2) $\pi^k_{m,h}$ is for steps from h to H while π_m is for steps from 1 to H. This definition mainly follows the certification process in Jin et al. (2021); Song et al. (2021). We refer the readers to their work for further details.

Appendix E. Performance Guarantee for DA-MAVL

As is stated in Section 4.1, proof of Theorem 1 is under assumption 1 and can be broken down into three steps. **In step one**, we bound the policy optimization regret for Subroutine 4 (Lemma 1). **In step two**, we establish the optimism and pessimism of our value estimates in Subroutine 3 (Lemma 2). **In step three**, we bound the gap between the optimistic and pessimistic value estimates and prove the main theorem (Theorem 1).

E.1. Step One: Proof of Lemma 1

Consider any fixed pair (m, h, s, n). Let R_n denote $R^n_{m,h}(s)$, and let a^* denote the optimal action for the first n visit of (h, s). Utilizing notations $l_i(a), \bar{l}_i(a)$, regret R_n can be rewritten as:

$$\begin{aligned}R_n = &\max_{a\in\mathcal{A}_m} \sum_{i=1}^n \alpha^i_n \Big[\mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a,a_{-m,h}) \\ a_{-m,h}\sim\hat{\pi}^{k_i}_{-m,h},\ s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \Big(r_{m,h}(s,\boldsymbol{a}) + \overline{V}^{k_i}_{m,h+1}(s') \Big) - \Big(r^{k_i}_{m,h} + \overline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1}) \Big) \Big] \\ =&H \sum_{i=1}^n \alpha^i_n \Big[l_i(a_{k_i}) - \bar{l}_i(a^*) \Big].\end{aligned} \tag{28}$$

We then decompose the regret as follows:

$$\begin{aligned}R_n &= H \sum_{i=1}^n \alpha^i_n \Big[l_i(a_{k_i}) - \bar{l}_i(a^*) \Big] \\ &= H \prod_{i=2}^n (1-\alpha_i) \cdot \sum_{i=1}^n w_i \Big[l_i(a_{k_i}) - \bar{l}_i(a^*) \Big] \\ &= H \prod_{i=2}^n (1-\alpha_i) \cdot R'_n.\end{aligned} \tag{29}$$

Recall that $\mathcal{M}_{k_i}$ denotes the set of unusable or unreceived visits of (h, s) for agent m at the beginning of the i-th visit. We then define two cumulative losses for i:

$$\begin{aligned} \hat{L}_i(a) &= \sum_{j \in [i-1] \setminus \mathcal{M}_{k_i}} w_j \hat{l}_j(a), \\ \tilde{L}_{i+1}(a) &= \sum_{j \in [i]} w_j \hat{l}_j(a) = \hat{L}_i(a) + w_i \hat{l}_i(a) + \sum_{j \in \mathcal{M}_{k_i}} w_j \hat{l}_j(a). \end{aligned} \tag{30}$$

Cumulative loss $\hat{L}_i(a)$ generates policy $\hat{\pi}_{k_i}$ in algorithm 1. Cumulative loss $\tilde{L}_{i+1}(a)$ is the cheating version of $\hat{L}_i(a)$, which includes all information of the first i visits. Then we have two corresponding policies:

$$\begin{aligned} \hat{\pi}_{k_i}(\cdot) &\propto \exp\Big[-(\eta_i / w_i) \hat{L}_i(\cdot) \Big], \\ \tilde{\pi}_{k_{i+1}}(\cdot) &\propto \exp\Big[-(\eta_i / w_i) \tilde{L}_{i+1}(\cdot) \Big]. \end{aligned} \tag{31}$$

Then the weighted regret R'_n can be further decomposed as follows:

$$\begin{aligned} R'_n &= \sum_{i=1}^{n} w_i \Big[l_i(a_{k_i}) - \bar{l}_i(a^*) \Big] \\ &= \underbrace{\sum_{i=1}^{n} w_i \Big[l_i(a_{k_i}) - \hat{l}_i(a_{k_i}) \hat{\pi}_{k_i}(a_{k_i}) \Big]}_{I_{1,1}} + \underbrace{\sum_{i=1}^{n} w_i \Big[\hat{l}_i(a_{k_i}) \hat{\pi}_{k_i}(a_{k_i}) - \hat{l}_i(a_{k_i}) \tilde{\pi}_{k_{i+1}}(a_{k_i}) \Big]}_{I_{1,2}} \\ &\quad + \underbrace{\sum_{i=1}^{n} w_i \Big[\hat{l}_i(a_{k_i}) \tilde{\pi}_{k_{i+1}}(a_{k_i}) - \hat{l}_i(a^*) \Big]}_{I_{1,3}} + \underbrace{\sum_{i=1}^{n} w_i \Big[\hat{l}_i(a^*) - \bar{l}_i(a^*) \Big]}_{I_{1,4}}. \end{aligned} \tag{32}$$

We then give the upper bounds for $I_{1,1}$, $I_{1,2}$, $I_{1,3}$, $I_{1,4}$ in the following lemmas.

Lemma 3 (The upper bound of $I_{1,1} + I_{1,2}$) *For $\forall (m, h, s, n) \in [M] \times [H] \times \mathcal{S} \times [K]$, the following equation holds with probability at least $1 - \delta/4$:*

$$I_{1,1} + I_{1,2} \le w_n d_{max} \iota + \sum_{i=1}^{n} w_i \gamma_{m,h}^i(s) \Big(2A + |\mathcal{M}_{m,h}^{k_h^i(s)}(s)| \Big). \tag{33}$$

Lemma 4 (The upper bound of $I_{1,3}$) *For $\forall (m, h, s, n) \in [M] \times [H] \times \mathcal{S} \times [K]$, the following equation holds:*

$$I_{1,3} \le \frac{w_n}{\eta_{m,h}^n(s)} \iota. \tag{34}$$

Lemma 5 (The upper bound of $I_{1,4}$) *For $\forall (m, h, s, n) \in [M] \times [H] \times \mathcal{S} \times [K]$, the following equation holds with probability at least $1 - \delta/4$:*

$$I_{1,4} \le \frac{w_n}{\eta_{m,h}^n(s)} \iota. \tag{35}$$

Combining the three lemmas and by union bound, the following equations hold for $\forall(m,h,s,n) \in [M]\times[H]\times\mathcal{S}\times[K]$ with probability at least $1-\delta/2$:

$$\begin{aligned} R_n =& H\cdot\prod_{i=2}^{n}(1-\alpha_i)\cdot\Big(I_{1,1}+I_{1,2}+I_{1,3}+I_{1,4}\Big) \\ \le& H\alpha_n d_{max}\iota + H\sum_{i=1}^{n}\alpha_n^i\gamma_i\Big(2A+|\mathcal{M}_{k_i}|\Big)+2H\frac{\alpha_n}{\eta_n}\iota \\ \le& \frac{4H^2}{n}\sum_{i=1}^{n}\Big(A+|\mathcal{M}_{k_i}|\Big)\sqrt{\frac{\iota}{iA+\mathcal{T}_i}}+4H^2\sqrt{\frac{nA+\mathcal{T}_n}{n^2}\iota}+\frac{2d_{max}H^2}{n}\iota \\ \le& 12H^2\sqrt{\frac{nA+\mathcal{T}_n}{n^2}\iota}+\frac{2d_{max}H^2}{n}\iota. \end{aligned} \tag{36}$$

Here the third line comes from Lemma 6, the last line utilizes Lemma 1 in Streeter and McMahan (2010), and the fact that $\mathcal{T}_i=\sum_{j=1}^{i}|\mathcal{M}_{k_j}|$.

E.1.1. SUPPORTING DETAILS

Lemma 6 *The following properties hold for* $\forall n\ge i\ge 1$*:*

- $\frac{1}{\sqrt{n}}\le\sum_{i=1}^{n}\frac{\alpha_n^i}{\sqrt{n}}\le\frac{2}{\sqrt{n}}$ *and* $\frac{1}{n}\le\sum_{i=1}^{n}\frac{\alpha_n^i}{n}\le\frac{2}{n}$,
- $\max_{i\in[n]}\alpha_n^i\le\frac{2H}{n}$,
- $\sum_{n=1}^{\infty}\alpha_n^i=1+\frac{1}{H}$,
- $\alpha_n^i=w_i\prod_{j=2}^{n}(1-\alpha_j)$.

Proof Here the first three lines is from Lemma 4.1 in Jin et al. (2021), while the last line is from Proof of Corollary 19 in Jin et al. (2021). ■

Lemma 7 *The following property hold for* $\forall(m,h,s,n)\in[M]\times[H]\times\mathcal{S}\times[K]$*:*

$$|\mathcal{M}_{m,h}^{k_h^n(s)}(s)| = n-e_{m,h}^n(s)\le d_{max}. \tag{37}$$

Similar property hold for $\forall(m,h,s,k)\in[M]\times[H]\times\mathcal{S}\times[K]$*:*

$$|\mathcal{M}_{m,h}^k(s)| = \overline{n}_{m,h}^k(s)-\underline{n}_{m,h}^k(s)-1\le d_{max}. \tag{38}$$

Proof We first proof Equation (37). Consider any fixed pair (m,h,s,n). Recall that k_n denotes the episode when the n-th visit of (h,s) happens, e_n is from Equation (2) and (3).

If $|\mathcal{M}_{k_n}|=0$, all first $n-1$ visits are received, which gives $e_n=n$. The lemma clearly holds:

$$|\mathcal{M}_{k_n}| = n-e_n=0. \tag{39}$$

If $|\mathcal{M}_{k_n}|>0$, e_n is the first unreceived visit. According to the definitions of unusable and unrecieved visits, all visits with happening order $e_n,\ldots,n-1$ are unusable and unreceived visits.

Consequently the count of unusable and unreceived visits are $n-e_n$. On the other hand, Algorithm 1 ensures all unusable and unreceived visits are included in set $\mathcal{M}_{k_n}$. Therefore,

$$|\mathcal{M}_{k_n}| = n - e_n - 1. \tag{40}$$

We prove the other half of the equation by contradiction. Suppose $e_n \leq n - d_{max} - 1$, then the e_n-th visit has been delayed for at least $d_{max} + 1$ episodes. This contradict with Assumption 1. Therefore $e_n \geq n - d_{max}$, which completes the proof.

We now prove Equation (38). At episode k, visits with happening order $1, \ldots, \underline{n}_k$ are usuable visits, while visits with happening order $\underline{n}_k + 1, \ldots, \overline{n}_k - 1$ are unusable or unreceived visits. This directly implies:

$$|\mathcal{M}_k| = \overline{n}_k - \underline{n}_k. \tag{41}$$

The other half of the equation follows directly from Equation (37). ■

Lemma 8 *For* $\forall(m, h, s, n) \in [M] \times [H] \times \mathcal{S} \times [K]$, *the following inequality holds:*

$$I_{1,1} \leq \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_m} w_i \gamma_{m,h}^i(s) \hat{l}_{m,h}^i(s, a). \tag{42}$$

Proof [Proof of Lemma 8] Consider any fixed pair of (m, h, s, n).

$$\begin{aligned} I_{1,1} &= \sum_{i=1}^{n} w_i \Big[l_i(a_{k_i}) - \hat{l}_i(a_{k_i}) \hat{\pi}_{k_i}(a_{k_i}) \Big] \\ &= \sum_{i=1}^{n} w_i \Big[\hat{\pi}_{k_i}(a_{k_i}) + \gamma_i - \hat{\pi}_{k_i}(a_{k_i}) \Big] \hat{l}_i(a_{k_i}) \\ &= \sum_{i=1}^{n} w_i \gamma_i \hat{l}_i(a_{k_i}) \\ &= \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_m} w_i \gamma_i \hat{l}_i(a). \end{aligned} \tag{43}$$

Here the second and the last line follows directly from the definitions of $l_i(a)$ and $\hat{l}_i(a)$. ■

Lemma 9 *For* $\forall(m, h, s, n) \in [M] \times [H] \times \mathcal{S} \times [K]$, *the following equation holds:*

$$I_{1,2} \leq \sum_{j=1}^{n} \sum_{a \in \mathcal{A}_m} \hat{l}_{m,h}^j(s, a) \Big(\eta_{m,h}^j(s) w_j + \sum_{i : j \in \mathcal{M}_{m,h}^{k_i}(s)} \eta_{m,h}^i(s) w_i \mathbb{I}\Big\{ a_{k_i} = a \Big\} \Big). \tag{44}$$

