Feasible Q-Learning for Average Reward Reinforcement Learning

Ying Jin
Stanford University

Jose Blanchet
Stanford University

Ramki Gummadi
Google

Zhengyuan Zhou[2]
New York University

Abstract

Average reward reinforcement learning (RL) provides a suitable framework for capturing the objective (i.e. long-run average reward) for continuing tasks, where there is often no natural way to identify a discount factor. However, existing average reward RL algorithms with sample complexity guarantees are not feasible, as they take as input the (unknown) mixing time of the Markov decision process (MDP). In this paper, we make initial progress towards addressing this open problem. We design a feasible average-reward Q-learning framework that requires no knowledge of any problem parameter as input. Our framework is based on discounted Q-learning, while we dynamically adapt the discount factor (and hence the effective horizon) to progressively approximate the average reward. In the synchronous setting, we solve three tasks: (i) learn a policy that is ϵ-close to optimal, (ii) estimate optimal average reward with ϵ-accuracy, and (iii) estimate the bias function (similar to Q-function in discounted case) with ϵ-accuracy. We show that with carefully designed adaptation schemes, (i) can be achieved with $\widetilde{O}(\frac{SAt_{\text{mix}}^8}{\epsilon^8})$ samples, (ii) with $\widetilde{O}(\frac{SAt_{\text{mix}}^5}{\epsilon^5})$ samples, and (iii) with $\widetilde{O}(\frac{SAB}{\epsilon^9})$ samples, where t_{mix} is the mixing time, and $B > 0$ is an MDP-dependent constant. To our knowledge, we provide the first finite-sample guarantees that are polynomial in $S, A, t_{\text{mix}}, \epsilon$ for a feasible variant of Q-learning. That said, the sample complexity bounds have tremendous room for improvement, which we leave for the community's best minds. Preliminary simulations verify that our framework is effective without prior knowledge of parameters as input.

1 Introduction

Reinforcement learning (RL) has achieved remarkable success in simulated environments such as beating world human champions in playing poker, chess and Go (Brown and Sandholm, 2018; Schrittwieser et al., 2020; Silver et al., 2018; Mnih et al., 2015; Schaeffer et al., 1992; Campbell et al., 2002). Such empirical successes–and the excitements generated therefrom–have motivated a remarkably fruitful line of research on the sample complexity of various RL algorithms, which characterizes how many samples (from a generative model such as a simulator) are needed to obtain an ϵ-optimal policy.

Regarding sample complexity, an important setting is the **discounted infinite-horizon** RL, where the goal is to maximize the (infinite) sum of all future γ-discounted rewards for a given factor $0 < \gamma < 1$. Broadly speaking, there are (at least) two main classes of methods to tackle the problem: *model-based* methods and *model-free* methods. Model-free algorithms learn to select actions without model estimation. Compared with model-based ones (Azar et al., 2013; Sidford et al., 2018b,a; Wang, 2020; Agarwal et al., 2020), they are often more computationally efficient, have less storage overhead, and are easy to generalize to RL with function approximation (Sutton and Barto, 2018; François-Lavet et al., 2018). In particular, Q-learning (Watkins and Dayan, 1992), as the prototypical model-free algorithm, has been studied[3] extensively in discounted infinite-horizon RL (Kearns and Singh, 1999; Even-Dar et al., 2003; Beck and Srikant, 2012; Wainwright, 2019a; Chen et al., 2020; Li et al., 2021).

[2]This work is generously supported by National Science Foundation grants CCF-2312205 and CCF-2312204. Correspond to ying531@stanford.edu and zz26@nyu.edu.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

[3]Earlier works on discounted Q-learning focused on characterizing its asymptotic convergence without any finite-sample guarantees; see Jaakkola et al. (1994); Tsitsiklis (1994); Szepesvári (1998); Borkar and Meyn (2000).

Such remarkable research efforts notwithstanding, **average-reward infinite-horizon** RL remains far from being resolved: As recognized by the RL community (Sutton and Barto, 2018; Wan et al., 2021; Mahadevan, 1996; Dewanto et al., 2020), average reward RL, which provides a more natural objective for many continuing tasks, is a much more challenging problem and has remained largely under-explored. The sparsity of theoretical understanding also limits the implementation. Indeed, discount factors are often used in practical instantiations of RL algorithms, even when the final objective of interest is clearly the average undiscounted reward in the long term. This is partly due to the fact that algorithms to directly optimize average reward are much more challenging to characterize than the discounted ones. As a result, popular RL algorithms make an ad-hoc choice of a fixed discount factor that are only partially understood (Tang et al., 2021). However, many RL applications do not have a natural discount factor that can be endogenously identified.

Classical results (Blackwell, 1962; Mahadevan, 1996) connect the discounted and average reward formulations via fundamental relations between optimal value functions *asymptotically*, yet there is only limited finite-sample guarantees related to algorithmic transfer between the two frameworks, given the ubiquitous practice of implementing the former to solve the latter. Recent works leverage similar relations to devise algorithms that optimizes the γ-discounted reward to approximate optimal average reward for a carefully designed, fixed γ (Jin and Sidford, 2020, 2021). However, these (model-based) algorithms that have finite-sample guarantees all require knowing the mixing time t_{mix} and are hence *infeasible*. In parallel, finite-sample guarantees for model-free methods (sometimes called R-learning, see Section 1.1) for average reward are even more scarce; the only one we know of is very recent (Zhang et al., 2021), and the algorithm therein also relies on unknown parameters as input. As such, and in light of the merits of **model-free** RL algorithms mentioned before, we are naturally led to the open question:

> *Can we design a feasible model-free average-reward RL algorithm with finite-sample guarantees for the sample complexity?*

In this paper, we provide new theoretical results that highlight an important algorithmic role for the discounted formulation as a subroutine in solving several learning tasks in the average reward scenario. In contrast to the typical use of a fixed discount factor, we highlight the importance of a carefully selected schedule for progressively finer discount factors to obtain a solution to the average reward objective. Our high-level idea is in concordance with Hordijk and Tijms (1975) which provides an asymptotic analysis for such paradigm when planning with knowledge of the MDP; in contrast, we provide finite-sample analysis for a learning problem, which requires judiciously chosen schemes of both discount factors and learning rates to tackle random data.

1.1 Related Work

We include more detailed discussion of related work in this part, extending Section 1.2 in the main text.

Model-based average-reward RL. (Kearns and Singh, 2002) is an early work on average-reward RL that proposes a model-based learning algorithm and establishes a sample complexity bound of $\widetilde{O}(\frac{S^5\text{poly}(A)t_{mix}^5}{\epsilon^6})$, where poly$(A)$ is an unspecified polynomial of A. From an algorithmic standpoint, an important issue of the algorithm is that it requires the knowledge of both t_{mix} and the optimal average reward[4]. More recently, several different model-based average reward RL algorithms have been proposed Wang (2017); Zhang and Xie (2023); Jin and Sidford (2020, 2021), with the latter two achieving state-of-the-art sample-complexity bounds of $\widetilde{O}(\frac{SAt_{mix}^2}{\epsilon^2})$ and $\widetilde{O}(\frac{SAt_{mix}}{\epsilon^3})$, respectively. All three algorithms rely on knowing t_{mix} or similar quantities. Jin and Sidford (2021) further provides a lower bound of $\widetilde{O}(\frac{SAt_{mix}}{\epsilon^2})$ for average-reward RL when t_{mix} is known. Consequently, while having the pleasing optimal dependence on SA, a minimax optimal algorithm for average-reward RL is not yet known even under known t_{mix}, an impractical assumption to begin with.

Model-free average-reward RL. More recently, driven by the merits of model-free algorithms, Zhang et al. (2021) provides the very first finite-sample analysis of a average reward Q-learning variant, yielding a sample complexity bound of $\widetilde{O}(\frac{SAJ^3}{(1-\delta)^5\epsilon^2})$, where J and δ are two unknown MDP-dependent constants that may arbitrarily depend on t_{mix}. This pioneering bound itself is valuable in light of the difficulty in characterizing finite-sample guarantees of Q-learning algorithms for average reward RL. However, the issue is that Zhang et al. (2021) assumes J and δ are known and given, rendering the algorithm infeasible.

Feasible average-reward RL with asymptotic guarantees. The only feasible average-reward Q-learning (sometimes also called R-learning) variants

[4]How to learn this quantity efficiently also remains under-explored. We provide an answer on this as well.

that we know of are Wan et al. (2021); Abounadi et al. (2001), which attempt to directly learn the Q-values (called bias function in average-reward RL) associated with the Bellman equation for average reward MDP (discussed in more detail in Section 2). However, for both algorithms, only asymptotic consistency is established. It is largely unclear whether finite-sample sample complexity bounds can be established for R-learning algorithms, since the underlying update is not a contraction. Remotely related is Hordijk and Tijms (1975), which shares a similar spirit as our methods for gradually adjusting discount factors to approximate average reward; however, it requires the exact knowledge of the MDP and hence solves a planning – rather than learning – problem (see the next subsection for more details), and only provide asymptotic analysis.

Other metrics for average-reward RL. Finally, we mention in passing that the literature has also studied other metrics such as regret Jaksch et al. (2010); Jin et al. (2018); Dong et al. (2019); Fruit et al. (2020); Dong et al. (2021); Wei et al. (2020), which is not the focus here. While there are online-to-batch tricks to turn regrets to sample complexity (e.g., via a randomly sample from history of policies), they can not tell *which policy* is ϵ-close to optimal; this deviates from our goal of designing a practically feasible algorithm and analyzing its sample complexity. Instead, we focus on algorithms with a deterministic output given data.

1.2 Our Contributions and Related Work

First, we design a feasible average-reward Q-learning algorithmic framework that requires no knowledge of any problem parameter. In contrast to the aforementioned Q-learning variants for average-reward RL Wan et al. (2021); Abounadi et al. (2001); Zhang et al. (2021) that all aim to directly learn the bias function (which determines the optimal policy; see Section 2) in the average-reward Bellman equation, our algorithmic framework uses discounted Q-learning but *dynamically* adjusts the discount factor towards 1 – and hence gradually enlarging the effective horizon – to yield progressively finer approximations of the average-reward setting. This idea is quite simple and intuitive; however, the challenge lies in designing the specific horizon adaptation scheme (involving the simultaneous adjustment of learning rate and discount factor; see Algorithm 1) such that the finite-sample analysis goes through. Note that unlike in the discounted setting with a fixed γ, we now have a "moving target" problem as the discount factor is constantly shifting and weaker contraction, thereby making the analysis much more difficult. On this note, an early work back to 1970s Hordijk and Tijms (1975) shares a similar idea of adjusting discount factors; however, it only provides asymptotic guarantees (and does so without any specific adaptation scheme); further, it requires the exact knowledge of the MDP and hence solves a planning problem.