Proof [Proof of Lemma 9] Consider any fixed pari (m, h, s, n).

$$\begin{aligned} I_{1,2} &= \sum_{i=1}^{n} w_i \Big[\hat{l}_i(a_{k_i}) \hat{\pi}_{k_i}(a_{k_i}) - \hat{l}_i(a_{k_i}) \tilde{\pi}_{k_{i+1}}(a_{k_i}) \Big] \\ &= \sum_{i=1}^{n} w_i \hat{l}_i(a_{k_i}) \hat{\pi}_{k_i}(a_{k_i}) \cdot \Big[1 - \frac{\tilde{\pi}_{k_{i+1}}(a_{k_i})}{\hat{\pi}_{k_i}(a_{k_i})} \Big]. \end{aligned} \tag{45}$$

By the definition of $\tilde{\pi}_{k_{i+1}}(a_{k_i})$ and $\hat{\pi}_{k_i}(a_{k_i})$, we have:

$$\begin{aligned}\frac{\tilde{\pi}_{k_{i+1}}(a_{k_i})}{\hat{\pi}_{k_i}(a_{k_i})} &= \frac{\exp\Big\{-(\eta_i/w_i)\tilde{L}_{i+1}(a_{k_i})\Big\}}{\exp\Big\{-(\eta_i/w_i)\hat{L}_i(a_{k_i})\Big\}} \cdot \frac{\sum_a \exp\Big\{-(\eta_i/w_i)\hat{L}_i(a)\Big\}}{\sum_a \exp\Big\{-(\eta_i/w_i)\tilde{L}_{i+1}(a)\Big\}} \\ &\geq \frac{\exp\Big\{-(\eta_i/w_i)\hat{L}_i(a_{k_i}) - (\eta_i/w_i)\sum_{j\in\mathcal{M}_{k_i}} w_j\hat{l}_j(a_{k_i}) - (\eta_i w_i/w_i)\hat{l}_i(a_{k_i})\Big\}}{\exp\Big\{-(\eta_i/w_i)\hat{L}_i(a_{k_i})\Big\}} \\ &= \exp\Big\{-\frac{\eta_i}{w_i}\sum_{j\in\mathcal{M}_{k_i}} w_j\hat{l}_j(a_{k_i}) - \eta_i\hat{l}_i(a_k)\Big\} \\ &\geq 1 - \frac{\eta_i}{w_i}\sum_{j\in\mathcal{M}_{k_i}} w_j\hat{l}_j(a_{k_i}) - \eta_i\hat{l}_i(a_{k_i}).\end{aligned} \tag{46}$$

Here the second line is due to the fact $\tilde{L}_{i+1}(a) \geq \hat{L}_i(a)$. Substituting into Equation (45), we get:

$$\begin{aligned}I_{1,2} &\leq \sum_{i=1}^n w_i\hat{l}_i(a_{k_i})\hat{\pi}_{k_i}(a_{k_i}) \cdot \Big(\frac{\eta_i}{w_i}\sum_{j\in\mathcal{M}_{k_i}} w_j\hat{l}_j(a_{k_i}) + \eta_i\hat{l}_i(a_{k_i})\Big) \\ &= \sum_{i=1}^n \eta_i \sum_{j\in\mathcal{M}_{k_i}} w_j\hat{l}_j(a_{k_i}) + \sum_{i=1}^n \eta_i w_i\hat{l}_i(a_{k_i}) \\ &= \sum_{i=1}^n \eta_i \sum_{j\in\mathcal{M}_{k_i}} \sum_{a\in\mathcal{A}_m} w_j\hat{l}_j(a)\mathbb{I}\Big\{a_{k_i} = a\Big\} + \sum_{i=1}^n \sum_{a\in\mathcal{A}_m} \eta_i w_i\hat{l}_i(a) \\ &= \sum_{j=1}^n \sum_{a\in\mathcal{A}_m} \hat{l}_j(a) \sum_{i:j\in\mathcal{M}_{k_i}} \eta_i w_j\mathbb{I}\Big\{a_{k_i} = a\Big\} + \sum_{i=1}^n \sum_{a\in\mathcal{A}_m} \eta_i w_i\hat{l}_i(a) \\ &= \sum_{j=1}^n \sum_{a\in\mathcal{A}_m} \hat{l}_j(a) \sum_{i:j\in\mathcal{M}_{k_i}} \eta_i w_i\mathbb{I}\Big\{a_{k_i} = a\Big\} + \sum_{i=1}^n \sum_{a\in\mathcal{A}_m} \eta_i w_i\hat{l}_i(a) \\ &\leq \sum_{j=1}^n \sum_{a\in\mathcal{A}_m} \hat{l}_j(a)\Big(\eta_j w_j + \sum_{i:j\in\mathcal{M}_{k_i}} \eta_i w_i\mathbb{I}\Big\{a_{k_i} = a\Big\}\Big).\end{aligned} \tag{47}$$

Here the first line is due to the following fact:

$$\hat{l}_i(a_{k_i})\hat{\pi}_{k_i}(a_{k_i}) \leq \frac{\hat{\pi}_{k_i}(a_{k_i})}{\hat{\pi}_{k_i}(a_{k_i}) + \gamma_i} \leq 1. \tag{48}$$

and the last inequality is because $j \leq i$ and because w_i monotonically increases with i. ■

Proof [Proof of Lemma 3] Consider any fixed pair of (m, h, s, n). With results of Lemma 8, Lemma 9, we directly have:

$$\begin{aligned} I_{1,1} + I_{1,2} \leq & \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_m} \hat{l}_i(a) \Big[2w_i\gamma_i + \sum_{j:i \in \mathcal{M}_{k_j}} \eta_j w_j \mathbb{I}\Big\{ a_{k_j} = a \Big\} \Big] \\ \leq & \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_m} \bar{l}_i(a) \Big[2w_i\gamma_i + \sum_{j:i \in \mathcal{M}_{k_j}} \eta_j w_j \mathbb{I}\Big\{ a_{k_j} = a \Big\} \Big] \\ & + \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_m} \Big[\hat{l}_i(a) - \bar{l}_i(a) \Big] \cdot \Big[2w_i\gamma_i + \sum_{j:i \in \mathcal{M}_{k_j}} \eta_j w_j \mathbb{I}\Big\{ a_{k_j} = a \Big\} \Big] \\ \leq & \sum_{i=1}^{n} 2w_i\gamma_i A + \sum_{i=1}^{n} \sum_{j:i \in \mathcal{M}_{k_j}} \eta_j w_j \\ & + \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_m} \Big[\hat{l}_i(a) - \bar{l}_i(a) \Big] \cdot \Big[2w_i\gamma_i + \sum_{j:i \in \mathcal{M}_{k_j}} \eta_j w_j \mathbb{I}\Big\{ a_{k_j} = a \Big\} \Big]. \end{aligned} \tag{49}$$

Here the last line is because $\bar{l}_i(a) \leq 1$.

For the second term of the last line, switching the summation gives:

$$\sum_{i=1}^{n} \sum_{j:i \in \mathcal{M}_{k_j}} \eta_j w_j = \sum_{j=1}^{n} \eta_j w_j \sum_{i:i \in \mathcal{M}_{k_j}} 1 = \sum_{j=1}^{n} \eta_j w_j |\mathcal{M}_{k_j}|. \tag{50}$$

Therefore

$$\begin{aligned} I_{1,1} + I_{1,2} \leq & \sum_{i=1}^{n} w_i\gamma_i (2A + |\mathcal{M}_{k_i}|) \\ & + \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_m} \Big[\hat{l}_i(a) - \bar{l}_i(a) \Big] \cdot \Big[2w_i\gamma_i + \sum_{j:i \in \mathcal{M}_{k_j}} \eta_j w_j \mathbb{I}\Big\{ a_{k_i} = a \Big\} \Big]. \end{aligned} \tag{51}$$

Notice that

$$2w_i\gamma_i + \sum_{j:i \in \mathcal{M}_{k_j}} \eta_j w_j \mathbb{I}\Big\{ a_{k_i} = a \Big\} \leq w_n \Big(2\gamma_i + \eta_i d_{max} \Big) \leq 2w_n d_{max} \gamma_i. \tag{52}$$

where the first inequality is because $i \geq j$ and i can be delayed for at most d_{max} episodes. Then by Lemma 4.3 in Gyorgy and Joulani (2021), the following equation holds for any fixed pair $(m, h, s, n) \in [M] \times [H] \times \mathcal{S} \times [K]$ with probability at least $1 - \delta/(4MHSK)$:

$$\sum_{i=1}^{n} \sum_{a \in \mathcal{A}_m} \Big[\hat{l}_i(a) - \bar{l}_i(a) \Big] \cdot \Big[2w_i\gamma_i + \sum_{j:i \in \mathcal{M}_{k_j}} \eta_j w_j \mathbb{I}\Big\{ a_j = a \Big\} \Big] \leq w_n d_{max} \iota. \tag{53}$$

By union bound, the above equation holds for all $(m, h, s, n) \in [M] \times [H] \times \mathcal{S} \times [K]$ with probability at least $1 - \delta/4$. Substituting the above results into Equation (49), we have the following equation holds with probability at least $1 - \delta/4$:

$$I_{1,1} + I_{1,2} \leq \sum_{i=1}^{n} w_i \gamma_i \Big(2A + |\mathcal{M}_{k_i}|\Big) + w_n d_{max} \iota. \tag{54}$$

■

Proof [Proof of Lemma 4] Consider any fixed pair of (m, h, s, n).

$$I_{1,3} = \sum_{i=1}^{n} w_i \Big[\hat{l}_i(a_{k_i}) \tilde{\pi}_{k_{i+1}}(a_{k_i}) - \hat{l}_i(a^*)\Big]. \tag{55}$$

To bound the term, we apply Theorem 3 in Joulani et al. (2017) with:

$$\begin{aligned} p_i(\pi) &= 0, & \forall i \in [n], \\ q_0(\pi) &= \frac{w_0}{\eta_0} \sum_{a \in \mathcal{A}_m} \pi(a) \log \pi(a), & \\ q_i(\pi) &= (\frac{w_i}{\eta_i} - \frac{w_{i-1}}{\eta_{i-1}}) \sum_{a \in \mathcal{A}_m} \pi(a) \log \pi(a), & \forall i \in [n]. \end{aligned} \tag{56}$$

Here we define $w_0 = \eta_0 = 1$. Finally, we have:

$$I_{1,3} \leq -\sum_{i=1}^{n} q_i(\tilde{\pi}_{k_{i+1}}) \leq \log A \sum_{i=1}^{n} (\frac{w_i}{\eta_i} - \frac{w_{i-1}}{\eta_{i-1}}) \leq \frac{w_n}{\eta_n} \iota. \tag{57}$$

■

Proof [Proof of Lemma 5] The lemma follows directly from Lemma 4.3 in Gyorgy and Joulani (2021). ■

E.2. Step Two: Proof of Lemma 2

Proof [Proof of lemma 2] We first prove the optimism part of the lemma for any fixed pair (m, h, s, k). Conditioned on the successful event of Lemma 1, which holds for probability at least $1 - \delta/2$, we prove the lemma by induction. For $k = 0$, it is clear that:

$$\overline{V}^{0}_{m,h}(s) = H + 1 - h \geq V^{\dagger, \pi^1_{-m,h}}_{m,h}(s). \tag{58}$$

Recall that $\overline{\beta}_{\underline{n}_k} = 12\sqrt{\frac{\underline{n}_k A + \mathcal{T}_{\underline{n}_k}}{\underline{n}_k^2}\iota} + \frac{4H^2 d_{max}}{\underline{n}_k}\iota$. Let $N_k = \max_m \underline{n}_{m,h}^k(s)$. Suppose the optimism part holds for all $k' < k$. Then for episode k and any (m, h, s):

$$
\begin{aligned}
\overline{V}_{m,h}^k(s) &= \alpha_{\underline{n}_k}^0 \cdot H + \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \left(r_{m,h}^{k_i} + \overline{V}_{m,h+1}^{k_i}(s_{h+1}^{k_i})\right) + \overline{\beta}_{\underline{n}_k} \\
&= \alpha_{\underline{n}_k}^0 \cdot H + H\sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \left[1 - l_i(a_{m,h}^{k_i})\right] + \overline{\beta}_{\underline{n}_k} \\
&\geq \alpha_{\underline{n}_k}^0 \cdot H + H\sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \left[1 - \bar{l}_i(a^*)\right] + \left(\overline{\beta}_{\underline{n}_k} - R_{\underline{n}_k}\right) \\
&= \alpha_{\underline{n}_k}^0 \cdot H + \frac{2d_{max}H^2}{\underline{n}_k}\iota + \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a^*,a_{-m,h}) \\ a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i},\ s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \left[r_{m,h}(s,\boldsymbol{a}) + \overline{V}_{m,h+1}^{k_i}(s')\right].
\end{aligned}
\tag{59}
$$

Here the third line is becuase of Lemma 1.