Second, we offer two concrete instantiations of our framework under (distinct) judiciously chosen adaptation schemes. Using the first scheme, stopping our algorithm at any iteration T yields a policy whose average reward differs from the optimal by at most $\widetilde{O}(t_{\text{mix}}/T^{1/8})$. Under the synchronous setting we consider (see Section 2), it translates to a sample complexity bound of $\widetilde{O}(SAt_{mix}^8/\epsilon^8)$. Furthermore, if the goal is to simply *estimate* the optimal average reward, then we can do so at a faster rate by our second scheme: stopping at any iteration T yields an estimation accuracy of $\widetilde{O}(t_{\text{mix}}/T^{1/5})$, which translates to a sample complexity of $\widetilde{O}(SAt_{\text{mix}}^5/\epsilon^5)$. These results also indicate that our algorithm is any time: we do not need to know T beforehand when running the algorithm[5]. To the best of our knowledge, these are the first finite-sample guarantees for a feasible Q-learning algorithm. That said, we believe these bounds can be improved; and we view our results as an open invitation for further progress in feasible average-reward RL.

Additionally, although not the main focus here, we also extend our algorithmic framework to estimate the bias function in the average-reward MDP Bellman equation (Section 5). As mentioned before, existing average-reward Q-learning algorithms Wan et al. (2021); Abounadi et al. (2001); Zhang and Ji (2019); Zhang et al. (2021) learn a policy through estimating the bias function. Viewed through this lens, we provide an alternative way for estimating this quantity that has a sample complexity of $\widetilde{O}(\frac{SAB}{\epsilon^9})$. Similar to Zhang et al. (2021) (which is the only work that has a finite-sample guarantee for estimating the bias function), our bound has an unpleasing dependence on an unknown MDP-dependent constant B. However, different from Zhang et al. (2021), our algorithm itself is feasible (and does not need to know B or any other problem parameter). Note that the two bounds – ours and the one in Zhang et al. (2021) – are incomparable as all of those problem-dependent constants are unknown.

Finally, through preliminary simulations, we verify that our algorithm learns the near-optimal policy with satisfactory convergence, and performs well across MDPs with various values of mixing time without the need of any prior knowledge of the mixing time.

[5]Algorithms without this property cannot *provably* adapt to additional samples beyond an initially chosen T.

2 Problem Setup

We consider an infinite-horizon tabular MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$, with finite state space $\mathcal{S} = \{1, \dots, S\}$, finite action space $\mathcal{A} = \{1, \dots, A\}$, transition probability $P\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ (i.e., $P(s' \,|\, s, a)$ is the probability of transiting to s' from a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$), and reward function $r\colon \mathcal{S} \times \mathcal{A} \to [0, 1]$ (i.e., $r(s, a)$ is the immediate reward at state $s \in \mathcal{S}$ if action $a \in \mathcal{A}$ is taken). We let $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ denote a policy, i.e., $\pi(a \,|\, s)$ is the probability of taking a at state s. When π is a deterministic policy, $\pi(s)$ denotes the action chosen at state s.

Learning objective. Given a policy π, we define its long-term average reward as

$$V^\pi(s) = \liminf_{T\to\infty} \mathbb{E}_\pi \left[\frac{1}{T} \sum_{k=1}^{T} r(s_k, a_k) \,\middle|\, s_1 = s \right]$$

for all $s \in \mathcal{S}$. Here $\mathbb{E}_\pi$ denotes the expectation over the trajectory $\{(s_k, a_k)\}_{k\geq 0}$ of the MDP under policy π. Under sufficient generality[7], the standard MDP theory (Puterman, 2014) shows that for any policy π, there exists a constant $J^\pi \in [0, 1]$ such that $V^\pi(s) = J^\pi$ for all $s \in \mathcal{S}$. The long-term average reward of π initialized at any state-action pair also equals J^π, that is, $J^\pi \equiv \liminf_{T\to\infty} \mathbb{E}_\pi \big[\frac{1}{T} \sum_{k=1}^{T} r(s_k, a_k) \,\big|\, s_1 = s, a_1 = a \big]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. We use $J^\pi \in [0, 1]$ to denote the average reward of policy π.

The optimal policy π^* for average reward attains $J^{\pi^*} = \max_\pi J^\pi$, and we denote $J^* = J^{\pi^*}$. Furthermore, there exists a function $q^*\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ so that the following Bellman equation holds for all $(s, a) \in \mathcal{S}\times\mathcal{A}$:

$$J^* + q^*(s, a) = r(s, a) + \mathbb{E}_{s'\sim P(\cdot \,|\, s,a)} \big[v^*(s') \big], \qquad (1)$$

where $v^*(s') = \max_{a'\in\mathcal{A}} q^*(s', a')$ for all $s' \in \mathcal{S}$. The optimal policy π^* is greedy w.r.t. q^*, i.e., $\pi^*(s) = \operatorname{argmax}_{a\in\mathcal{A}} q^*(s, a)$ for all $s \in \mathcal{S}$. The solution (q^*, v^*) to (1) is unique up to a constant; one solution is

$$v^*(s) = \mathbb{E}_{\pi^*} \left[\sum_{k=1}^{\infty} \big(r(s_k, a_k) - J^* \big) \,\middle|\, s_1 = s \right], \qquad (2)$$

$$q^*(s, a) = \mathbb{E}_{\pi^*} \left[\sum_{k=1}^{\infty} \big(r(s_k, a_k) - J^* \big) \,\middle|\, s_1 = s, a_1 = a \right],$$

where v^* and q^* are called the *value bias function and q-value bias function* respectively (Puterman, 2014).

In this work, we will consider three tasks: (1) learning the optimal reward J^* up to ϵ-accuracy, (2) learning a policy whose average reward is ϵ-close to J^*, and (3) learning the bias functions q^* and v^* up to constants.

[7] Puterman (2014) shows that if the state space $\mathcal{S}$ is finite or countable, then the limit (instead of liminf) $V^\pi(s)$ exists; if the chain induced by π is irreducible or has a single recurrent class, then $V^\pi(s)$ is a constant function. We operate in this scenario.

2.1 Sampling Scheme: Synchronous Setting with a Generative Model

Throughout this paper, we work under the synchronous scenario with a generative model (or simulator) (Even-Dar et al., 2003). That is, we consider algorithms that proceed in multiple iterations. In each iteration t, we receive an independent sample $s' \sim P(\cdot \,|\, s, a)$ for *all* state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$. Despite being relatively simple, it serves as an idealistic sampling protocol that has received much attention for various RL algorithms throughout the RL literature (Kearns et al., 2002; Kakade, 2003; Even-Dar et al., 2003; Azar et al., 2013; Beck and Srikant, 2012; Sidford et al., 2018a,b; Wainwright, 2019a,b; Yang and Wang, 2019; Zanette et al., 2019; Agarwal et al., 2020; Li et al., 2021). We focus on this setting as a starting point for studying sample complexity of feasible average-reward RL algorithms.

2.2 Mixing time and other related notions

Definition 2.1. The mixing time t_{mix} of an MDP is

$$\max_\pi \min \Big\{ t\colon \max_{q\in\Delta(\mathcal{S})} d_{\text{TV}}\big((P^\pi)^t(q), \nu^\pi \big) \leq 1/4 \Big\}, \quad (3)$$

where $d_{\text{TV}}(\mu, \nu) := \frac{1}{2} \sum_{s\in\mathcal{S}} |\mu(s) - \nu(s)|$ is the total variation distance, and $(P^\pi)^t(q)$ is the distribution of s_t induced by policy π with initial distribution $s_0 \sim q$. We suppose for any policy π there exists a stationary distribution $\nu^\pi \in \Delta(\mathcal{S})$; otherwise $t_{\text{mix}} = \infty$.

The notion of mixing time (3) is widely adopted in the literature of average reward MDPs (Wei et al., 2020; Jin and Sidford, 2021), and our assumption is standard inthe liteature Wang (2017); Jin and Sidford (2020). While we adopt such notion to be consistent with the literature, our analysis framework also works under other regularity conditions.

Remark 2.2. Our results remain true if the mixing time (3) is replaced by the reward averaging time similar to De Farias and Van Roy (2006); Dong et al. (2021); the latter is also closely related to the notion of averaging time considered in Kearns and Singh (2002). To be specific, the reward averaging time is defined as $\tau = \max_\pi \sup_{T\geq 1} \big\{ T \cdot |J^\pi - V^\pi(s, T)| \big\}$, where $V^\pi(s, T) = \mathbb{E}_\pi \big[\frac{1}{T} \sum_{t=1}^{T} r(s_t, a_t) \,\big|\, s_1 = s \big]$ is the T-step average reward value function of π. Our results carry over to this setting with t_{mix} replaced by $\tau < \infty$ in the upper bound.

Remark 2.3. Our framework also applies to *weakly communicating* MDPs (Wei et al., 2020). Define the

span of the optimal γ-discounted value function as $\mathrm{sp}(\gamma) = \big(\max_s V_\gamma^*(s) - \min_s V_\gamma^*(s)\big)/(1-\gamma)$, where V_γ^* is the optimal discounted value function with a scaling; the exact definition of V_γ^* is deferred to (4). It is argued in Wei et al. (2020) that the span is bounded by the diameter of the MDP (Lattimore and Szepesvári, 2020) for weakly-communicating MDPs. Our bound applies to MDPs with bounded spans with t_{mix} replaced by $\sup_{\gamma\in(0,1)} \mathrm{sp}(\gamma)$.