On the other hand,

$$
\begin{aligned}
V_{m,h}^{\dagger,\pi_{-m,h}^k}(s) &= \max_{\mu_h}\max_{\mu_{h+1:H}} \sum_{i=1}^{N_k} \alpha_{N_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a,a_{-m,h}) \\ a\sim\mu_h,a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i},s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \left(r_{m,h}(s,\boldsymbol{a}) + V_{m,h+1}^{\mu_{h+1:H},\pi_{-m,h+1}^{k_i}}(s')\right) \\
&\leq \max_{\mu_h} \sum_{i=1}^{N_k} \alpha_{N_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a,a_{-m,h}) \\ a\sim\mu_h,a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i},s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \left(r_{m,h}(s,\boldsymbol{a}) + V_{m,h+1}^{\dagger,\pi_{-m,h+1}^{k_i}}(s')\right) \\
&\leq \max_{\mu_h} \sum_{i=1}^{\underline{n}_k} \alpha_{N_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a,a_{-m,h}) \\ a\sim\mu_h,a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i},s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \left(r_{m,h}(s,\boldsymbol{a}) + V_{m,h+1}^{\dagger,\pi_{-m,h+1}^{k_i}}(s')\right) \\
&\quad + \max_{\mu_h} \sum_{i=\underline{n}_k+1}^{N_k} \alpha_{N_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a,a_{-m,h}) \\ a\sim\mu_h,a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i},s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \left(r_{m,h}(s,\boldsymbol{a}) + V_{m,h+1}^{\dagger,\pi_{-m,h+1}^{k_i}}(s')\right) \\
&\leq \max_{\mu_h} \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a,a_{-m,h}) \\ a\sim\mu_h,a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i},s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \left(r_{m,h}(s,\boldsymbol{a}) + V_{m,h+1}^{\dagger,\pi_{-m,h+1}^{k_i}}(s')\right) + \sum_{i=\underline{n}_k+1}^{N_k} \alpha_{N_k} H \\
&\leq \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a^*,a_{-m,h}) \\ a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i},s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \left(r_{m,h}(s,\boldsymbol{a}) + \overline{V}_{m,h+1}^{k_i}(s')\right) + 2H^2\frac{N_k - \underline{n}_k}{N_k} \\
&\leq \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a^*,a_{-m,h}) \\ a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i},s'\sim\mathbb{P}_h(s,\boldsymbol{a})}} \left(r_{m,h}(s,\boldsymbol{a}) + \overline{V}_{m,h+1}^{k_i}(s')\right) + 2H^2\frac{d_{max}}{N_k}.
\end{aligned}
\tag{60}
$$

Here the second line is because of the convexity of the maximum, the last line is due to the fact that $N_k - \underline{n}_k \le (\overline{n}_k - 1) - \underline{n}_k = |\mathcal{M}_k| \le d_{max}$, where the first equation is because $\underline{n}_k$ which is the direct result of Lemma 7.

Finally, combining equations 59 and 60, we finish the proof of the induction and prove the optimism part of the lemma.

Now we prove the pessimism part. For a fixed pair (m, h, s, k), the following hold with probability at least $1 - \delta/(2MHSK)$:

$$
\begin{aligned}
\underline{V}^k_{m,h}(s) &= \sum_{i=1}^{\underline{n}_k} \alpha^i_{\underline{n}_k} \left(r^{k_i}_{m,h} + \underline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1}) \right) - \underline{\beta}_{\underline{n}_k} \\
&\le \sum_{i=1}^{\underline{n}_k} \alpha^i_{\underline{n}_k} \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\hat{\pi}^{k_i}_h, s'\sim\mathbb{P}_h(\cdot|s,\boldsymbol{a})} \left(r_{m,h}(s,\boldsymbol{a}) + \underline{V}^{k_i}_{m,h+1}(s') \right) + 2\sqrt{\frac{H^3}{\underline{n}_k}\iota} - \underline{\beta}_{\underline{n}_k} \\
&\le \sum_{i=1}^{\underline{n}_k} \alpha^i_{\underline{n}_k} \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\hat{\pi}^{k_i}_h, s'\sim\mathbb{P}_h(\cdot|s,\boldsymbol{a})} \left(r_{m,h}(s,\boldsymbol{a}) + \underline{V}^{k_i}_{m,h+1}(s') \right) - 2H^2\frac{d_{max}}{\underline{n}_k}\iota \\
&= \sum_{i=1}^{\underline{n}_k} \alpha^i_{N_k} \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\hat{\pi}^{k_i}_h, s'\sim\mathbb{P}_h(\cdot|s,\boldsymbol{a})} \left(r_{m,h}(s,\boldsymbol{a}) + \underline{V}^{k_i}_{m,h+1}(s') \right) - 2H^2\frac{d_{max}}{\underline{n}_k}\iota \\
&\quad + \left(1 - \prod_{j=\underline{n}_k+1}^{N_k}(1-\alpha_j)\right) \sum_{i=1}^{\underline{n}_k} \alpha^i_{\underline{n}_k} \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\hat{\pi}^{k_i}_h, s'\sim\mathbb{P}_h(\cdot|s,\boldsymbol{a})} \left(r_{m,h}(s,\boldsymbol{a}) + \underline{V}^{k_i}_{m,h+1}(s') \right) \\
&\le \sum_{i=1}^{\underline{n}_k} \alpha^i_{N_k} \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\hat{\pi}^{k_i}_h, s'\sim\mathbb{P}_h(\cdot|s,\boldsymbol{a})} \left(r_{m,h}(s,\boldsymbol{a}) + \underline{V}^{k_i}_{m,h+1}(s') \right) - 2H^2\frac{d_{max}}{\underline{n}_k}\iota + (N_k - \underline{n}_k)\alpha_{\underline{n}_k} \sum_{i=1}^{\underline{n}_k} \alpha^i_{\underline{n}_k} H \\
&\le \sum_{i=1}^{\underline{n}_k} \alpha^i_{N_k} \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\hat{\pi}^{k_i}_h, s'\sim\mathbb{P}_h(\cdot|s,\boldsymbol{a})} \left(r_{m,h}(s,\boldsymbol{a}) + \underline{V}^{k_i}_{m,h+1}(s') \right) - 2H^2\frac{d_{max}}{\underline{n}_k}\iota + 2H^2\frac{d_{max}}{\underline{n}_k} \\
&\le \sum_{i=1}^{N_k} \alpha^i_{N_k} \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\hat{\pi}^{k_i}_h, s'\sim\mathbb{P}_h(\cdot|s,\boldsymbol{a})} \left(r_{m,h}(s,\boldsymbol{a}) + \underline{V}^{k_i}_{m,h+1}(s') \right) \\
&= V^{\pi^k_h}_{m,h}(s).
\end{aligned} \tag{61}
$$

Here the second line follows Azuma's inequality with propbability at least $1 - \delta/(2MHSK)$. While the last line is the definition of $V^{\pi^k_h}_{m,h}(s)$. Finally, by union bound over all $(m, h, s, k) \in [M] \times [H] \times \mathcal{S} \times [K]$, we finish the proof. ■

E.3. Proof of Theorem 1

Proof [Proof of theorem 1] Consider any fixed pair (m, h). We start by upper bounding the term $\sum_{k=1}^{K} (\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h)$. For any episode k, we slightly overload notations $\underline{n}_k = \underline{n}^k_{m,h}(s^k_h)$, $k_i = k^i_h(s^k_h)$, $\overline{\beta}_n = \overline{\beta}^n_{m,h}(s^k_h)$. Then we have:

$$
\begin{aligned}
&(\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h) \\
\leq&\alpha^0_{\underline{n}_k} \cdot H + \sum_{i=1}^{\underline{n}_k} \alpha^i_{\underline{n}_k} \Big[r^{k_i}_{m,h} + \overline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1}) \Big] + \overline{\beta}_{\underline{n}_k} \\
&\qquad - \sum_{i=1}^{\underline{n}_k} \alpha^i_{\underline{n}_k} \Big[r^{k_i}_{m,h} + \underline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1}) \Big] - \underline{\beta}_{\underline{n}_k} \\
\leq&\alpha^0_{\underline{n}_k} \cdot H + \sum_{i=1}^{\underline{n}_k} \alpha^i_{\underline{n}_k} (\overline{V}^{k_i}_{m,h+1} - \underline{V}^{k_i}_{m,h+1})(s^{k_i}_{h+1}) + \overline{\beta}_{\underline{n}_k} - \underline{\beta}_{\underline{n}_k}.
\end{aligned}
\tag{62}
$$

Taking the summation over episode k gives:

$$
\begin{aligned}
&\sum_{k=1}^{K} (\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h) \\
=&\sum_{k=1}^{K} (\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h) \cdot \mathbb{I}\Big\{ \underline{n}_k < d_{max} \Big\} + \sum_{k=1}^{K} (\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h) \cdot \mathbb{I}\Big\{ \underline{n}_k \geq d_{max} \Big\} \\
\leq&H \sum_{k=1}^{K} \mathbb{I}\Big\{ \underline{n}_k < d_{max} \Big\} + \underbrace{\sum_{k=1}^{K} \sum_{i=1}^{\underline{n}_k} \alpha^i_{\underline{n}_k} (\overline{V}^{k_i}_{m,h+1} - \underline{V}^{k_i}_{m,h+1})(s^{k_i}_{h+1})}_{I_{2,1}} \\
&\qquad + \underbrace{\sum_{k=1}^{K} (\overline{\beta}_{\underline{n}_k} + \underline{\beta}_{\underline{n}_k}) \cdot \mathbb{I}\Big\{ \underline{n}_k \geq d_{max} \Big\}}_{I_{2,2}}.
\end{aligned}
\tag{63}
$$

We then give the upper bounds for $I_{2,1}$ and $I_{2,2}$ in the following lemmas.

Lemma 10 (The upper bound of $I_{2,1}$) *For* $\forall (m,h) \in [M] \times [H]$,

$$
I_{2,1} \leq (1 + \frac{1}{H}) \sum_{k=1}^{K} \Big(\overline{V}^k_{m,h+1} - \underline{V}^k_{m,h+1} \Big)(s^k_{h+1}) + 4 d_{max} H^2 S \iota. \tag{64}
$$

Lemma 11 (The upper bound of $I_{2,2}$) *For* $\forall (m,h) \in [M] \times [H]$,

$$
I_{2,2} \leq 12 d_{max} H^2 S \iota^2 + 56 H^2 \sqrt{SAK\iota} + 24 H^2 \sqrt{S \mathcal{T}_K} \iota^2. \tag{65}
$$

Substituting into previous equations, we get:

$$\begin{aligned}
&\sum_{k=1}^{K}(\overline{V}_{m,h}^{k} - \underline{V}_{m,h}^{k})(s_h^k) \\
\leq & H\sum_{k=1}^{K}\mathbb{I}\Big\{\underline{n}_k < d_{max}\Big\} + (1+\frac{1}{H})\sum_{k=1}^{K}\Big(\overline{V}_{m,h+1}^{k} - \underline{V}_{m,h+1}^{k}\Big)(s_{h+1}^k) \\
& \qquad + 4d_{max}H^2S\iota + 12d_{max}H^2S\iota^2 + 56H^2\sqrt{SAK\iota} + 24H^2\sqrt{S\mathcal{T}_K}\iota^2 \\
\leq & 2d_{max}HS + (1+\frac{1}{H})\sum_{k=1}^{K}\Big(\overline{V}_{m,h+1}^{k} - \underline{V}_{m,h+1}^{k}\Big)(s_{h+1}^k) \\
& \qquad + 16d_{max}H^2S\iota^2 + 56H^2\sqrt{SAK\iota} + 24H^2\sqrt{S\mathcal{T}_K}\iota^2
\end{aligned} \tag{66}$$

.