3 Algorithm: Dynamic Horizon Q-Learning

3.1 Recap: Discounted Q-Learning

We start with a brief recap on the celebrated Q-learning algorithm for discounted value functions. Given any discount factor $\gamma \in (0,1)$ and any policy π, with rewards $r_k = r(s_k, a_k)$, we denote the *rescaled* γ-discounted value function and Q-function of π as

$$V_\gamma^\pi(s) = (1-\gamma)\mathbb{E}_\pi\Big[\sum_{k=1}^\infty \gamma^{t-1} r_k \,\Big|\, s_1 = s\Big],$$
$$Q_\gamma^\pi(s,a) = (1-\gamma)\mathbb{E}_\pi\Big[\sum_{k=1}^\infty \gamma^{k-1} r_k \,\Big|\, s_1 = s, a_1 = a\Big]$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. The optimal discounted value function V_γ^* and Q-function Q_γ^* are

$$V_\gamma^*(s) = \max_\pi V_\gamma^\pi(s), \quad Q_\gamma^*(s,a) = \max_\pi Q_\gamma^\pi(s,a) \quad (4)$$

for $(s,a) \in \mathcal{S}\times\mathcal{A}$. For preparation, we also denote the *unscaled* γ-discounted value and Q-functions of π as

$$v_\gamma^\pi = V_\gamma^\pi/(1-\gamma), \quad q_\gamma^\pi = Q_\gamma^\pi/(1-\gamma),$$

and the optimal ones as

$$v_\gamma^* = V_\gamma^*/(1-\gamma), \quad q_\gamma^* = Q_\gamma^*/(1-\gamma). \quad (5)$$

The optimal policy for γ-discounted reward is denoted as π_γ^*. It is well-known that π_γ^* is a deterministic policy with $\pi_\gamma^*(s) = \mathrm{argmax}_{a\in\mathcal{A}} Q_\gamma^*(s,a) = \mathrm{argmax}_{a\in\mathcal{A}} q_\gamma^*(s,a)$.

In synchronous setting, the Q-learning algorithm for γ-discounted rewards maintains an estimate $Q_t\colon \mathcal{S}\times\mathcal{A}\to[0,1]$ for the optimal Q-function Q_γ^*. In each iteration t, it updates *all* entries of the estimate at once, according to $Q_t = (1-\eta_t)Q_{t-1} + \eta_t \mathcal{T}_t(Q_{t-1})$. Here $\eta_t \in (0,1]$ is the learning rate or the step size, and $\mathcal{T}_t$ is the empirical Bellman operator depending on the samples collected in the t-th iteration (with proper scaling): $\mathcal{T}_t(Q)(s,a) = (1-\gamma)r(s,a) + \gamma \max_{a'\in\mathcal{A}} Q(s',a')$, where $s' \sim P(\cdot\,|\,s,a)$ is the independent sample collected for (s,a) from the generative model.

3.2 Dynamic Horizon Q-Learning Framework

In the existing works of Wei et al. (2020); Jin and Sidford (2021), algorithms for the discounted setting are applied with a properly chosen discounted factor γ that depends on the *known* mixing time and a prespecified sample size. Rather distinct from them, we avoid the knowledge of the mixing time and a prespecified sample size by employing a series of dynamic discount factors.

To be specific, given a sequence of discount factors $\{\gamma_t\}_{t\ge1}$, we maintain an estimate $Q_t\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ in the t-th iteration. In each iteration, the algorithm updates all entries of the Q-function estimate via

$$Q_t(s,a) = (1-\eta_t)Q_{t-1}(s,a) + \eta_t\big[(1-\gamma_t)r(s,a) + \gamma_t \max_{a'\in\mathcal{A}} Q_{t-1}\big(s_t(s,a),a'\big)\big] \quad (6)$$

for all state-action pairs $(s,a)\in\mathcal{S}\times\mathcal{A}$. Here $s_t(s,a)$ is the independent sample from the generative model, γ_t is the discount factor and $\eta_t \in (0,1]$ is the learning rate in the t-th iteration. Correspondingly, we define the estimate of value function in the t-th iteration as $V_t(s) := \max_{a\in\mathcal{A}} Q_t(s,a)$ for all $s\in\mathcal{S}$, and the associated greedy policy as

$$\pi_t(s) := \operatorname*{argmax}_{a\in\mathcal{A}} Q_t(s,a), \quad (7)$$

so that $V_t(s) = Q_t(s,\pi_t(s))$ for all $s\in\mathcal{S}$. The complete algorithm is summarized as follows.

Algorithm 1 Dynamic Horizon Q-Learning

$\{\eta_t\}_{t\ge1}$, $\{\gamma_t\}_{t\ge1}$.Initialization: $Q_0 \equiv 0$. **for** $t = 1, 2, \ldots$ **do**
Generate $s_t(s,a) \sim P(\cdot\,|\,s,a)$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.
Set $Q_t(s,a) = (1-\eta_t)Q_{t-1}(s,a) + \eta_t[(1-\gamma_t)r(s,a) + \gamma_t V_{t-1}(s_t(s,a))]$, $\forall (s,a)\in\mathcal{S}\times\mathcal{A}$.
Set $V_t(s) = \max_{a\in\mathcal{A}} Q_t(s,a)$ for all $s\in\mathcal{S}$.
end for

3.3 Theory for estimating optimal reward

We provide theoretical guarantee of Algorithm 1 for learning the optimal reward J^* as follows. The proof of Theorem 3.1 is sketched in Section 4.1 and 4.2, while detailed in Appendix A.1.

Theorem 3.1. *In Algorithm 1, we set* $\eta_t = \big(1 + \frac{c_1 t^{3/5}}{(\log t)^3}\big)^{-1}$, $\gamma_t = 1 - t^{-1/5}$, $t \ge 2$, *for some constant* $c_1 > 0$ *and set* $\eta_1 = \eta_2$, $\gamma_1 = \gamma_2$. *Let* $\varepsilon \in (0,1)$ *and* $\delta \in (0,1)$. *Suppose* T *is sufficiently large such that* $T/\log T \ge 300$, $T^{1/5}(\log T)^2 \ge 4c_2$, $(\log T)^2 \ge 12(10+c_1)$, $c_2 T^{1/5} \ge 24(10+c_1)(\log T)^{2/5}$

and $T^{2/5} \geq \frac{64(\log T)^4}{9c_1}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$ *for some constant* $c_2 > 0$. *. Then with probability at least* $1-\delta$, *after* T *iterations, Algorithm 1 achieves*

$$\big|V_T(s) - J^*\big| \leq \frac{c \cdot t_{mix}(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}}$$

simultaneously for all $s \in \mathcal{S}$, *where* $J^* \in [0,1]$ *is the optimal average reward defined in Section 2, and* $c > 0$ *is an absolute constant that only depends on* c_1, c_2.

We take a moment to discuss the general idea of our learning framework. At a high level, we approximate the ultimate targets (i.e., the optimal average reward J^*) with a sequence of proxies $\{Q^*_{\gamma_t}\}$ that our estimates $\{Q_t\}$ eventually converge to. In particular,

$$\big|Q_t(\cdot,\cdot) - J^*\big| \leq \big|Q^*_{\gamma_t}(\cdot,\cdot) - J^*\big| + \big|Q_t(\cdot,\cdot) - Q^*_{\gamma_t}(\cdot,\cdot)\big| \tag{8}$$

for all inputs in $\mathcal{S}\times\mathcal{A}$. The first term in (8) is the approximation error by discounted Q-functions that is controlled as follows (proof in Appendix A.2).

Lemma 3.2. $\big|V^*_\gamma(s) - J^*\big| \leq 3(1-\gamma)t_{mix}$ *and* $\big|Q^*_\gamma(s,a) - J^*\big| \leq 3(1-\gamma)t_{mix}$ *for all* $(s,a) \in \mathcal{S}\times\mathcal{A}$.

The second term of (8) is the estimation error of $\{Q_t\}$ to the dynamic targets $\{Q^*_{\gamma_t}\}$. We provide the sketch of a recursive analysis for this term in Section 4.1.

3.4 Theory for Learning ϵ-Optimal Policy

Algorithm 1 can be adapted to learn a policy whose average reward is ϵ-close to optimal. The idea is still to take the greedy policy from our estimate Q_t. However, due to the subtlety between Q-function approximation and the value of the greedy policy, we need another adaptation scheme of $\{\gamma_t\}$ and $\{\eta_t\}$. A partial proof sketch of Theorem 3.3 is in Section 4.1; the detailed proof is in Appendix B.

Theorem 3.3. *In Algorithm 1, we set* $\eta_t = \big(1 + \frac{c_1 t^{5/8}}{(\log t)^2}\big)^{-1}$, $\gamma_t = 1 - t^{-1/8}$, $t \geq 2$ *for some constant* $c_1 > 0$ *and set* $\eta_1 = \eta_2$, $\gamma_1 = \gamma_2$. *Let* π_t *be the greedy policy with respect to* Q_t *from Algorithm 1 for all* $t \geq 1$. *Let* $\varepsilon \in (0,1)$ *and* $\delta \in (0,1)$. *Suppose* T *is sufficiently large such that* $(\log T)^2 \geq 11(2+c_1)/2$, $T/\log T \geq 100$, $c_2 T^{1/4} \geq 4(2+c_1)(\log T)^{11/8}$ *and* $T^{5/8} \geq \frac{64(\log T)^3}{9c_1}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$ *for some constant* $c_2 > 0$. *Then with probability at least* $1-\delta$, *after* T *iterations,* π_T *in Algorithm 1 satisfies*

$$\big|J^{\pi_T} - J^*\big| \leq \frac{c \cdot t_{mix}(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/8}}.$$

4 Analysis Framework

We provide the analysis framework for Theorems 3.1 and 3.3. In Section 4.1, we provide a general decomposition of the estimation error; we then give the proof sketch for Theorem 3.1 in Section 4.2. Theorem 3.3 follows similar ideas, with details in Appendix B.

Notations. Bold letters denote vectors and matrices. For any matrix $\boldsymbol{M}$, we denote $\|\boldsymbol{M}\|_1 = \max_i \sum_j |M_{ij}|$ and $\|\boldsymbol{M}\|_\infty = \max_{i,j}|M_{ij}|$. For vectors $\boldsymbol{a} = [a_i], \boldsymbol{b} = [b_i] \in \mathbb{R}^n$, $\boldsymbol{a} \leq \boldsymbol{b}$ (resp. $\boldsymbol{a} \geq \boldsymbol{b}$) means $a_i \leq b_i$ (resp. $a_i \geq b_i$) for all i. We let $\boldsymbol{a}\circ\boldsymbol{b} = [a_i b_i] \in \mathbb{R}^n$. For vector $\boldsymbol{a} = [a_i]$, we denote $|\boldsymbol{a}| = [|a_i|]$. For a group of vectors $\{\boldsymbol{a}_i \colon i \in \mathcal{I}\}$, we denote $\max_{i\in\mathcal{I}} \boldsymbol{a}_i$ as the vector of entrywise maximum, that is, $[\max_{i\in\mathcal{I}} \boldsymbol{a}_i]_j = \max_{i\in\mathcal{I}}[\boldsymbol{a}_i]_j$, $\forall j$. We use vector $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent reward functions, so that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, the (s,a)-th entry of $\boldsymbol{r}$ is given by $r(s,a)$. We represent value and Q-functions in vectors: for example, the s-th entry of $\boldsymbol{V}^\pi \in \mathbb{R}^{|\mathcal{S}|}$ is given by $V^\pi(s)$; we define $\boldsymbol{Q}_t, \boldsymbol{Q}^\pi, \boldsymbol{Q}^\pi_\gamma, \boldsymbol{Q}^*_\gamma, \mathbf{q}_t, \mathbf{q}^\pi, \mathbf{q}^*, \mathbf{q}^\pi_\alpha, \mathbf{q}^*_\alpha \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and $\boldsymbol{V}_t, \boldsymbol{V}^\pi, \boldsymbol{V}^\pi_\gamma, \boldsymbol{V}^*_\gamma, \mathbf{v}_t, \mathbf{v}^\pi, \mathbf{v}^*, \mathbf{v}^\pi_\alpha, \mathbf{v}^*_\alpha \in \mathbb{R}^{|\mathcal{S}|}$, analogously. We use a matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ to represent the probability transition kernel P, whose (s,a)-row $\boldsymbol{P}_{s,a}$ represents the vector $P(\cdot\,|\,s,a)$. For any vector $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$, we define $Var_{\boldsymbol{P}}(\boldsymbol{V}) = \boldsymbol{P}(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}\boldsymbol{V})\circ(\boldsymbol{P}\boldsymbol{V}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, that is, the (s,a)-entry of $\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V})$ is $\mathrm{Var}(V(s'))$ for $s' \sim P(\cdot\,|\,s,a)$. We also define the square probability transition matrix $\boldsymbol{P}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ (resp. $\boldsymbol{P}_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$) induced by a deterministic policy π over the state-action pairs (resp. states) as $\boldsymbol{P}^\pi := \boldsymbol{P}\boldsymbol{\Pi}^\pi$, $\boldsymbol{P}_\pi := \boldsymbol{\Pi}^\pi\boldsymbol{P}$, where $\boldsymbol{\Pi}^\pi \in \{0,1\}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$ is the projection matrix associated with π, whose s-th row consists of $|\mathcal{S}|$ blocks each of length $|\mathcal{A}|$, among which the s-th block is $\boldsymbol{e}^\top_{\pi(s)}$, and $\boldsymbol{e}_i$ is the i-th standard basis vector. Given samples $s_t(s,a) \sim P(\cdot\,|\,s,a)$ collected in the t-th iteration, we define the empirical transition matrix $\boldsymbol{P}_t \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ by $\boldsymbol{P}_t\big((s,a),s'\big) = \mathbb{1}\{s' = s_t(s,a)\}$.

4.1 Framework of Analysis for Algorithm 1

In this section, we provide a sketch of analysis for Algorithm 1. To begin with, we denote $\boldsymbol{\Delta}_t = \boldsymbol{Q}_t - \boldsymbol{Q}^*_{\gamma_t}$, which is the estimation error of $\boldsymbol{Q}_t$ for the rescaled optimal Q-function with γ_t in the t-th iteration. Our updating rule (6) in the t-th iteration admits the representation:

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t\big[(1-\gamma_t)\boldsymbol{r} + \gamma_t\boldsymbol{P}_t\boldsymbol{V}_{t-1}\big].$$

Employing the Bellman equation $\boldsymbol{Q}^*_{\gamma_t} = (1-\gamma_t)\boldsymbol{r} + \gamma_t\boldsymbol{P}\boldsymbol{V}^*_{\gamma_t}$ (note the rescaling of the value and Q-functions compared to conventional notations),

$$\begin{aligned}
\boldsymbol{\Delta}_t &= \boldsymbol{Q}_t - \boldsymbol{Q}^*_{\gamma_t}\\
&= (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t\big[(1-\gamma_t)\boldsymbol{r} + \gamma_t\boldsymbol{P}_t\boldsymbol{V}_{t-1}\big] - \boldsymbol{Q}^*_{\gamma_t}\\
&= (1-\eta_t)\big[\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^*_{\gamma_t}\big]\\
&\quad + \eta_t\big[(1-\gamma_t)\boldsymbol{r} + \gamma_t\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{Q}^*_{\gamma_t}\big]
\end{aligned}$$

$$
\begin{aligned}
&= (1-\eta_t)\big[\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^*_{\gamma_t}\big] \\
&\quad + \eta_t\big[(1-\gamma_t)\boldsymbol{r} + \gamma_t \boldsymbol{P}_t \boldsymbol{V}_{t-1} - (1-\gamma_t)\boldsymbol{r} - \gamma_t \boldsymbol{P}\boldsymbol{V}^*_{\gamma_t}\big] \\
&= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + (1-\eta_t)\big[\boldsymbol{Q}^*_{\gamma_{t-1}} - \boldsymbol{Q}^*_{\gamma_t}\big] \\
&\quad + \eta_t\gamma_t\big[\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^*_{\gamma_t}\big]. \qquad (9)
\end{aligned}
$$

The decomposition (9) is similar to what appears in the standard analysis of discounted Q-learning updates, such as Li et al. (2021); Wainwright (2019b); however, we note a few key technical challenges due to a dynamic discount factor and a moving target $\boldsymbol{Q}^*_{\gamma_t}$, which require considerably more efforts and techniques to address.

(1) First, due to the dynamic discount factor $\gamma_t \to 1$, the contraction of (9) is much weaker than that with a fixed discount factor, so that the last term $\eta_t\gamma_t\big[\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^*_{\gamma_t}\big]$ is much more difficult to control. As a result, γ_t cannot converge to 1 too fast.

(2) Due to the moving target $\boldsymbol{Q}^*_{\gamma_t}$, there is a bias $(1-\eta_t)\big[\boldsymbol{Q}^*_{\gamma_{t-1}} - \boldsymbol{Q}^*_{\gamma_t}\big]$ in (9). It is large if γ_t converges to 1 too quickly or if η_t is too small. A similar bias occurs in $\eta_t\gamma_t\big[\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^*_{\gamma_t}\big]$. In light of (1) and (2) and the bias in Lemma 3.2 (large if γ_t converges to 1 too slowly), we need a careful choice of γ_t to balance different sources of bias.

(3) Similar to the discounted setting, the learning rate η_t needs to balance the bias from earlier updates and variance from recent updates. Moreover, as η_t is coupled with γ_t in the third term, we have to admit a less aggressive learning rate due to additionally balancing the bias&variance in random update with η_t and the bias from γ_t.

To summarize, in sharp distinction from the discounted Q-learning, we need to address the bias due to $\gamma_t \to 1$ in conjunction with the statistical error in random learning update. As a result, our dynamic Q-learning algorithm typically yield slower convergence compared with the discounted counterpart.

Our analysis partially builds on the recent progress in sharp analysis for discounted Q-learning Li et al. (2021). However, we additionally tackle the bias due to $\gamma_t \neq \gamma_{t-1}$, which leads to distinct analysis. The last term in the bracket in (9) can be further decomposed as

$$
\begin{aligned}
\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^*_{\gamma_t} =& (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1} + \boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*_{\gamma_{t-1}}) \\
&+ \boldsymbol{P}(\boldsymbol{V}^*_{\gamma_{t-1}} - \boldsymbol{V}^*_{\gamma_t}),
\end{aligned}
$$

where

$$
\begin{aligned}
&\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*_{\gamma_{t-1}}) = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^*_{\gamma_{t-1}}}\boldsymbol{Q}^*_{\gamma_{t-1}} \\
&\le \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}^*_{\gamma_{t-1}} = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1}, \\
&\boldsymbol{P}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*_{\gamma_{t-1}}) = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^*_{\gamma_{t-1}}}\boldsymbol{Q}^*_{\gamma_{t-1}} \\
&\ge \boldsymbol{P}^{\pi^*_{\gamma_{t-1}}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^*_{\gamma_{t-1}}}\boldsymbol{Q}^*_{\gamma_{t-1}} = \boldsymbol{P}^{\pi^*_{\gamma_{t-1}}}\boldsymbol{\Delta}_{t-1}.
\end{aligned}
$$

Above two inequalities use the fact that $P^*_{\gamma t-1}$ is greedy w.r.t. $\boldsymbol{Q}^*_{\gamma_{t-1}}$, while π_{t-1} is greedy w.r.t. $\boldsymbol{Q}_{t-1}$. Plugging them back into (9) leads to

$$
\begin{aligned}
\boldsymbol{\Delta}_t &\le (1-\eta_t)\boldsymbol{\Delta}_t \\
&\quad + \boldsymbol{d}_t + \eta_t\gamma_t\big[\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big]; \quad (11a) \\
\boldsymbol{\Delta}_t &\ge (1-\eta_t)\boldsymbol{\Delta}_t \\
&\quad + \boldsymbol{d}_t + \eta_t\gamma_t\big[\boldsymbol{P}^{\pi^*_{\gamma_{t-1}}}\boldsymbol{\Delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}\big], \quad (11b)
\end{aligned}
$$

where we define the switching error in the t-th iteration

$$
\boldsymbol{d}_t = (1-\eta_t)[\boldsymbol{Q}^*_{\gamma_{t-1}} - \boldsymbol{Q}^*_{\gamma_t}] + \eta_t\gamma_t\boldsymbol{P}(\boldsymbol{V}^*_{\gamma_{t-1}} - \boldsymbol{V}^*_{\gamma_t}).
$$

Applying (11a) and (11b) recursively, we arrive at

$$
\begin{aligned}
\boldsymbol{\Delta}_t &\le \sum_{i=1}^{t} \eta_i^{(t)}\gamma_i\big[(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1} + \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{i-1}\big] \\
&\qquad + \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\boldsymbol{d}_i/\eta_i, \\
\boldsymbol{\Delta}_t &\ge \sum_{i=1}^{t} \eta_i^{(t)}\gamma_i\big[(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1} + \boldsymbol{P}^{\pi^*_{\gamma_{i-1}}}\boldsymbol{\Delta}_{i-1}\big] \\
&\qquad + \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\boldsymbol{d}_i/\eta_i,
\end{aligned} \qquad (12)
$$

where we define $\eta_t^{(t)} = \eta_t$, $\eta_0^{(t)} = \prod_{j=1}^{t}(1-\eta_j)$ and $\eta_i^{(t)} = \eta_i \cdot \prod_{j=i+1}^{t}(1-\eta_j)$ for $i \ge 1$.