To see why the last line holds, we notice that the $(d_{max}-1)$-th visit of any (h,s) will be received when the $2d_{max}$-th visit of (h,s) happens. This leads to

$$H\sum_{k=1}^{K}\mathbb{I}\Big\{\underline{n}_k < d_{max}\Big\} \leq H\sum_{s\in\mathcal{S}}2d_{max} \leq 2d_{max}HS. \tag{67}$$

Iterating over h, we get:

$$\begin{aligned}
&\sum_{k=1}^{K}(\overline{V}_{m,1}^{k} - \underline{V}_{m,1}^{k})(s_1) \\
\leq & 18d_{max}H^3S\iota^2 + 56H^3\sqrt{SAK\iota} + 24H^3\sqrt{S\mathcal{T}_K}\iota^2 \\
& \qquad + (1+\frac{1}{H})^H\sum_{k=1}^{K}\Big(\overline{V}_{m,H+1}^{k} - \underline{V}_{m,H+1}^{k}\Big)(s_{H+1}^k) \\
\lesssim & d_{max}H^3S\iota^2 + H^3\sqrt{SAK\iota} + H^3\sqrt{S\mathcal{T}_K}\iota^2.
\end{aligned} \tag{68}$$

Following from Lemma 2, the following inequality holds with probability at least $1-\delta$:

$$\max_{m\in[M]}\sum_{k=1}^{K}\Big(V_{m,1}^{\dagger,\pi_{-m,k}} - V_{m,1}^{\pi_k}\Big)(s_1) \lesssim d_{max}H^3S\iota^2 + H^3\sqrt{SAK\iota} + H^3\sqrt{S\mathcal{T}_K}\iota^2. \tag{69}$$

By the definition of policy π, we have the following equation when $K \geq d_{max}^2S\iota^3$:

$$\max_{m\in[M]}\Big(V_{m,1}^{\dagger,\pi_{-m}} - V_{m,1}^{\pi}\Big)(s_1) \lesssim H^3\sqrt{S\mathcal{T}_K/K^2}\iota^2 + H^3\sqrt{SA\iota/K}. \tag{70}$$

■

E.3.1. SUPPORTING DETAILS

Proof [Proof of Lemma 10.] Consider any fixed pair (m,h), we define the following set $X_n(s)$:

$$X_n(s) = \Big\{x : s_h^x = s, \underline{n}_{m,h}^x(s) \geq n\Big\}. \tag{71}$$

Intuitively speaking, it collects episodes where (h,s) is visited and the n-th visit of (h,s) is usuable.

Rearranging the summation gives:

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{i=1}^{\underline{n}_k}\alpha_{\underline{n}_k}^{i}\Big(\overline{V}_{m,h+1}^{k_i}-\underline{V}_{m,h+1}^{k_i}\Big)(s_{h+1}^{k_i})\\ =&\sum_{k=1}^{K}\Big(\overline{V}_{m,h+1}^{k}-\underline{V}_{m,h+1}^{k}\Big)(s_{h+1}^{k})\sum_{x\in X_{\overline{n}_k}(s_h^k)}\alpha_{\underline{n}_x}^{\overline{n}_k}.\end{aligned} \tag{72}$$

For any episode $x \in X_{\overline{n}_k}(s_h^k)$, there are at most d_{max} unreceived and unusable visits of (h, s_h^k) according to Lemma 7. So we have $\underline{n}_x \geq \overline{n}_x - d_{max} + 1$. On the other hand, according to the definition of $X_{\overline{n}_k}(s_h^k)$, $\underline{n}_x \geq \overline{n}_k$. This gives:

$$\overline{n}_x \geq \underline{n}_x \geq \max\{\overline{n}_k, \overline{n}_x - d_{max} + 1\}. \tag{73}$$

Notice here $\overline{n}_x$ strictly increases with x.

Based on the above observations, we conclude that for the i-th episode ($i \leq d_{max}$) in $X_{\overline{n}_k}(s_h^k)$, if denoted as x, we have:

$$\underline{n}_x \geq \overline{n}_k. \tag{74}$$

For the i-th element ($i > d_{max}$) in $X_{\overline{n}_k}(s_h^k)$, if denoted as x, we have:

$$\begin{aligned}\underline{n}_x &\geq \overline{n}_x - d_{max} + 1 \geq \overline{n}_k + (i-1) - d_{max} + 1\\ &= \overline{n}_k + (i - d_{max}).\end{aligned} \tag{75}$$

Substituting the above two inequalities on $\underline{n}_x$ into Equation (72), then:

$$\begin{aligned}
&\sum_{k=1}^{K}\left(\overline{V}_{m,h+1}^{k}-\underline{V}_{m,h+1}^{k}\right)(s_{h+1}^{k})\sum_{x\in X_{\overline{n}_k}(s_h^k)}\alpha_{\underline{n}_x}^{\overline{n}_k}\\
\leq&\sum_{k=1}^{K}\left(\overline{V}_{m,h+1}^{k}-\underline{V}_{m,h+1}^{k}\right)(s_{h+1}^{k})\left[d_{max}\alpha_{\overline{n}_k}+\sum_{i=d_{max}+1}^{\infty}\alpha_{\overline{n}_k+i-d_{max}}^{\overline{n}_k}\right]\\
\leq&\sum_{k=1}^{K}\left(\overline{V}_{m,h+1}^{k}-\underline{V}_{m,h+1}^{k}\right)(s_{h+1}^{k})\left[d_{max}\alpha_{\overline{n}_k}+(1+\frac{1}{H})\right]\\
\leq&(1+\frac{1}{H})\sum_{k=1}^{K}\left(\overline{V}_{m,h+1}^{k}-\underline{V}_{m,h+1}^{k}\right)(s_{h+1}^{k})+d_{max}H\sum_{k=1}^{K}\alpha_{\overline{n}_k}\\
\leq&(1+\frac{1}{H})\sum_{k=1}^{K}\left(\overline{V}_{m,h+1}^{k}-\underline{V}_{m,h+1}^{k}\right)(s_{h+1}^{k})+2d_{max}H^2\sum_{s\in\mathcal{S}}\sum_{i=1}^{\overline{n}_{m,h}^{K}(s)}\frac{1}{i}\\
\leq&(1+\frac{1}{H})\sum_{k=1}^{K}\left(\overline{V}_{m,h+1}^{k}-\underline{V}_{m,h+1}^{k}\right)(s_{h+1}^{k})+2d_{max}H^2S(\ln K+1)\\
\leq&(1+\frac{1}{H})\sum_{k=1}^{K}\left(\overline{V}_{m,h+1}^{k}-\underline{V}_{m,h+1}^{k}\right)(s_{h+1}^{k})+4d_{max}H^2S\iota.
\end{aligned} \tag{76}$$

Here the second line follows the monotonicity of $\{\alpha_n^i\}_{n\in[K]}$, the third and fifth line follows Lemma 6, and the last line is because $\overline{n}_{m,h}^{K}(s)$. ■

Proof [Proof of Lemma 11.] Consider any fixed pair (m,h). We inherit the definition of $X_n(s)$ from the proof of Lemma 10. We first bound term $\sum_{k=1}^{K}\overline{\beta}_{\underline{n}_k}\cdot\mathbb{I}\left\{\underline{n}_k\geq d_{max}\right\}$

$$\begin{aligned}
&\sum_{k=1}^{K}\overline{\beta}_{\underline{n}_k}\cdot\mathbb{I}\left\{\underline{n}_k\geq d_{max}\right\}\\
=&4d_{max}H^2\sum_{k=1}^{K}\frac{\iota}{\underline{n}_k}\cdot\mathbb{I}\left\{\underline{n}_k\geq d_{max}\right\}+12H^2\sum_{k=1}^{K}\sqrt{\frac{\underline{n}_kA+\mathcal{T}_{m,h}^{\underline{n}_k}(s_h^k)}{\underline{n}_k^2}\iota}\cdot\mathbb{I}\left\{\underline{n}_k\geq d_{max}\right\}\\
\leq&4d_{max}H^2\sum_{k=1}^{K}\frac{\iota}{\underline{n}_k}\cdot\mathbb{I}\left\{\underline{n}_k\geq d_{max}\right\}+12H^2\sum_{k=1}^{K}\sqrt{\frac{A}{\underline{n}_k}\iota}\cdot\mathbb{I}\left\{\underline{n}_k\geq d_{max}\right\}\\
&\quad+12H^2\cdot\sum_{k=1}^{K}\sqrt{\frac{\mathcal{T}_{m,h}^{\underline{n}_k}(s_h^k)}{\underline{n}_k^2}\iota}\cdot\mathbb{I}\left\{\underline{n}_k\geq d_{max}\right\}.
\end{aligned} \tag{77}$$

For the first term:

$$d_{max}H^2\sum_{k=1}^{K}\frac{\iota}{\underline{n}_k}\cdot\mathbb{I}\left\{\underline{n}_k\geq d_{max}\right\}=d_{max}H^2\iota\sum_{s\in\mathcal{S}}\sum_{x\in X_{d_{max}}(s)}\frac{1}{\underline{n}_x}. \tag{78}$$

Here set $X_{d_{max}}(s)$ collects all episodes when (h, s) is visited and the d_{max}-th visit of (h, s) is usable. Following the analysis for Lemma 10, for the i-th element $(i \leq d_{max})$ in set $X_{d_{max}}(s)$, if denoted as x, we have:

$$\underline{n}_x \geq d_{max}. \tag{79}$$

For the i-th element $(i > d_{max})$ in $X_{d_{max}}(s)$, if denoted as x, we have:

$$\underline{n}_x \geq (d_{max} + i) - d_{max} = i. \tag{80}$$

These inequalities lead to:

$$\sum_{x \in X_{d_{max}}(s)} \frac{1}{\underline{n}_x} \leq 1 + \sum_{i=d_{max}+1}^{\overline{n}_{m,h}^K(s)} \frac{1}{i} \leq \ln \overline{n}_{m,h}^K(s) + 2 \leq 3\iota, \tag{81}$$

where the first inequality is due to the monotinicity of $1/n$. So the first term is bounded as follows:

$$d_{max} H^2 \sum_{k=1}^{K} \frac{\iota}{\underline{n}_k} \cdot \mathbb{I}\Big\{\underline{n}_k \geq d_{max}\Big\} \leq 3 d_{max} H^2 S \iota^2. \tag{82}$$

For the second term:

$$\begin{aligned} H^2 \sum_{k=1}^{K} \sqrt{\frac{A}{\underline{n}_k}\iota} \cdot \mathbb{I}\Big\{\underline{n}_k \geq d_{max}\Big\} \leq & H^2 \sqrt{A\iota} \cdot \sum_{s \in \mathcal{S}} \sum_{x \in X_{d_{max}}(s)} \sqrt{\frac{1}{\underline{n}_x}} \\ \leq & H^2 \sqrt{A\iota} \cdot \sum_{s \in \mathcal{S}} \Big[1 + \sum_{i=d_{max}+1}^{\overline{n}_{m,h}^K(s)} \frac{1}{\sqrt{i}}\Big] \\ \leq & 2H^2 \sqrt{A\iota} \sum_{s \in \mathcal{S}} (\sqrt{\overline{n}_{m,h}^K(s)} + 1) \\ \leq & 2H^2 \sqrt{A\iota} \sqrt{S \sum_{s \in \mathcal{S}} \overline{n}_{m,h}^K(s)} + 2H^2 S \sqrt{A\iota} \\ \leq & 2H^2 \sqrt{SAK\iota} + 2H^2 S \sqrt{A\iota} \\ \leq & 4H^2 \sqrt{SAK\iota}. \end{aligned} \tag{83}$$

Here the second inequalities follows similar analysis of the first term, the fifth inequality holds because $\sum_{s \in \mathcal{S}} \overline{n}_{m,h}^K(s) = K$ and the last inequality holds because $K \geq d_{max}^2 S \iota^3$.

For the third term:

$$
\begin{aligned}
H^2 \sum_{k=1}^{K} \sqrt{\frac{\mathcal{T}_{m,h}^{\underline{n}_k}(s_h^k)}{\underline{n}_k^2}\iota} \cdot \mathbb{I}\left\{\underline{n}_k > d_{max}\right\} \leq & H^2 \sum_{s\in\mathcal{S}} \sum_{x\in X_{d_{max}}(s)} \sqrt{\frac{\mathcal{T}_{m,h}^{\underline{n}_x}(s)}{\underline{n}_x^2}\iota} \\
\leq & H^2 \sum_{s\in\mathcal{S}} \sqrt{\mathcal{T}_{m,h}^{\underline{n}_{m,h}^K(s)}(s)\iota} \sum_{x\in X_{d_{max}}(s)} \frac{1}{\underline{n}_x} \\
\leq & H^2 \sum_{s\in\mathcal{S}} \sqrt{\mathcal{T}_{m,h}^{\underline{n}_{m,h}^K(s)}(s)\iota} \cdot \left[1 + \sum_{i=d_{max}+1}^{\overline{n}_{m,h}^K(s)} \frac{1}{i}\right] \\
\leq & 2H^2 \sum_{s\in\mathcal{S}} \sqrt{\mathcal{T}_{m,h}^{\underline{n}_{m,h}^K(s)}(s)\iota^2} \\
\leq & 2H^2 \sqrt{S\mathcal{T}_K\iota^2}.
\end{aligned}
\tag{84}
$$

Here the last line utilizes the definition of $\mathcal{T}_K$.

Finally, we bound the term $\sum_{k=1}^{K} \underline{\beta}_{\underline{n}_k} \cdot \mathbb{I}\left\{\underline{n}_k > d_{max}\right\}$ as follows:

$$
\begin{aligned}
\sum_{k=1}^{K} \underline{\beta}_{\underline{n}_k} \cdot \mathbb{I}\left\{\underline{n}_k > d_{max}\right\} = & 2H^2 \sum_{k=1}^{K} \sqrt{\frac{\iota}{\underline{n}_k}}\, \mathbb{I}\left\{\underline{n}_k > d_{max}\right\} \\
\leq & 4H^2\sqrt{\iota} \sum_{s\in\mathcal{S}} (\sqrt{\overline{n}_{m,h}^K(s)} + 1) \\
\leq & 8H^2\sqrt{SK\iota}
\end{aligned}
\tag{85}
$$

Here the second inequality utilizes the analysis of Equation (83).

Combining all above four terms, we have:

$$
\begin{aligned}
& \sum_{k=1}^{K} (\overline{\beta}_{\underline{n}_k} + \underline{\beta}_{\underline{n}_k}) \cdot \mathbb{I}\left\{\underline{n}_k \geq d_{max}\right\} \\
\leq & 12 d_{max} H^2 S\iota^2 + 48H^2\sqrt{SAK\iota} + 8H^2\sqrt{SK\iota} + 24H^2\sqrt{S\mathcal{T}_K\iota^2} \\
\leq & 12 d_{max} H^2 S\iota^2 + 56H^2\sqrt{SAK\iota} + 24H^2\sqrt{S\mathcal{T}_K\iota^2}
\end{aligned}
\tag{86}
$$

■

Appendix F. Performance Guarantee for DA-MAVL with Reward Skipping

Proof of Theorem 2 is under assumption 2 and consists of four steps. We inherit all notations from previous section, except that they refer to variables in DA-MAVL with reward skipping.

STEP ONE: Bound the 'Policy Optimization Regret'. For any pair (m, h, s, n), upper bound of $R_{m,h}^n(s)$ is established:

Lemma 12 *Suppose assumption 2 holds. For* $\forall(m,h,s,n) \in [M]\times[H]\times\mathcal{S}\times[K]$*, the following inequality holds with probability at least* $1-\delta/2$*:*

$$R^n_{m,h}(s) \leq 20H^2C\sqrt{\frac{\mathcal{T}_n}{n^2}}\iota + 14H^2\sqrt{\frac{A}{n}}\iota. \tag{87}$$

Lemma 12 extends Lemma 1 to cases with infinite delays. In the proof of this lemma, we have to upper bound the regret by **the largest possible delay** and **the number of reward skips** instead of the maximum delay d_{max}, since the delays may be infinite. We then highlight their upper bounds, i.e. Lemma 16 and Lemma 17, which play significant roles in showing that the influence of the delays can be bounded by term $H^2C\sqrt{\mathcal{T}^n_{m,h}(s)/n^2}$.

STEP TWO: Optimism and Pessimism. Utilizing regret $R^n_{m,h}(s)$, we carefully design the bonuses (Equation (21)) and show that value estimates in Algorithm 9 are optimistic and pessimistic:

Lemma 13 *For* $\forall(m,h,s,k) \in [M]\times[H]\times\mathcal{S}\times[K]$*, the following inequality holds with probability at least* $1-\delta$*:*

$$\overline{V}^k_{m,h}(s) \geq V^{\dagger,\pi^k_{-m,h}}_{m,h}(s), \quad \underline{V}^k_{m,h}(s) \leq V^{\pi^k_h}_{m,h}(s). \tag{88}$$

In the proof of this lemma, we separately consider the skipped and unskipped visits. Utilizing the upper bounds on the largest possible delay, we can show the bonuses can make up for the performance degradation of skipping visits and ensure optimism and pessimism.

STEP THREE: Next, we bound the gap between the optimistic and pessimistic value estimates:

$$\sum_{k=1}^{K}(\overline{V}^k_{m,1} - \underline{V}^k_{m,1})(s^k_h). \tag{89}$$

The gap, together with Optimism and Pessimism, leads to the following bound:

Lemma 14 *Suppose assumption 2 holds. For* $\forall\delta \in (0,1)$ *and* $\forall K \in \mathbb{N}$*, let* $\iota = \log\left(4HSAK/\delta\right)$*. Let policy* π *be the output of Algorithm 2 after running Algorithm 9 for* K *episodes. The following equation holds with probability at least* $1-\delta$*:*

$$\sum_{k=1}^{K}\left(V^{\dagger,\pi_{-m,k}}_{m,1} - V^{\pi_k}_{m,1}\right)(s_1) \lesssim CH^3 \max_h \sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}^{\overline{n}^K_{m,h}(s)}_{m,h}(s)\iota^2} + H^3\sqrt{SAK\iota}. \tag{90}$$

STEP FOUR: Bound the CCE-gap. Finally, we give an upper bound on $\sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}^{\overline{n}^K_{m,h}(s)}_{m,h}(s)}$.

Lemma 15 *For* $\forall(m,h,s,n) \in [M]\times[H]\times\mathcal{S}\times[K]$*, the following inequality hold:*

$$\sqrt{\mathcal{T}^n_{m,h}(s)} \leq 2\min_{\mathcal{L}\in[K]}\left\{|\mathcal{L}| + \sqrt{\mathcal{T}^{n,\mathcal{L}}_{m,h}(s)}\right\} + 64C^2. \tag{91}$$

Consequently, we can show that term $\sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}^{\overline{n}^K_{m,h}(s)}_{m,h}(s)}$ can be upper bounded :

$$\sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}^{\overline{n}^K_{m,h}(s)}_{m,h}(s)} \leq 2\min_{\mathcal{L}\in[K]}\left\{S|\mathcal{L}| + \sqrt{S\mathcal{T}^{K,\mathcal{L}}_{m,h}}\right\} + 64C^2S. \tag{92}$$

Intuitively, this upper bound shows that the influence of the skipped large delays is only reflected by some constant $|\mathcal{L}|$. As a direct consequence, we complete the proof of Theorem 2.

F.1. Step One: Proof of Lemma 12

Consider any fixed pair (m, h, s, n). Recall that $\mathcal{O}_n$ denote that set of skipped visits of (h, s) during the first n visits of (h, s). We first decompose the policy optimization regret R_n as follows:

$$\begin{aligned} R_n &= H \sum_{i=1}^{n} \alpha_n^i \Big[l_i(a_{k_i}) - \bar{l}_i(a^*) \Big] \\ &= H \sum_{i \in [n] \setminus \mathcal{O}_n} \alpha_n^i \Big[l_i(a_{k_i}) - \bar{l}_i(a^*) \Big] + H \sum_{i \in \mathcal{O}_n} \alpha_n^i \Big[l_i(a_{k_i}) - \bar{l}_i(a^*) \Big] \\ &\le \alpha_n H |\mathcal{O}_n| + H \sum_{i \in [n]} \alpha_n^i \Big[l_i(a_{k_i}) - \bar{l}_i(a^*) \Big] \cdot \mathbb{I}\{ i \notin \mathcal{O}_n \}. \end{aligned} \tag{93}$$

Here the last line holds because of the monotonicity of $\{\alpha_n^i\}_{i \in [n]}$.

Consider the delays and losses defined as follows:

$$\begin{aligned} {d'}_{m,h}^{i}(s) &= \begin{cases} k_h^{\min_j \{ i \in \mathcal{O}_j \}}(s) - k_h^i(s) - 1, & i \in \mathcal{O}_n \\ d_{m,h}^i(s), & i \notin \mathcal{O}_n \end{cases}, \\ {l'}_{m,h}^{i}(s,a) &= \begin{cases} 0, & i \in \mathcal{O}_n \\ l_{m,h}^i(s,a), & i \notin \mathcal{O}_n \end{cases}, \\ \bar{l'}_{m,h}^{i}(s,a) &= \mathbb{E}[{l'}_{m,h}^{i}(s,a)] = \begin{cases} 0, & i \in \mathcal{O}_n \\ \bar{l}_{m,h}^i(s,a), & i \notin \mathcal{O}_n \end{cases}. \end{aligned} \tag{94}$$

When a fixed (m, h, s) is considered in the context, the above notations are abbreviated as ${d'}_i$, ${l'}_i(a)$ and $\bar{l'}_i(a)$.

Then the second term of R_n is exactly the policy optimization regret (without reward skipping) with delays $\{{d'}_i\}_{i \in [n]}$ and losses $\{{l'}_i(a)\}_{i \in [n], a \in \mathcal{A}_m}$ and $\{\bar{l'}_i(a)\}_{i \in [n], a \in \mathcal{A}_m}$. Let $d'_{max} = \max_{i \in [n]} \min \{ {d'}_i, n - i \}$ be the maximum delay during the first n visits of (h, s). Then we directly apply results of Lemma 1 and get the following with probability at least $1 - \delta/2$:

$$\begin{aligned} R_n &= \alpha_n H |\mathcal{O}_n| + H \sum_{i \in [n]} \alpha_n^i \Big[l_i(a_{k_i}) - \bar{l}_i(a^*) \Big] \cdot \mathbb{I}\{ i \notin \mathcal{O}_n \} \\ &\le 2H^2 \frac{|\mathcal{O}_n|}{n} + 12 H^2 \sqrt{\frac{nA + \mathcal{T}_n}{n^2} \iota} + 2H^2 \frac{d'_{max}}{n} \iota. \end{aligned} \tag{95}$$

We then upper bound $\mathcal{O}_n$ and d'_{max}:

Lemma 16 *For* $\forall (m, h, s, n, i) \in [M] \times [H] \times \mathcal{S} \times [K] \times [n]$, *let* ${d'}_{max,m,h}^{n}(s)$ *denote the maximal delay of* (h, s) *during the first* n *visits of* (h, s)*:*

$${d'}_{max,m,h}^{n}(s) = \max_{i \in [n]} \min \{ {d'}_{m,h}^{i}(s), n - i \}. \tag{96}$$

the following equation holds:

$$\begin{aligned} {d'}_{max,m,h}^{n}(s) &\le \sqrt[4]{4 \mathcal{T}_{m,h}^{n}(s)} + 1, \\ \phi_{m,h}^{i,n}(s) &\le \sqrt{\mathcal{T}_{m,h}^{n}(s)} + \sqrt[4]{4 \mathcal{T}_{m,h}^{n}(s)} + 1. \end{aligned} \tag{97}$$

Lemma 17 *For* $\forall(m,h,s,n)\in[M]\times[H]\times\mathcal{S}\times[K]$*, the following equation holds:*

$$|\mathcal{O}^n_{m,h}(s)| \le 2C\sqrt{\mathcal{T}^n_{m,h}(s)}. \tag{98}$$

Finally, with Lemma 16 and Lemma 17, we have:

$$\begin{aligned} R_n &\le 4H^2C\sqrt{\frac{\mathcal{T}_n}{n^2}} + 12H^2\sqrt{\frac{A}{n}}\iota + 12H^2\sqrt{\frac{\mathcal{T}_n}{n^2}}\iota + 4H^2\sqrt[4]{\frac{\mathcal{T}_n}{n^4}}\iota + \frac{2H^2}{n}\iota \\ &\le 20H^2C\sqrt{\frac{\mathcal{T}_n}{n^2}}\iota + 12H^2\sqrt{\frac{A}{n}}\iota + \frac{2H^2}{n}\iota. \\ &\le 20H^2C\sqrt{\frac{\mathcal{T}_n}{n^2}}\iota + 14H^2\sqrt{\frac{A}{n}}\iota \end{aligned} \tag{99}$$

F.1.1. SUPPORTING DETAILS

Proof [Proof of Lemma 16] Consider any fixed pair (m,h,s,n). Recall $\phi_{i,n}$ is the skipping metric of the i-th visit when the n-th visit happens. It can be written as in Equation (23). Let d'_{max} denote $d'^{n}_{max,m,h}(s)$. Let $i_0 = \arg\max_{i\in[n]}\min\{d'_i, n-i\}$ denote the visit with the largest delay.