We now proceed to bound (12). Let $\beta \in (0,1)$ be a constant whose value will be specified later. Writing $t_\beta = \lfloor (1-\beta)t \rfloor$, the upper bound in (12) is

$$
\boldsymbol{\Delta}_t \le \boldsymbol{\zeta}_t + \boldsymbol{\xi}_t + \sum_{i=1+t_\beta}^{t}\eta_i^{(t)}\gamma_i\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} + \boldsymbol{\delta}_t, \quad (13)
$$

where

$$
\begin{aligned}
\boldsymbol{\zeta}_t &:= \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t_\beta}\eta_i^{(t)}\gamma_i\big[(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1} + \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{i-1}\big], \\
\boldsymbol{\xi}_t &:= \sum_{i=1+t_\beta}^{t}\eta_i^{(t)}\gamma_i(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}, \\
\boldsymbol{\delta}_t &:= \sum_{i=1}^{t}\eta_i^{(t)}\boldsymbol{d}_i/\eta_i.
\end{aligned}
$$

The convergence rates for these quantities then depend on the choice of $\{\eta_t\}$ and $\{\gamma_t\}$, which we analyze in a case-by-case fashion for all our theoretical results.

4.2 Sketch of Analysis for Theorem 3.1

As an example of our analysis, we now bound the terms $\boldsymbol{\delta}_t$, $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$ in the decomposition (13), which leads to

a recursive bound on $\boldsymbol{\Delta}_t$ for Theorem 3.1. To begin with, for the constant $c_2 > 0$ in Theorem 3.1, we set

$$\beta = \frac{c_2}{T^{1/5}(\log T)^2}.$$

The lemmas throughout this subsection will be under the same conditions as Theorem 3.1.

Bounding switching error. First, the dynamic discount factors leads to the switching error $\boldsymbol{\delta}_t$, since the estimation targets $\boldsymbol{Q}^*_{\gamma_t}$ keeps moving.

Lemma 4.1. *Let T satisfy the conditions of Theorem 3.1. Then $\|\boldsymbol{\delta}_t\|_\infty \le 2(\frac{\log T}{T})^{2/5}$ for all t obeying $T/\log T \le t \le T$.*

Bounding $\boldsymbol{\zeta}_t$. The second term $\boldsymbol{\zeta}_t$ is the cumulative estimation error up to $t_\beta = \lfloor (1-\beta)t \rfloor$. It can be bounded via appropriate contraction with appropriate choices of t_β. We bound its ℓ_∞-norm as follows. The proof of Lemma 4.2 is in Appendix A.5.

Lemma 4.2. *Let T satisfy the conditions of Theorem 3.1. Then $\|\boldsymbol{\zeta}_t\|_\infty \le \frac{2}{T}$ for all $T/\log T \le t \le T$.*

Bounding $\boldsymbol{\xi}_t$. Finally, $\boldsymbol{\xi}_t$ is a random error term, for which we derive a high-probability bound, adapting the sharp analysis strategies introduced in Li et al. (2021). The proof of Lemma 4.3 is in Appendix A.6.

Lemma 4.3. *For any fixed t obeying $T/\log T \le t \le T$, it holds with probability at least $1-\delta$ that*

$$|\boldsymbol{\xi}_t| \le 5\sqrt{\frac{(\log T)^{\frac{19}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{c_1 T^{3/5}}} \times \sqrt{\Big(\max_{\lfloor (1-\beta)t \rfloor < i \le t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-1/5}\mathbf{1} \Big)}. \tag{14}$$

A recursive bound. With the above three bounds in place, we put them together and obtain a recursive bound on $\boldsymbol{\Delta}_t$, whose detailed proof is in Appendix A.3.

Proposition 4.4. *Under the conditions in Theorem 3.1, with probability at least $1-\delta$,*

$$\boldsymbol{\Delta}_t \le \sqrt{\frac{c_3(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}(T\log T)^{1/5}}} \times \sqrt{\big(1 + 9t_{mix}(\log T)^{1/5} + T^{1/5} \max_{\lfloor t/2 \rfloor \le i < t} \|\boldsymbol{\Delta}_i\|_\infty\big)} \cdot \mathbf{1} \tag{15}$$

holds simultaneously for all $\frac{3T}{2\log T} \le t \le T$, where $c_3 > 0$ is a constant that only depends on c_1, c_2.

The techniques of Proposition 4.4 are related to Li et al. (2021), but controlling multiple sources of bias and variance relies on quite different ideas.

A similar lower bound is in Proposition 4.5, whose proof is essentially the same as that of Proposition 4.4 hence omitted here.

Proposition 4.5. *Suppose T satisfies the conditions in Theorem 3.1. Then with probability at least $1-\delta$,*

$$\boldsymbol{\Delta}_t \ge -\sqrt{\frac{c_3(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}}(T\log T)^{-1/5}} \times \sqrt{\big(1 + 9t_{mix}(\log T)^{1/5} + T^{1/5} \max_{\lfloor t/2 \rfloor \le i < t} \|\boldsymbol{\Delta}_i\|_\infty\big)}\mathbf{1} \tag{16}$$

holds simultaneously for all t obeying $\frac{3T}{2\log T} \le t \le T$, where $c_3 > 0$ is a constant that only depends on c_1, c_2.

Finally, solving the recursive bounds proves Theorem 3.1; see Appendix A.1 for details.

5 Extension to Learning Bias Function

We now propose a variant of our framework that learns (up to a constant) $q^*(s,a)$ in (1) hence the optimal policy for average reward. In particular, our approach overcomes the non-contraction of the empirical Bellman updates that are often considered in the literature (Abounadi et al., 2001; Zhang et al., 2021), enabling finite-sample analysis of convergence to solutions to (1).

Our new algorithm maintains an estimate $q_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for all $t \ge 1$; in each iteration t, it updates all entries of the estimate for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ via

$$q_t(s,a) = (1-\theta_t)q_{t-1}(s,a) + \theta_t\big[r(s,a) + \alpha_t \max_{a' \in \mathcal{A}} q_{t-1}\big(s_t(s,a), a'\big)\big]. \tag{17}$$

Here $s_t(s,a)$ is the independent sample from the generative model, and α_t and θ_t are the discount factor and learning rate in the t-th iteration, respectively. In addition, we define the value function $v_t : \mathcal{S} \to \mathbb{R}$ in the t-th iteration by $v_t(s) := \max_{a \in \mathcal{A}} q_t(s,a)$ for all $s \in \mathcal{S}$. See Algorithm 2 for a formal statement.

Algorithm 2 Dynamic Horizon q-Learning

$\{\theta_t\}_{t\ge1}$, $\{\alpha_t\}_{t\ge0}$.Initialization: $q_0 \equiv 0$. **for** $t = 1, 2, \ldots$ **do**

3: Generate $s_t(s,a) \sim P(\cdot \mid s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

4: Update $q_t(s,a) = (1-\theta_t)q_{t-1}(s,a) + \theta_t\big[r(s,a) + \alpha_t v_{t-1}(s_t(s,a))\big]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

5: Set $v_t = \max_{a \in \mathcal{A}} q_t(s,a)$ for all $s \in \mathcal{S}$.

6: **end for**

The distinction between this procedure and Algorithm 1 is that we use $r(s,a)$ instead of $(1-\alpha_t)r(s,a)$ in the update. As a result, our estimates approximate another set of dynamic targets: the unscaled discounted

value functions $\{q^*_{\alpha_t}\}$ and $\{v^*_{\alpha_t}\}$ (c.f. (5)), which further approximate solutions to (1).

Lemma 5.1. *There exists a constant $B > 0$ which only depends on the underlying MDP such that $\left|v^*_\alpha(s) - \frac{J^*}{1-\alpha} - v^*(s)\right| \le (1-\alpha)B$ and $\left|q^*_\alpha(s,a) - \frac{J^*}{1-\alpha} - q^*(s,a)\right| \le (1-\alpha)B$ for all state-action pairs $(s,a) \in \mathcal{S}\times\mathcal{A}$, where J^* is the optimal average reward, q^*_α and v^*_α are the unscaled optimal α-discounted Q- (value) functions, and q^* and v^* are the bias functions defined in (2).*

The proof of Lemma 5.1 is in Appendix D.2. Thus,

$$\begin{aligned}&\left|q_t(s,a) - q^*(s,a) - J^*/(1-\alpha_t)\right| \\ &\le (1-\alpha)B + \left|q_t(s,a) - q^*_{\alpha_t}(s,a)\right|,\end{aligned}$$

where, similar to the previous case, it remains to control the estimation error to the dynamic targets.

Theorem 5.2. *In Algorithm 2, we set $\theta_t = \left(1 + \frac{c_1 t^{2/3}}{(\log t)^2}\right)^{-1}$, $\alpha_t = 1 - t^{-1/9}$, $t \ge 2$, for some constant $c_1 > 0$, and set $\theta_1 = \theta_2$, $\alpha_1 = \alpha_2$. Let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Suppose T is sufficiently large such that $T/\log T \ge \max\{e^{2+c_1}, 100, 64(2+c_1)^3\}$, $c_3(\log T)^5 \ge 5$, $c_1 T^3 \ge 3(\log T)^5$, $2c_3 T \ge (1+3/c_1)^2(\log T)^2$ and $c_2 \log T \ge 5(2+c_1)(\log T)^{1/3} + \log\log T$ for some constant $c_2 > 0$ and $c_3 = 16/c_1 + 144/c_1^2$. Then with probability at least $1-\delta$, after T iterations, Algorithm 2 achieves*

$$\left|q_T(s,a) - q^*(s,a) - \frac{J^*}{1-\alpha_T}\right| \le \frac{B + c(\log T)^4(\log\frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{1/9}},$$

$$\text{and } \left|v_T(s) - v^*(s) - \frac{J^*}{1-\alpha_T}\right| \le \frac{B + c(\log T)^4(\log\frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{1/9}}$$

simultaneously for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, where $c > 0$ is an absolute constant that only depends on c_1, c_2, B is the constant given in Lemma 5.1, q^* and v^* are the bias functions defined in (2).