We first bound $\phi_{i,n}$. Since $\phi_{i,n}$ grows monotonically with $\min\{d'_i, n-i\}$, we only need to consider $\phi_{i_o,n}$. If $i_0 \notin \mathcal{O}_n$, namely the i_0-th visit of (h,s) has not been skipped, then:

$$\phi_{i_0,n} \le \sqrt{\mathcal{T}_n}. \tag{100}$$

On the other hand, from the definition of $\phi_{i_0,n}$,

$$\phi_{i_0,n} = \sum_{j=1}^{d'_{max}} j \ge \frac{d'^2_{max}}{2}. \tag{101}$$

This gives $d'_{max} \le \sqrt[4]{4\mathcal{T}_n}$.

If $i_0 \in \mathcal{O}_n$, namely the i_0-th visit of (h,s) has been skipped, suppose it is skipped during the n_0-th visit of (h,s). Then $d'_{max} = n_0 - i_0$. Then:

$$\phi_{i_0,n_0-1} \le \sqrt{\mathcal{T}_{n_0-1}} \le \sqrt{\mathcal{T}_n}. \tag{102}$$

On the other hand, from the definition of ϕ_{i_0,n_0-1},

$$\phi_{i_0,n_0-1} = \sum_{j=1}^{n_0-1-i_0} j \ge \frac{(n_0-1-i_0)^2}{2} = \frac{(d'_{max}-1)^2}{2}. \tag{103}$$

This is equivalent to $d'_{max} \le \sqrt[4]{4\mathcal{T}_n} + 1$. Combining Equation (102) and (103) we have:

$$\begin{aligned} \phi_{i_0,n} =& \phi_{i_0,n_0} = \phi_{i_0,n_0-1} + d'_{max} \\ \le& \sqrt{\mathcal{T}_n} + \sqrt[4]{4\mathcal{T}_n} + 1. \end{aligned} \tag{104}$$

Here the equation is because $\phi_{i_0,n}$ stops increasing after the i_0-th visit is skipped.

Combining the above two cases finishes the proof. ∎

Proof [Proof of Lemma 17] Proof of this Lemma largely follows proof of Lemma 4 in Zimmert and Seldin (2020). Consider any fixed pair (m, h, s, n). Denote all visits in $\mathcal{O}_n$ in order as $i_1, i_2, \ldots, i_{|\mathcal{O}_n|}$. And suppose they are skipped during visits $i'_1, i'_2, \ldots, i'_{|\mathcal{O}_n|}$. Then following the skipping rule, we have:

$$\begin{aligned} \phi_{i_x,n} \geq & \sqrt{\mathcal{T}_{i'_x}} \geq \sqrt{\sum_{y=1}^{i_x} \phi_{y,i'_x}/C} \geq \sqrt{\sum_{y=1}^{x} \phi_{i_y,i'_x}/C} = \sqrt{\sum_{y=1}^{x} \phi_{i_y,n}/C} \\ \geq & \frac{\sqrt{\phi_{i_x,n} + \sum_{y=1}^{x-1} \phi_{i_y,n}}}{\sqrt{C}}. \end{aligned} \tag{105}$$

Here the second inequality is because of Lemma 18, the first equation is because $\phi_{i,n}$ stops increasing after the i-th visit is skipped.

Solve the inequality on $\phi_{i_x,n}$ gives:

$$\phi_{i_x,n} \geq \frac{1 + \sqrt{1 + 4C\sum_{y=1}^{x-1} \phi_{i_y,n}}}{2C}. \tag{106}$$

By induction, we can easily prove:

$$\phi_{i_x,n} \geq \frac{x}{2C}. \tag{107}$$

This directly gives:

$$\sum_{x=1}^{|\mathcal{O}_n|} \phi_{i_x,n} \geq \frac{|\mathcal{O}_n|^2}{4C}. \tag{108}$$

On the other hand, we have $\sum_{x=1}^{|\mathcal{O}_n|} \phi_{i_x,n} \leq \sum_{x=1}^{n} \phi_{x,n} \leq C\mathcal{T}_n$ from Lemma 18. Combining the two inequalities gives the desired result. ■

Lemma 18 *For* $\forall(m, h, s, n) \in [M] \times [H] \times \mathcal{S} \times [K]$, *the following equation holds:*

$$\sum_{i=1}^{n} \phi_{m,h}^{i,n}(s) \leq C\mathcal{T}_{m,h}^{n}(s). \tag{109}$$

Proof [Proof of Lemma 18] Consider any fixed pair (m, h, s, n). Recall that sequence $\{\phi_{i,n}\}_{i,n\in[K]}$ in algorithm 9 can be written:

$$\phi_{i,n} = \sum_{j=i+1}^{n} (j - i) \cdot \mathbb{I}\{i \in \arg \mathcal{M}_{k_j}\}. \tag{110}$$

Then simple algebra gives:

$$\begin{aligned}\sum_{i=1}^{n-1}\phi_{i,n} &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(j-i)\cdot\mathbb{I}\{i\in\arg\mathcal{M}_{k_j}\}\\ &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{I}\{i\in\arg\mathcal{M}_{k_j}\}\sum_{x=i+1}^{j}1\\ &= \sum_{i=1}^{n-1}\sum_{x=i+1}^{n}\sum_{j=x}^{n}\mathbb{I}\{i\in\arg\mathcal{M}_{k_j}\}\\ &= \sum_{x=2}^{n}\sum_{i=1}^{x-1}\sum_{j=x}^{n}\mathbb{I}\{i\in\arg\mathcal{M}_{k_j}\}.\end{aligned} \tag{111}$$

Notice that $i\in\arg\mathcal{M}_{k_j}$ implies the i-th visit is not usuable when the j-th visit happens. Since $x\geq i$, we have:

$$x\in\arg\mathcal{M}_{k_j}. \tag{112}$$

This observation gives:

$$\sum_{i=1}^{n-1}\phi_{i,n} = \sum_{x=2}^{n}\sum_{j=x}^{n}\mathbb{I}\Big\{x\in\arg\mathcal{M}_{k_j}\Big\}\cdot\Big[\sum_{i=1}^{x-1}\mathbb{I}\Big\{i\in\arg\mathcal{M}_{k_j}\Big\}\Big]. \tag{113}$$

On the other hand, we have:

$$\begin{aligned}\sum_{i=1}^{x-1}\mathbb{I}\{i\in\arg\mathcal{M}_{k_j}\} \leq \sum_{i=1}^{j-1}\mathbb{I}\{i\in\arg\mathcal{M}_{k_j}\} &= \Big|\{i\leq j-1: d_i+k_i\geq k_j\}\Big|\\ &\leq \Big|\{i\leq j: d_i+i\geq j\}\Big| \leq C.\end{aligned} \tag{114}$$

where the second inequality follows Assumption 2 and the fact that $j-i\leq k_j-k_i$. Substituting into Equation (113):

$$\begin{aligned}\sum_{i=1}^{n}\phi_{i,n} &\leq C\sum_{x=2}^{n}\sum_{j=x}^{n}\mathbb{I}\Big\{x\in\arg\mathcal{M}_{k_j}\Big\}\\ &= C\sum_{j=2}^{n}\sum_{x=2}^{j}\mathbb{I}\Big\{x\in\arg\mathcal{M}_{k_j}\Big\}\\ &= C\sum_{j=2}^{n}|\mathcal{M}_{k_j}|\\ &\leq C\mathcal{T}_n.\end{aligned} \tag{115}$$

■

F.2. Step Two: Proof of Lemma 13

Proof [Proof of lemma 13] Consider any fixed pari (m, h, s). Conditioned on the successful event of Lemma 12, which holds for probability at least $1 - \delta/2$, we first prove the optimism part by induction. For $k = 0$, it clear that for any (m, h, s) :

$$\overline{V}^{0}_{m,h}(s) = H + 1 - h \geq V^{\dagger,\pi^{1}_{-m,h}}_{m,h}(s). \tag{116}$$

Let $N_k = \max_m \underline{n}^{k}_{m,h}(s)$. Suppose the lemma holds for all $k' < k$. Then for episode k:

$$\begin{aligned}
\overline{V}^{k}_{m,h}(s) &= \alpha^{0}_{\underline{n}_k} \cdot H + \sum_{i=1}^{\underline{n}_k} \alpha^{i}_{\underline{n}_k} \left(r^{k_i}_{m,h} + \overline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1}) \right) \cdot \mathbb{I}\{i \notin \mathcal{O}_{\overline{n}_k}\} \\
&\quad + \sum_{i=1}^{\underline{n}_k} \alpha^{i}_{\underline{n}_k} H \cdot \mathbb{I}\{i \in \mathcal{O}_{\overline{n}_k}\} + \overline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s) \\
&\geq \alpha^{0}_{\underline{n}_k} \cdot H + \sum_{i=1}^{\underline{n}_k} \alpha^{i}_{\underline{n}_k} \left(r^{k_i}_{m,h} + \overline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1}) \right) + \overline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s) \\
&= \alpha^{0}_{\underline{n}_k} \cdot H + 4H^2 C \frac{\sqrt{\mathcal{T}^{\overline{n}_k}_{m,h}(s)}}{\underline{n}_k} \iota + 4H^2 \sqrt{\frac{A}{\underline{n}_k} \iota} \\
&\quad + \sum_{i=1}^{\underline{n}_k} \alpha^{i}_{\underline{n}_k} \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a^*, a_{-m,h}) \\ a_{-m,h} \sim \hat{\pi}^{k_i}_{-m,h},\, s' \sim \mathbb{P}_h(s, \boldsymbol{a})}} \left[r_{m,h}(s, \boldsymbol{a}) + \overline{V}^{k_i}_{m,h+1}(s') \right].
\end{aligned} \tag{117}$$

Here the last line follows directly from proof of Lemma 12.

On the other hand, the following holds according to Lemma 2:

$$\begin{aligned}
V^{\dagger,\pi^{k}_{-m,h}}_{m,h}(s) &\leq \sum_{i=1}^{\underline{n}_k} \alpha^{i}_{\underline{n}_k} \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a^*, a_{-m,h}) \\ a \sim \mu_h, a_{-m,h} \sim \hat{\pi}^{k_i}_{-m,h}}} \left(r_{m,h}(s, \boldsymbol{a}) + \overline{V}^{k_i}_{m,h+1}(s') \right) + \frac{2(N_k - \underline{n}_k)H^2}{N_k} \\
&\leq \sum_{i=1}^{\underline{n}_k} \alpha^{i}_{\underline{n}_k} \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a^*, a_{-m,h}) \\ a \sim \mu_h, a_{-m,h} \sim \hat{\pi}^{k_i}_{-m,h}}} \left(r_{m,h}(s, \boldsymbol{a}) + \overline{V}^{k_i}_{m,h+1}(s') \right) + 2H^2 \frac{\overline{n}_k - \underline{n}_k}{\overline{n}_k}.
\end{aligned} \tag{118}$$

Here the last line is because $\overline{n}_k > N_k$. From Lemma 7 and Lemma 16, we have:

$$\overline{n}_k - \underline{n}_k = |\mathcal{M}_k| + 1 \leq d'^{\overline{n}_k}_{max,m,h}(s) + 1 \leq \sqrt[4]{4\mathcal{T}^{\overline{n}_k}_{m,h}(s)} + 2. \tag{119}$$

This leads to:

$$\begin{aligned} V_{m,h}^{\dagger,\pi_{-m,h}^k}(s) &\le \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a^*,a_{-m,h}) \\ a\sim\mu_h, a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i}}} \left(r_{m,h}(s,\boldsymbol{a}) + \overline{V}_{m,h+1}^{k_i}(s')\right) + 2H^2 \frac{\sqrt[4]{4\mathcal{T}_{m,h}^{\overline{n}_k}(s)} + 2}{\overline{n}_k} \\ &\le \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a^*,a_{-m,h}) \\ a\sim\mu_h, a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i}}} \left(r_{m,h}(s,\boldsymbol{a}) + \overline{V}_{m,h+1}^{k_i}(s')\right) + 4H^2 \frac{\sqrt{\mathcal{T}_{m,h}^{\overline{n}_k}(s)}}{\overline{n}_k} + 4H^2\frac{1}{\overline{n}_k} \\ &\le \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \mathop{\mathbb{E}}_{\substack{\boldsymbol{a}=(a^*,a_{-m,h}) \\ a\sim\mu_h, a_{-m,h}\sim\hat{\pi}_{-m,h}^{k_i}}} \left(r_{m,h}(s,\boldsymbol{a}) + \overline{V}_{m,h+1}^{k_i}(s')\right) + 4H^2 \frac{\sqrt{\mathcal{T}_{m,h}^{\overline{n}_k}(s)}}{\underline{n}_k} + 4H^2\sqrt{\frac{A}{\overline{n}_k}\iota}. \end{aligned} \tag{120}$$

which directly leads to $\overline{V}_{m,h}^k(s) \ge V_{m,h}^{\dagger,\pi_{-m,h}^k}(s)$ and finishes the induction.

We now prove the pessimism part of the lemma. For $\forall(m,h,s,k)\in[M]\times[H]\times\mathcal{S}\times[K]$, the following hold with probability at least $1-\delta/(2MHSK)$:

$$\begin{aligned} \underline{V}_{m,h}^k(s) &= \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \left(r_{m,h}^{k_i} + \underline{V}_{m,h+1}^{k_i}(s_{h+1}^{k_i})\right)\cdot\mathbb{I}\{i\notin\mathcal{O}_{\overline{n}_k}\} - \underline{\beta}_{\underline{n}_k,\overline{n}_k} \\ &\le \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \left(r_{m,h}^{k_i} + \underline{V}_{m,h+1}^{k_i}(s_{h+1}^{k_i})\right) - \underline{\beta}_{\underline{n}_k,\overline{n}_k} \\ &\le \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\hat{\pi}_h^{k_i}, s'\sim\mathbb{P}_h(\cdot|s,\boldsymbol{a})} \left(r_{m,h}(s,\boldsymbol{a}) + \underline{V}_{m,h+1}^{k_i}(s')\right) \\ &\quad + 2H^2\frac{\overline{n}_k-\underline{n}_k}{\underline{n}_k} + 2\sqrt{\frac{H^3}{\underline{n}_k}\iota} - \underline{\beta}_{\underline{n}_k,\overline{n}_k} \\ &\le \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\hat{\pi}_h^{k_i}, s'\sim\mathbb{P}_h(\cdot|s,\boldsymbol{a})} \left(r_{m,h}(s,\boldsymbol{a}) + \underline{V}_{m,h+1}^{k_i}(s')\right) \\ &\quad + 2H^2\frac{\sqrt[4]{4\mathcal{T}_{m,h}^{\overline{n}_k}(s)}+2}{\underline{n}_k} + 2\sqrt{\frac{H^3}{\underline{n}_k}\iota} - \underline{\beta}_{\underline{n}_k,\overline{n}_k} \\ &= V_{m,h}^{\pi_h^k}(s). \end{aligned} \tag{121}$$

Here the third line follows directly from the proof of Lemma 2, while the last line is from the definition of $\underline{\beta}_{\underline{n}_k,\overline{n}_k}$ and Equation (119). Finally by taking the union bound over all $(m,h,s,k)\in[M]\times[H]\times\mathcal{S}\times[K]$, we finish the proof. ■

F.3. Step Three: Proof of Lemma 14

Consider any fixed pair (m,h). For episode k, we slightly overload the notations and let $\underline{n}_k = \underline{n}^k_{m,h}(s^k_h)$, $\underline{n}_k = \overline{n}^k_{m,h}(s^k_h)$, $k_i = k^i_h(s^k_h)$, $d'_{max}(s) = d'^{\overline{n}^K_{m,h}(s)}_{max,m,h}(s) \leq \sqrt[4]{4\mathcal{T}^{\overline{n}^K_{m,h}(s)}_{m,h}}(s) + 1$. Then:

$$
\begin{aligned}
&\sum_{k=1}^{K}(\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h) \\
=&\sum_{k=1}^{K}(\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h)\cdot\mathbb{I}\Big\{\underline{n}_k < d'_{max}(s^k_h)\Big\} + \sum_{k=1}^{K}(\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h)\cdot\mathbb{I}\Big\{\underline{n}_k \geq d'_{max}(s^k_h)\Big\} \\
\leq& H\sum_{k=1}^{K}\mathbb{I}\Big\{\underline{n}_k < d'_{max}(s^k_h)\Big\} + \sum_{k=1}^{K}(\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h)\cdot\mathbb{I}\Big\{\underline{n}_k \geq d'_{max}(s^k_h)\Big\}.
\end{aligned}
\tag{122}
$$

To bound the second term, notice that:

$$
\begin{aligned}
&(\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h) \\
\leq&\alpha^0_{\underline{n}_k}H + \sum_{i=1}^{\underline{n}_k}\alpha^i_{\underline{n}_k}H\cdot\mathbb{I}\{i\in\mathcal{O}_{\overline{n}_k}\} + \sum_{i=1}^{\underline{n}_k}\alpha^i_{\underline{n}_k}\Big(r^{k_i}_{m,h} + \overline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1})\Big)\cdot\mathbb{I}\{i\notin\mathcal{O}_{\overline{n}_k}\} + \overline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s^k_h) \\
&\qquad - \sum_{i=1}^{\underline{n}_k}\alpha^i_{\underline{n}_k}\Big(r^{k_i}_{m,h} + \underline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1})\Big)\cdot\mathbb{I}\{i\notin\mathcal{O}_{\overline{n}_k}\} + \underline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s^k_h) \\
\leq&\alpha^0_{\underline{n}_k}H + \overline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s^k_h) + \underline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s^k_h) + \sum_{i=1}^{\underline{n}_k}\alpha^i_{\underline{n}_k}H\cdot\mathbb{I}\{i\in\mathcal{O}_{\overline{n}_k}\} \\
&\qquad + \sum_{i=1}^{\underline{n}_k}\alpha^i_{\underline{n}_k}\Big(\overline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1}) - \underline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1})\Big)\cdot\mathbb{I}\{i\notin\mathcal{O}_{\overline{n}_k}\} \\
\leq&\alpha^0_{\underline{n}_k}H + \overline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s^k_h) + \underline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s^k_h) + 2\alpha_{\underline{n}_k}CH\sqrt{\mathcal{T}^{\overline{n}_k}_{m,h}(s^k_h)} \\
&\qquad + \sum_{i=1}^{\underline{n}_k}\alpha^i_{\underline{n}_k}\Big(\overline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1}) - \underline{V}^{k_i}_{m,h+1}(s^{k_i}_{h+1})\Big).
\end{aligned}
\tag{123}
$$

Here the last line follows Lemma 17. Substituting into Equation (122), we get:

$$
\begin{aligned}
&\sum_{k=1}^{K}(\overline{V}^k_{m,h} - \underline{V}^k_{m,h})(s^k_h) \\
\leq&\underbrace{H\sum_{k=1}^{K}\mathbb{I}\Big\{\underline{n}_k < d'_{max}(s^k_h)\Big\}}_{I_{3,1}} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{\underline{n}_k}\alpha^i_{\underline{n}_k}(\overline{V}^{k_i}_{m,h+1} - \underline{V}^{k_i}_{m,h+1})(s^{k_i}_{h+1})}_{I_{3,2}} \\
&\qquad + \underbrace{\sum_{k=1}^{K}(\overline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s^k_h) + \underline{\beta}^{\underline{n}_k,\overline{n}_k}_{m,h}(s^k_h) + 4CH^2\frac{\sqrt{\mathcal{T}^{\overline{n}_k}_{m,h}(s^k_h)}}{\underline{n}_k})\cdot\mathbb{I}\Big\{\underline{n}_k \geq d'_{max}(s^k_h)\Big\}}_{I_{3,3}}.
\end{aligned}
\tag{124}
$$

The three terms are bounded by the following Lemmas:

Lemma 19 *For any fixed pair* $(m,h) \in [M] \times [H]$,

$$I_{3,1} \le 4H \sum_{s\in\mathcal{S}} \sqrt[4]{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)} + 2HS. \tag{125}$$

Lemma 20 *For any fixed pair* $(m,h) \in [M] \times [H]$,

$$I_{3,2} \le (1+\frac{1}{H}) \sum_{k=1}^{K} \left(\overline{V}_{m,h+1}^k - \underline{V}_{m,h+1}^k\right)(s_{h+1}^k) + 8H^2 \sum_{s\in\mathcal{S}} \sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)\iota} + 4H^2 S\iota. \tag{126}$$

Lemma 21 *For any fixed pair* $(m,h) \in [M] \times [H]$,

$$I_{3,3} \le 96CH^2 \sum_{s\in\mathcal{S}} \sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)\iota^2} + 96H^2\sqrt{SAK\iota}. \tag{127}$$

This Lemma leads to:

$$\begin{aligned}
&\sum_{k=1}^{K} (\overline{V}_{m,h}^k - \underline{V}_{m,h}^k)(s_h^k) \\
\le&(1+\frac{1}{H}) \sum_{k=1}^{K} \left(\overline{V}_{m,h+1}^k - \underline{V}_{m,h+1}^k\right)(s_{h+1}^k) \\
&\quad + 108CH^2 \sum_{s\in\mathcal{S}} \sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)\iota^2} + 96H^2\sqrt{SAK\iota} + 6H^2S\iota
\end{aligned} \tag{128}$$

Iterating over h, we get:

$$\begin{aligned}
&\sum_{k=1}^{K} (\overline{V}_{m,1}^k - \underline{V}_{m,1}^k)(s_1) \\
\le&108CH^3 \max_h \sum_{s\in\mathcal{S}} \sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)\iota^2} + 96H^3\sqrt{SAK\iota} + 6H^3S\iota \\
\lesssim&CH^3 \max_h \sum_{s\in\mathcal{S}} \sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)\iota^2} + H^3\sqrt{SAK\iota}
\end{aligned} \tag{129}$$

Finally, conditioned on the successful event of Lemma 13, which has probability at least $1-\delta$, the Lemma 14 holds.

F.3.1. SUPPORTING DETAILS

Proof [Proof of Lemma 19] Consider any fixed pair (m,h). We inherit the definition of $X_n(s)$ from the proof of Lemma 10 to denote all episodes where (h,s) is visited and the n-th visit of (h,s) is received or been skipped. We then have:

$$I_{3,1} = H \sum_{k=1}^{K} \mathbb{I}\left\{\underline{n}_k < d'_{max}(s_h^k)\right\} = H \sum_{s\in\mathcal{S}} \sum_{x\in X_0(s)} \mathbb{I}\left\{\underline{n}_x < d'_{max}(s)\right\} \tag{130}$$

Notice that the $d'_{max}(s)$-th visit will be received or skipped when the $(2d'_{max}(s)+1)$-th visit happens. So we must have for every $s \in \mathcal{S}$:

$$\sum_{x \in X_0(s)} \mathbb{I}\left\{\underline{n}_x < d'_{max}(s)\right\} \leq 2d'_{max}(s) \tag{131}$$

Consequently,

$$I_{3,1} \leq H \sum_{s \in \mathcal{S}} 2d'_{max}(s) \leq 4H \sum_{s \in \mathcal{S}} \sqrt[4]{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)} + 2HS \tag{132}$$