See Appendix C.1 for a sketch of analysis, and Appendix C.2 for a detailed proof.

With the bias function estimator q_T, a natural idea for the task of policy learning is to take the greedy policy with respect to q_T:

$$\pi_T(s) = \operatorname*{argmax}_{a\in\mathcal{A}} \in q_T(s,a).$$

Its suboptimality from the true optimal reward turns out to be of the same scale as the estimation error bound of q_T. We thus have its near-optimality property. The proof of Corollary 5.3 is in Appendix C.3.

Corollary 5.3. *Under the same conditions of Theorem 5.2, the greedy policy π_T with respect to q_T obeys*

$$J^* - J^{\pi_T} \le \frac{2B + 2c(\log T)^4(\log\frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{1/9}}.$$

6 Empirical Validation

We support our theoretical results with preliminary simulations. We design MDPs with $|\mathcal{S}| = 10$ and $|\mathcal{A}| = 8$ according to the construction in Jin and Sidford (2021), which is the hardest instance (in the sense of information-theoretic lower bound) for learning a policy. We vary the mixing time $t_{\text{mix}} \in \{O(10), O(100), O(1000)\}$, which is obtained by tuning the parameter $\gamma \in \{0.1, 0.01, 0.001\}$ in Jin and Sidford (2021). On each MDP instance, we run our algorithm for Theorem 3.3 in 200 independent experiments for $T = 10^7$. We keep the same scheduling of $\gamma_t = 1 - t^{-1/8}$ and $\eta_t = (1 + \frac{t^{5/8}}{(\log t)^2})^{-1}$ for all instances, without adjusting algorithm inputs using any prior knowledge.

We compute the frequency where the algorithm finds the optimal action for each state at $t \in \{10, 10^2, \ldots, 10^6, 10^7\}$ in Figure 1 (for the first 4 states for visualization). The results suggest that our algorithm performs stably well, finding the optimal policy within a reasonable time of training (which might be better than what is predicted by our theory). As promised, it does not need any prior information as input, but works well for MDPs with different t_{mix}. Interestingly, the recovery frequency turns out to be relatively stable across different values of t_{mix}.

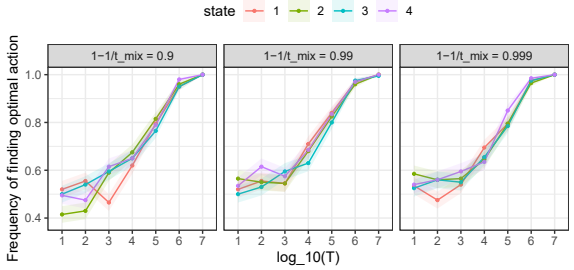

Figure 1: Frequency of recovering optimal action.

Conclusions

In this paper, we provide a feasible framework for average-reward RL that, distinct from existing works, does not require any problem-dependent parameters as input. Our results highlight the algorithmic role of sequentially adjusted discounted factors, along with a carefully selected adaptation scheme, in achieving several average reward learning objectives. We provide finite-sample guarantees for three popular learning tasks. We envision this work as initial progress towards algorithmic design and theoretical understanding of feasible model-free average-reward RL, and an invitation for subsequent efforts in these aspects.

References

Abounadi, J., Bertsekas, D., and Borkar, V. S. (2001). Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698.

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR.

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349.

Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size q-learning. *Systems & control letters*, 61(12):1203–1208.

Blackwell, D. (1962). Discrete dynamic programming. *The Annals of Mathematical Statistics*, pages 719–726.

Borkar, V. S. and Meyn, S. P. (2000). The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469.

Brown, N. and Sandholm, T. (2018). Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424.

Campbell, M., Hoane Jr, A. J., and Hsu, F.-h. (2002). Deep blue. *Artificial intelligence*, 134(1-2):57–83.

Chen, Z., Theja Maguluri, S., Shakkottai, S., and Shanmugam, K. (2020). Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv e-prints*, pages arXiv–2002.

De Farias, D. P. and Van Roy, B. (2006). A cost-shaping linear program for average-cost approximate dynamic programming with performance guarantees. *Mathematics of Operations Research*, 31(3):597–620.

Dewanto, V., Dunn, G., Eshragh, A., Gallagher, M., and Roosta, F. (2020). Average-reward model-free reinforcement learning: a systematic review and literature mapping. *arXiv preprint arXiv:2010.08920*.

Dong, S., Van Roy, B., and Zhou, Z. (2019). Provably efficient reinforcement learning with aggregated states. *arXiv preprint arXiv:1912.06366*.

Dong, S., Van Roy, B., and Zhou, Z. (2021). Simple agent, complex environment: Efficient reinforcement learning with agent state. *arXiv preprint arXiv:2102.05261*.

Even-Dar, E., Kakade, S. M., and Mansour, Y. (2009). Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736.

Even-Dar, E., Mansour, Y., and Bartlett, P. (2003). Learning rates for q-learning. *Journal of machine learning Research*, 5(1).

Feinberg, E. A. and Shwartz, A. (2012). *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., and Pineau, J. (2018). An introduction to deep reinforcement learning. *arXiv preprint arXiv:1811.12560*.

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118.

Fruit, R., Pirotta, M., and Lazaric, A. (2020). Improved analysis of ucrl2 with empirical bernstein inequality. *arXiv preprint arXiv:2007.05456*.

Hordijk, A. and Tijms, H. (1975). A modified form of the iterative method of dynamic programming. *The Annals of Statistics*, pages 203–208.

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*.

Jin, Y. and Sidford, A. (2020). Efficiently solving mdps with stochastic mirror descent. In *International Conference on Machine Learning*, pages 4890–4900. PMLR.

Jin, Y. and Sidford, A. (2021). Towards tight bounds on the sample complexity of average-reward mdps. *arXiv preprint arXiv:2106.07046*.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *IN PROC. 19TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, pages 267–274.

Kakade, S. M. (2003). *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom).

Kearns, M., Mansour, Y., and Ng, A. Y. (2002). A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine learning*, 49(2):193–208.

Kearns, M. and Singh, S. (1999). Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, pages 996–1002.

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Li, G., Cai, C., Chen, Y., Gu, Y., Wei, Y., and Chi, Y. (2021). Is q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1):159–195.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

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

Schaeffer, J., Culberson, J., Treloar, N., Knight, B., Lu, P., and Szafron, D. (1992). A world championship caliber checkers program. *Artificial Intelligence*, 53(2-3):273–289.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018a). Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31.

Sidford, A., Wang, M., Wu, X., and Ye, Y. (2018b). Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.

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

Szepesvári, C. (1998). The asymptotic convergence-rate of q-learning. In *Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 1064–1070.

Tang, Y., Rowland, M., Munos, R., and Valko, M. (2021). Taylor expansions of discount factors. In *Proceedings of the 29th International Conference on Machine Learning (ICML-21)*.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202.

Wainwright, M. J. (2019a). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for q-learning. *arXiv preprint arXiv:1905.06265*.

Wainwright, M. J. (2019b). Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*.

Wan, Y., Naik, A., and Sutton, R. S. (2021). Learning and planning in average-reward markov decision processes. In *International Conference on Machine Learning*, pages 10653–10662. PMLR.

Wang, M. (2017). Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*.

Wang, M. (2020). Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR.

Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR.

Zanette, A., Kochenderfer, M. J., and Brunskill, E. (2019). Almost horizon-free structure-aware best policy identification with a generative model. *Advances in Neural Information Processing Systems*, 32.

Zhang, S., Zhang, Z., and Maguluri, S. T. (2021). Finite sample analysis of average-reward td learning and q-learning. *Advances in Neural Information Processing Systems*, 34.

Zhang, Z. and Ji, X. (2019). Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, 32.

Zhang, Z. and Xie, Q. (2023). Sharper model-free reinforcement learning for average-reward markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable] (Explanation: we will do so when it is possible to de-anonymize the paper.)
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Technical Proofs Regarding Theorem 3.1

In this section, we provide detailed proofs for results regarding Theorem 3.1. Appendix A.1 provides the detailed proof for Theorem 3.1, and the remaining subsections provide proofs for intermediate results in the analysis sketch of Section 4.

Throughout the section, the learning rates and discount factors are

$$\eta_t = \frac{1}{1+\frac{c_1 t^{3/5}}{(\log t)^3}}, \quad \gamma_t = 1 - t^{-1/5},$$

and we choose the proportion

$$\beta = \frac{c_2}{T^{1/5}(\log T)^2}$$

for the constant $c_2 > 0$ in Theorem 3.1.

A.1 Proof of Theorem 3.1

Proof of Theorem 3.1. We solve the recursive bounds in Propositions 4.4 and 4.5 to obtain the final high-probability bound. For any positive integer k, we define

$$u_k = \max\Big\{\|\boldsymbol{\Delta}_i\|_\infty : 2^k \frac{3T}{2\log T} \le i \le T\Big\}.$$

A naive upper bound is $u_k \le 1$ for all k. Furthermore, by (15) and the definition of u_k, with probability at least $1-\delta$, it holds simultaneously for all $k \ge 1$ that

$$u_{k+1} \le \sqrt{\frac{c_3(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}}\Big(\frac{(\log T)^{-1/5}+9t_{\mathrm{mix}}}{T^{1/5}} + \frac{u_k}{(\log T)^{1/5}}\Big)}\mathbf{1}$$

for some absolute constant $c_3 > 0$. If there exists some k such that

$$u_k \le c_4\big((\log T)^{-1/5}+9t_{\mathrm{mix}}\big)\frac{(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}}$$

for some sufficiently large constant $c_4 > 0$, then

$$\begin{aligned} u_{k+1} &\le \sqrt{\frac{c_3(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}}\frac{(\log T)^{-1/5}+9t_{\mathrm{mix}}}{T^{1/5}}\Big(1 + c_4(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)} \\ &\le c_4\big((\log T)^{-1/5}+9t_{\mathrm{mix}}\big)\frac{(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}} \end{aligned}$$

as well. By induction, the above bound holds for all $j \ge k$, hence

$$\|\boldsymbol{\Delta}_T\|_\infty \le c_4\big((\log T)^{-1/5}+9t_{\mathrm{mix}}\big)\frac{(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{S}|T}{\delta}}{T^{1/5}}.$$

On the other hand, suppose $u_j > c_4((\log T)^{-1/5}+9t_{\mathrm{mix}})\frac{(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}}$ for any $1 \le j \le k$. Then we know

$$u_{j+1} \le \sqrt{\frac{2c_3(\log T)^4 \log\frac{|\mathcal{A}||\mathcal{S}|T}{\delta}}{T^{1/5}}u_j}, \quad \text{for all } 1 \le j \le k,$$

which implies

$$\log u_{j+1} \le \frac{1}{2}\log u_j + \frac{1}{2}\log\theta, \quad \text{where } \theta = 2c_3 \cdot T^{-1/5}(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$$

for some constant $c_6 > 0$. Recursively applying this relation for $1 \le j < k$ yields

$$2^k \log u_k \le \log u_0 + \sum_{j=0}^{k-1} 2^j \log \theta \le (2^k - 1) \log \theta,$$

where we used the fact that $u_0 \le 1$. Hence

$$\begin{aligned} u_k \le \theta^{1-1/2^k} &= \Big(2c_3 \cdot T^{-1/5} (\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)^{1-1/2^k} \\ &= \Big(2c_3 \cdot (\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)^{1-1/2^k} \cdot T^{-1/5} \cdot T^{\frac{2^{-k}}{5}} \\ &\le 2c_3 \cdot \frac{(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}} \cdot \exp\big(\log T \cdot 2^{-k}/5\big). \end{aligned}$$

Now we can set $2^k \ge \log T/5$, so that the above upper bound translates to

$$\|\boldsymbol{\Delta}_T\|_\infty \le u_k \le 6c_3 \cdot \frac{(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}}.$$

Combining the two cases above, we arrive at

$$\|\boldsymbol{\Delta}_T\|_\infty \le \big[c_4\big((\log T)^{-1/5} + 9t_{\mathrm{mix}}\big) + 6c_3\big] \frac{(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}}$$

with probability at least $1 - \delta$ for T obeying the conditions of Proposition 4.4. Furthermore, since $t_{\mathrm{mix}} \ge 1$, the above bound translates to

$$\|\boldsymbol{\Delta}_T\|_\infty \le c t_{\mathrm{mix}} \frac{(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/5}}$$

for some constant $c > 0$ that only depends on c_1, c_2, which completes the proof of Theorem 3.1. □

A.2 Proof of Lemma 3.2

Proof of Lemma 3.2. For any discounted factor $\gamma \in (0,1)$, the optimal policy π^*_γ with respect to γ-discounted reward attains the optimal Q-function, i.e., $\boldsymbol{Q}^*_\gamma = \boldsymbol{Q}^{\pi^*_\gamma}_\gamma$. Similarly, the optimal policy π^* with repsect to average reward satisfies $J^* = J^{\pi^*}$. Applying Lemma D.1 to π^*_γ and γ leads to

$$Q^*_\gamma(s,a) = Q^{\pi^*_\gamma}_\gamma(s,a) \le J^{\pi^*_\gamma} + 3(1-\gamma)t_{\mathrm{mix}} \le J^* + 3(1-\gamma)t_{\mathrm{mix}}$$

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. On the other hand, applying Lemma D.1 to π^* and γ leads to

$$Q^*_\gamma(s,a) \ge Q^{\pi^*}_\gamma(s,a) \ge J^{\pi^*} - 3(1-\gamma)t_{\mathrm{mix}} \ge J^* - 3(1-\gamma)t_{\mathrm{mix}}$$

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. The last steps of the two equations follow from the bounded reward averaging time for π^*_γ and π^*. Combining the two inequalities, we complete the proof of Lemma 3.2. □

A.3 Proof of Proposition 4.4

Proof of Proposition 4.4. Combining Lemma 4.1, 4.2 and 4.3, for sufficiently larget T obeying the conditions of Theorem 3.1 and any fixed t within $T/\log T \le t \le T$, it holds with probability at least $1 - \delta$ that

$$\begin{aligned} \boldsymbol{\Delta}_t \le \frac{2(\log T)^{2/5}}{T^{2/5}} + \frac{2}{T} + 5\sqrt{\frac{(\log T)^{\frac{19}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{c_1 T^{3/5}} \Big(\max_{\lfloor (1-\beta)t \rfloor < i \le t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-1/5}\mathbf{1}\Big)} \\ + \sum_{i=1+\lfloor (1-\beta)t \rfloor}^{t} \eta_i^{(t)} \gamma_i \boldsymbol{P}^{\pi_{t-1}} \boldsymbol{\Delta}_{i-1} \end{aligned}$$

$$\leq \Big(1+\frac{5}{\sqrt{c_1}}\Big)\sqrt{\frac{(\log T)^{\frac{19}{5}}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{3/5}}\Big(\max_{\lfloor(1-\beta)t\rfloor<i\leq t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1})+T^{-1/5}\mathbf{1}\Big)}$$
$$+\sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t}\eta_i^{(t)}\gamma_i\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}, \tag{18}$$

where the second inequality follows from the fact that

$$\frac{2(\log T)^{2/5}}{T^{2/5}}+\frac{2}{T}\leq\frac{4(\log T)^{2/5}}{T^{2/5}}\leq\sqrt{\frac{(\log T)^{\frac{19}{5}}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{4/5}}}$$

for T so large that $T/\log T\geq 300$. Once $(1-\beta)\geq 3/4$, for any fixed t obeying $3T/(2\log T)\leq t\leq T$, we apply (18) with a union bound over $\{k\colon \frac{2t}{3}\leq k\leq t\}$ — as a result, with probability at least $1-\delta$, one has

$$\boldsymbol{\Delta}_k\leq\sqrt{\boldsymbol{\varphi}_t}+\sum_{i=1+\lfloor(1-\beta)k\rfloor}^{k}\eta_i^{(k)}\gamma_i\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1},\qquad\text{for all }\frac{2t}{3}\leq k\leq t, \tag{19}$$

where we define

$$\boldsymbol{\varphi}_t=\Big(1+\frac{5}{\sqrt{c_1}}\Big)^2\cdot\frac{(\log T)^{\frac{19}{5}}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{3/5}}\Big(\max_{\lfloor t/2\rfloor<i\leq t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1})+T^{-1/5}\mathbf{1}\Big) \tag{20}$$
$$\geq 2\Big(1+\frac{5}{\sqrt{c_1}}\Big)^2\cdot\frac{(\log T)^{\frac{19}{5}}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta/T}}{T^{3/5}}\Big(\max_{\lfloor(1-\beta)k\rfloor<i\leq k}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1})+T^{-1/5}\mathbf{1}\Big).$$

Before proceeding to recursively bound $\boldsymbol{\Delta}_t$, we define $\{\alpha_i^{(t)}\}$ as

$$\alpha_i^{(t)}:=\frac{\eta_{i+1}^{(t)}}{\sum_{j=\lfloor(1-\beta)t\rfloor}^{t-1}\eta_{j+1}^{(t)}}, \tag{21}$$

which, following Li et al. (2021), satisfies

$$\alpha_i^{(t)}\geq\eta_{i+1}^{(t)}\quad\text{and}\quad\sum_{j=\lfloor(1-\beta)t\rfloor}^{t-1}\alpha_{j+1}^{(t)}=1 \tag{22}$$

for all t. When T is sufficiently large so that $1-\beta\geq 2/3$, We decompose (19) as

$$\boldsymbol{\Delta}_k\leq\sqrt{\boldsymbol{\varphi}_t}+\sum_{i=1+\lfloor(1-\beta)k\rfloor}^{k}\eta_i^{(k)}\gamma_i\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}=\sum_{i_1=\lfloor(1-\beta)k\rfloor}^{k-1}\Big(\alpha_{i_1}^{(k)}\sqrt{\boldsymbol{\varphi}_t}+\eta_{i_1+1}^{(k)}\gamma_{i_1+1}\boldsymbol{P}^{\pi_{i_1}}\boldsymbol{\Delta}_{i_1}\Big).$$

We recursively apply the above relation in a manner similar to Equation 68 of Li et al. (2021), yet with a sequence of dynamic discount factor $\{\gamma_t\}$:

$$\begin{aligned}
\boldsymbol{\Delta}_t&\leq\sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\Big(\alpha_{i_1}^{(t)}\sqrt{\boldsymbol{\varphi}_t}+\eta_{i_1+1}^{(t)}\gamma_{i_1+1}\boldsymbol{P}^{\pi_{i_1}}\boldsymbol{\Delta}_{i_1}\Big)\\
&\leq\sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\Big[\alpha_{i_1}^{(t)}\sqrt{\boldsymbol{\varphi}_t}+\eta_{i_1+1}^{(t)}\gamma_{i_1+1}\boldsymbol{P}^{\pi_{i_1}}\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\Big(\alpha_{i_2}^{(i_1)}\sqrt{\boldsymbol{\varphi}_t}+\eta_{i_2+1}^{(i_1)}\gamma_{i_2+1}\boldsymbol{P}^{\pi_{i_2}}\boldsymbol{\Delta}_{i_2}\Big)\Big]\\
&\leq\sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\alpha_{i_1}^{(t)}\sqrt{\boldsymbol{\varphi}_t}+\sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\alpha_{i_1}^{(t)}\alpha_{i_2}^{(i_1)}\big(\gamma_{i_1+1}\boldsymbol{P}^{\pi_{i_1}}\big)\sqrt{\boldsymbol{\varphi}_t}\\
&\quad+\sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\eta_{i_1+1}^{(t)}\eta_{i_2+1}^{(i_1)}\prod_{k=1}^{2}(\gamma_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\boldsymbol{\Delta}_{i_2}\Big)
\end{aligned}$$

$$\leq \sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1} \sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1} \alpha_{i_1}^{(t)}\alpha_{i_2}^{(i_1)}\{\boldsymbol{I}+\gamma_{i_1+1}\boldsymbol{P}^{\pi_{i_1}}\}\sqrt{\boldsymbol{\varphi}_t}$$
$$+ \sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1} \sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1} \eta_{i_1+1}^{(t)}\eta_{i_2+1}^{(i_1)} \prod_{k=1}^{2}(\gamma_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\boldsymbol{\Delta}_{i_2}, \tag{23}$$

where the third line uses $\eta_{i_1+1}^{(t)} \leq \alpha_i^{(t)}$ in (22), and the last line uses $\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\alpha_{i_2}^{(i_1)} = 1$ in (22). We shall further recursively apply the above relation. To begin with, we set

$$H := T^{1/5}(\log T)^2 \quad \text{and} \quad \alpha_{\{i_k\}_{k=1}^H} := \alpha_{i_1}^{(t)}\alpha_{i_2}^{(i_1)}\ldots\alpha_{i_H}^{(i_{H-1})} \geq 0$$

for any $t > i_1 > i_2 > \cdots > i_H$, which (according to (22)) satisfies

$$\alpha_{\{i_k\}_{k=1}^H} \geq \eta_{i_1+1}^{(t)}\eta_{i_2+1}^{(i_1)}\cdots\eta_{i_H+1}^{(i_{H-1})}.$$