The last line follows Lemma 16. ■

Proof [Proof of Lemma 20] Consider any fixed pair (m,h). We have:

$$\begin{aligned}
I_{3,2} =& \sum_{k=1}^{K} \sum_{i=1}^{\underline{n}_k} \alpha_{\underline{n}_k}^i (\overline{V}_{m,h+1}^{k_i} - \underline{V}_{m,h+1}^{k_i})(s_{h+1}^{k_i}) \\
\leq& \sum_{k=1}^{K} \left(\overline{V}_{m,h+1}^k - \underline{V}_{m,h+1}^k\right)(s_{h+1}^k)\left[d'_{max}(s_h^k)\alpha_{\overline{n}_k} + (1+\frac{1}{H})\right] \\
\leq& (1+\frac{1}{H}) \sum_{k=1}^{K} \left(\overline{V}_{m,h+1}^k - \underline{V}_{m,h+1}^k\right)(s_{h+1}^k) + 2H^2 \sum_{k=1}^{K} \frac{\sqrt[4]{4\mathcal{T}_{m,h}^{\overline{n}_K}(s_h^k)} + 1}{\overline{n}_k} \\
\leq& (1+\frac{1}{H}) \sum_{k=1}^{K} \left(\overline{V}_{m,h+1}^k - \underline{V}_{m,h+1}^k\right)(s_{h+1}^k) + 2H^2 \sum_{s \in \mathcal{S}} \sum_{i=1}^{\overline{n}_{m,h}^K(s)} \frac{\sqrt[4]{4\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)} + 1}{i} \\
\leq& (1+\frac{1}{H}) \sum_{k=1}^{K} \left(\overline{V}_{m,h+1}^k - \underline{V}_{m,h+1}^k\right)(s_{h+1}^k) + 2H^2 \sum_{s \in \mathcal{S}} \left(\sqrt[4]{4\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)} + 1\right) \sum_{i=1}^{\overline{n}_{m,h}^K(s)} \frac{1}{i} \\
\leq& (1+\frac{1}{H}) \sum_{k=1}^{K} \left(\overline{V}_{m,h+1}^k - \underline{V}_{m,h+1}^k\right)(s_{h+1}^k) + 4H^2 \sum_{s \in \mathcal{S}} \left(\sqrt[4]{4\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)} + 1\right)\iota \\
\leq& (1+\frac{1}{H}) \sum_{k=1}^{K} \left(\overline{V}_{m,h+1}^k - \underline{V}_{m,h+1}^k\right)(s_{h+1}^k) + 8H^2 \sum_{s \in \mathcal{S}} \sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)\iota} + 4H^2 S\iota.
\end{aligned} \tag{133}$$

Here the second line follows directly from proof of Lemma 20, the third line follows the definition of $d'_{max}(s_h^k)$. ■

Proof [Proof of Lemma 21] Consider any fixed pair of $(m,h) \in [M] \times [H]$. We inherit the definition of $X_n(s)$ from the proof of Lemma 10 to denote all episodes where (h,s) is visited and the n-th visit of (h,s) is received or been skipped. We have:

$$
\begin{aligned}
&\sum_{k=1}^{K}\Big(\overline{\beta}_{m,h}^{\underline{n}_k,\overline{n}_k}(s_h^k)+\underline{\beta}_{m,h}^{\underline{n}_k,\overline{n}_k}(s_h^k)+4CH^2\frac{\sqrt{\mathcal{T}_{m,h}^{\overline{n}_k}(s_h^k)}}{\underline{n}_k}\Big)\cdot\mathbb{I}\Big\{\underline{n}_k\geq d'_{max}(s_h^k)\Big\} \\
\leq&32H^2C\sum_{k=1}^{K}\sqrt{\frac{\mathcal{T}_{m,h}^{\overline{n}_k}(s_h^k)}{\underline{n}_k^2}\iota}\cdot\mathbb{I}\Big\{\underline{n}_k\geq d'_{max}(s_h^k)\Big\}+24H^2\sum_{k=1}^{K}\sqrt{\frac{A}{\underline{n}_k}\iota}\cdot\mathbb{I}\Big\{\underline{n}_k\geq d'_{max}(s_h^k)\Big\}
\end{aligned}
\tag{134}
$$

For the first term, we have:

$$
\begin{aligned}
&H^2C\sum_{k=1}^{K}\sqrt{\frac{\mathcal{T}_{m,h}^{\overline{n}_k}(s_h^k)}{\underline{n}_k^2}\iota}\cdot\mathbb{I}\Big\{\underline{n}_k\geq d'_{max}(s_h^k)\Big\} \\
\leq&H^2C\iota\sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)}\cdot\Big[\sum_{x\in X_{d'_{max}(s)}(s)}\frac{1}{\underline{n}_x}\Big] \\
\leq&3H^2C\sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)}\iota^2
\end{aligned}
\tag{135}
$$

Here the last line is from results in Equation (82).

Then for the second,

$$
\begin{aligned}
&H^2\sum_{k=1}^{K}\sqrt{\frac{A}{\underline{n}_k}\iota}\cdot\mathbb{I}\Big\{\underline{n}_k\geq d'_{max}(s_h^k)\Big\} \\
\leq&H^2\sum_{s\in\mathcal{S}}\sqrt{A\iota}\sum_{x\in X_{d'_{max}(s)}(s)}\sqrt{\frac{1}{\underline{n}_x}} \\
\leq&2H^2\sum_{s\in\mathcal{S}}\sqrt{A\overline{n}_{m,h}^K(s)\iota}+2H^2S\sqrt{A\iota} \\
\leq&4H^2\sqrt{SAK\iota}
\end{aligned}
\tag{136}
$$

where the third line follows Equation (85).

Substituting the above results into Equation (134), we have:

$$
\begin{aligned}
&\sum_{k=1}^{K}\Big(\overline{\beta}_{m,h}^{\underline{n}_k,\overline{n}_k}(s_h^k)+\underline{\beta}_{m,h}^{\underline{n}_k,\overline{n}_k}(s_h^k)+4CH^2\frac{\sqrt{\mathcal{T}_{m,h}^{\overline{n}_k}(s_h^k)}}{\underline{n}_k}\Big)\cdot\mathbb{I}\Big\{\underline{n}_k\geq d'_{max}(s_h^k)\Big\} \\
\leq&96CH^2\sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)}\iota^2+96H^2\sqrt{SAK\iota}
\end{aligned}
\tag{137}
$$

■

F.4. Step Four: Proof of Theorem 2

From Lemma 15, we can upper bound $\max_{m,h}\sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)}$ as follows:

$$
\begin{aligned}
\max_{m,h}\sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)} \leq & 2\max_{m,h}\sum_{s\in\mathcal{S}}\min_{\mathcal{L}}\left\{|\mathcal{L}|+\sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s),\mathcal{L}}(s)}\right\}+64C^2S \\
\leq & 2\max_{m,h}\min_{\mathcal{L}}\left\{S|\mathcal{L}|+\sum_{s\in\mathcal{S}}\sqrt{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s),\mathcal{L}}(s)}\right\}+64C^2S \\
= & 2\max_{m,h}\min_{\mathcal{L}}\left\{S|\mathcal{L}|+\sqrt{S\mathcal{T}_{m,h}^{K,\mathcal{L}}}\right\}+64C^2S,
\end{aligned} \tag{138}
$$

where $\mathcal{L}\subset[K]$. Then from Lemma 14, we have:

$$
\begin{aligned}
&\max_m\left(V_{m,1}^{\dagger,\pi_{-m}}-V_{m,1}^{\pi}\right)(s_1)=\max_m\frac{1}{K}\sum_{k=1}^{K}\left(V_{m,1}^{\dagger,\pi_{-m,k}}-V_{m,1}^{\pi_k}\right)(s_1) \\
\lesssim & CH^3\max_{m,h}\sum_{s\in\mathcal{S}}\sqrt{\frac{\mathcal{T}_{m,h}^{\overline{n}_{m,h}^K(s)}(s)}{K^2}}\iota^2+H^3\sqrt{\frac{SA}{K}\iota} \\
\lesssim & CH^3\max_{m,h}\min_{\mathcal{L}}\left\{S|\mathcal{L}|+\sqrt{S\mathcal{T}_{m,h}^{K,\mathcal{L}}}\right\}\iota^2+H^3\sqrt{\frac{SA}{K}\iota}
\end{aligned} \tag{139}
$$

F.5. Supporting Details

Proof [Proof of Lemma 15] Consider any fixed pair (m,h,s). When $\mathcal{T}_n\leq(8C)^4$, the desired result trivially holds. We now consider $\mathcal{T}_n\geq(8C)^4$. Then for any set $\mathcal{L}\subset[K]$, we have:

$$
\mathcal{T}_n=\sum_{i=1}^{n}i-e_i=\sum_{i=1}^{n}(i-e_i)\cdot\mathbb{I}\Big\{e_i\notin\mathcal{L}\Big\}+\sum_{i=1}^{n}(i-e_i)\cdot\mathbb{I}\Big\{e_i\in\mathcal{L}\Big\} \tag{140}
$$

For the first term, we have:

$$
\sum_{i=1}^{n}(i-e_i)\cdot\mathbb{I}\Big\{e_i\notin\mathcal{L}\Big\}\leq\sum_{i=1}^{n}(i-\arg\min_{j\notin\mathcal{L}}\left[d_{m,h}^j(s)+k_j>k_i-1\right])=\mathcal{T}_{n,\mathcal{L}} \tag{141}
$$

To see why this holds, consider the following cases. If $e_i\notin\mathcal{L}$, then:

$$
\begin{aligned}
&(i-e_i)\cdot\mathbb{I}\Big\{e_i\notin\mathcal{L}\Big\}=i-e_i \\
=&i-\arg\min_{j}\left[d_{m,h}^j(s)+k_j>k_i-1\right] \\
=&i-\arg\min_{j\notin\mathcal{L}}\left[d_{m,h}^j(s)+k_j>k_i-1\right]
\end{aligned} \tag{142}
$$

However, if $e_i \in \mathcal{L}$, then

$$(i - e_i) \cdot \mathbb{I}\Big\{e_i \notin \mathcal{L}\Big\} = 0 \leq i - \arg\min_{j \notin \mathcal{L}} \Big[d^j_{m,h}(s) + k_j > k_i - 1\Big] \tag{143}$$

Combining the two cases gives the desired result.

Now for the second term,

$$\begin{aligned}
&\sum_{i=1}^{n}(i - e_i) \cdot \mathbb{I}\Big\{e_i \in \mathcal{L}\Big\} \\
=& \sum_{i=1}^{n}\sum_{j \in \mathcal{L}}(i - j) \cdot \mathbb{I}\Big\{j = e_i\Big\} \\
=& \sum_{j \in \mathcal{L}}\sum_{i=j}^{n}(i - j) \cdot \mathbb{I}\Big\{j = e_i\Big\} \\
\leq& \sum_{j \in \mathcal{L}}\sum_{i=j}^{n}(i - j) \cdot \mathbb{I}\Big\{j \in \arg \mathcal{M}_{k_i}\Big\} \\
=& \sum_{j \in \mathcal{L}} \phi_{j,n}.
\end{aligned} \tag{144}$$

Here the last line follows Equation (23). For any visit $j \in \mathcal{O}_{\overline{n}_K}$, from Lemma 16, we have:

$$\phi_{j,n} \leq \sqrt{\mathcal{T}_n} + \sqrt[4]{4\mathcal{T}_n} + 1 \tag{145}$$

For any visit $j \notin \mathcal{O}_{\overline{n}_K}$, we have $\phi_{j,n} \leq \sqrt{\mathcal{T}_n}$. Combining the above two cases, we have:

$$\begin{aligned}
&\sum_{i=1}^{n}(i - e_i) \cdot \mathbb{I}\Big\{e_i \in \mathcal{L}\Big\} \\
\leq& |\mathcal{L}|\sqrt{\mathcal{T}_n} + |\mathcal{O}_K|\sqrt[4]{4\mathcal{T}_n} + |\mathcal{O}_K| \\
\leq& |\mathcal{L}|\sqrt{\mathcal{T}_n} + 4C(\mathcal{T}_n)^{3/4},
\end{aligned} \tag{146}$$

where the last line is due to Lemma 17.

Finally, combining the first and second term gives:

$$\mathcal{T}_n \leq \mathcal{T}_{n,\mathcal{L}} + |\mathcal{L}|\sqrt{\mathcal{T}_n} + 4C(\mathcal{T}_n)^{3/4} \tag{147}$$

Since $\mathcal{T}_n \geq (8C)^4$, we have $\mathcal{T}_n - 4C(\mathcal{T}_n)^{3/4} \geq \frac{1}{2}\mathcal{T}_n$. This implies:

$$\mathcal{T}_{n,\mathcal{L}} \geq \frac{1}{2}\mathcal{T}_n - |\mathcal{L}|\sqrt{\mathcal{T}_n} \tag{148}$$

which is equivalent to:

$$\min_{\mathcal{L} \in [n]} \left\{|\mathcal{L}| + \sqrt{\mathcal{T}_{n,\mathcal{L}}}\right\} \geq \min_{l} \left\{l + \sqrt{\frac{1}{2}\mathcal{T}_n - l\sqrt{\mathcal{T}_n}}\right\} \tag{149}$$

Since the right hand side is concave in l, its maximum is reached when $l = \frac{1}{2}\sqrt{\mathcal{T}_n}$. So we have:

$$\frac{1}{2}\sqrt{\mathcal{T}_n} \leq \min_{\mathcal{L}\in[n]} \left\{ |\mathcal{L}| + \sqrt{\mathcal{T}_{n,\mathcal{L}}} \right\} \tag{150}$$

■