We define the index set

$$\mathcal{I}_t = \big\{(i_1,\ldots,i_H)\colon \lfloor(1-\beta)t\rfloor \leq i_1 \leq t-1,\ \lfloor(1-\beta)i_{j-1}\rfloor \leq i_j \leq i_{j-1}-1,\ \forall 1\leq j<H\big\},$$

which satisfies

$$\sum_{(i_1,\ldots,i_H)\in\mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H} = 1.$$

By definition of β, we have

$$(1-\beta)^H \geq \exp(-\beta H) \geq \frac{2}{3}$$

for sufficiently small $c_2 \leq \log(3/2)$, which implies

$$i_1 > i_2 > \cdots > i_H \geq 2t/3, \quad \text{for all } (i_1,\ldots,i_H)\in\mathcal{I}_t.$$

Recursively invoking (23), we obtain

$$\boldsymbol{\Delta}_t \leq \sum_{(i_1,\ldots,i_H)\in\mathcal{I}_t} \alpha_{\{i_k\}_{k=1}^H}\bigg\{\Big(\boldsymbol{I}+\sum_{h=1}^{H-1}\prod_{k=1}^{h}(\gamma_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\Big)\sqrt{\boldsymbol{\varphi}_t} + \prod_{h=1}^{H}(\gamma_{i_h+1}\boldsymbol{P}^{\pi_{i_h}})|\boldsymbol{\Delta}_{i_H}|\bigg\}$$
$$\leq \max_{(i_1,\ldots,i_H)\in\mathcal{I}_t}\bigg\{\underbrace{\Big(\boldsymbol{I}+\sum_{h=1}^{H-1}\prod_{k=1}^{h}(\gamma_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\Big)\sqrt{\boldsymbol{\varphi}_t}}_{=:\boldsymbol{\beta}_1} + \underbrace{\prod_{h=1}^{H}(\gamma_{i_h+1}\boldsymbol{P}^{\pi_{i_h}})|\boldsymbol{\Delta}_{i_H}|}_{=:\boldsymbol{\beta}_2}\bigg\}, \tag{24}$$

where the second line uses $\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\alpha_{i_2}^{(i_1)} = 1$ in (22). In the following, we are to bound $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in (24) separately. The easier term is $\boldsymbol{\beta}_2$: noting that by definition of the discount factor $\{\gamma_t\}$, we have $\gamma_j \leq \gamma_T$ for all $2t/3 \leq j \leq T$, indicating

$$|\boldsymbol{\beta}_2| \leq \gamma_t^H \prod_{h=1}^{H}\boldsymbol{P}^{\pi_{i_h}}|\boldsymbol{\Delta}_H| \leq \gamma_T^H\bigg\|\prod_{h=1}^{H}\boldsymbol{P}^{\pi_{i_h}}\bigg\|_1\|\boldsymbol{\Delta}_H\|_\infty \overset{\text{(i)}}{\leq} \gamma_T^H \overset{\text{(ii)}}{\leq} \frac{1}{T}, \tag{25}$$

where (i) follows from the bounded magnitude $\|\boldsymbol{\Delta}_H\|_\infty \leq 1$ and the fact that $\prod_{h=1}^{H}\boldsymbol{P}^{\pi_{i_h}}$ is a probability transition matrix; (ii) follows from

$$\gamma_T^H = \big(1-T^{-1/5}\big)^{T^{1/5}(\log T)^2} \leq \exp\big(-(\log T)^2\big) \leq \frac{1}{T}.$$

Moving on to $\boldsymbol{\beta}_1$, its entrywise square can be upper bounded as

$$|\boldsymbol{\beta}_1|^2 = \bigg|\sum_{h=0}^{H-1}\prod_{k=1}^{h}(\gamma_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\sqrt{\boldsymbol{\varphi}_t}\bigg|^2 \leq \bigg|\sum_{h=0}^{H-1}\gamma_T^h\Big(\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\Big)\sqrt{\boldsymbol{\varphi}_t}\bigg|^2$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \left| \sum_{h=0}^{H-1} \gamma_T^{h/2}\gamma_T^{h/2} \sqrt{\textstyle\prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \boldsymbol{\varphi}_t} \right|^2 \\
&\overset{\text{(ii)}}{\leq} \sum_{h=0}^{H-1} \gamma_T^h \cdot \sum_{h=0}^{H-1} \gamma_T^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \boldsymbol{\varphi}_t \\
&\overset{\text{(iii)}}{\leq} \frac{1}{1-\gamma_T} \sum_{h=0}^{H-1} \gamma_T^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \frac{2(1+\frac{5}{\sqrt{c_1}})^2 (\log T)^{\frac{19}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta/T}}{T^{3/5}} \Big(\max_{\lfloor (1-\beta)k \rfloor < i \leq k} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-1/5}\mathbf{1} \Big) \\
&\overset{\text{(iv)}}{\leq} \frac{4(1+\frac{5}{\sqrt{c_1}})^2 (\log T)^{\frac{19}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{2/5}} \Big(\sum_{h=0}^{H-1} \gamma_T^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \max_{\lfloor t/2 \rfloor \leq i < t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) + \mathbf{1} \Big),
\end{aligned} \tag{26}
$$

where (i) follows from Jensen's inequality noting that $\prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}}$ is a probability transition matrix, (ii) follows from the Cauchy-Schwarz inequality, (iii) follows from the definition of $\boldsymbol{\varphi}_t$ in (20), and (iv) follows from the fact that $\prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \mathbf{1} = \mathbf{1}$ since $\prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}}$ is a probability transition matrix. We employ the following lemma to bound the first term of (26), whose proof is in Appendix D.4.

Lemma A.1. *Suppose $T \geq 160$, then it holds for all $T/\log T \leq t \leq T$ that*

$$
\sum_{h=0}^{H-1} \gamma_T^h \prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \max_{\lfloor t/2 \rfloor \leq i < t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) \leq \Big(7 + 72 t_{mix} (\log T)^{1/5} + 8T^{1/5} \max_{\lfloor t/2 \rfloor \leq i < t} \|\boldsymbol{\Delta}_i\|_\infty \Big) \mathbf{1}. \tag{27}
$$

Invoking Lemma A.1, we obtain an upper bound

$$
|\boldsymbol{\beta}_1|^2 \leq \frac{4(1+\frac{5}{\sqrt{c_1}})^2 (\log T)^{\frac{19}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{2/5}} \Big(8 + 72 t_{\mathrm{mix}} (\log T)^{1/5} + 8T^{1/5} \max_{\lfloor t/2 \rfloor \leq i < t} \|\boldsymbol{\Delta}_i\|_\infty \Big) \mathbf{1}. \tag{28}
$$

Recalling (24), by the upper bound (25) on $|\boldsymbol{\beta}_2|$ and (28) on $|\boldsymbol{\beta}_1|$, for any fixed t such that $3T/(2\log T) \leq t \leq T$,

$$
\begin{aligned}
\boldsymbol{\Delta}_t &\leq \frac{1}{T} + \sqrt{\frac{4(1+\frac{5}{\sqrt{c_1}})^2 (\log T)^{\frac{19}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{2/5}} \Big(8 + 72 t_{\mathrm{mix}} (\log T)^{1/5} + 8T^{1/5} \max_{\lfloor t/2 \rfloor \leq i < t} \|\boldsymbol{\Delta}_i\|_\infty \Big) \mathbf{1}} \\
&\leq 4\Big(1 + \frac{5}{\sqrt{c_1}}\Big) \sqrt{\frac{(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{2/5}} \Big(\frac{8}{(\log T)^{1/5}} + 72 t_{\mathrm{mix}} + \frac{8T^{1/5}}{(\log T)^{1/5}} \max_{\lfloor t/2 \rfloor \leq i < t} \|\boldsymbol{\Delta}_i\|_\infty \Big) \mathbf{1}}
\end{aligned} \tag{29}
$$

holds with probability at least $1-\delta$.

We now further take a union bound for (29) over $\{t\colon 3T/(2\log T) \leq t \leq T\}$, which leads to the simultaneous $(1-\delta)$ high-probability bound

$$
\boldsymbol{\Delta}_t \leq 16\Big(1 + \frac{5}{\sqrt{c_1}}\Big) \sqrt{\frac{(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{2/5}} \Big((\log T)^{-1/5} + 9 t_{\mathrm{mix}} + (T/\log T)^{1/5} \max_{\lfloor t/2 \rfloor \leq i < t} \|\boldsymbol{\Delta}_i\|_\infty \Big) \mathbf{1}}
$$

for all t such that $3T/(2\log T) \leq t \leq T$, as $\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta/T} \leq 2 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$. This completes the proof of Proposition 4.4.

□

A.4 Proof of Lemma 4.1

Proof of Lemma 4.1. Employing Lemma D.4, each term $\boldsymbol{d}_t$ can be bounded as

$$
\begin{aligned}
\|\boldsymbol{d}_t\|_\infty &\leq (1-\eta_t) \big\| \boldsymbol{Q}^*_{\gamma_{t-1}} - \boldsymbol{Q}^*_{\gamma_t} \big\|_\infty + \eta_t \gamma_t \big\| \boldsymbol{P}(\boldsymbol{V}^*_{\gamma_{t-1}} - \boldsymbol{V}^*_{\gamma_t}) \big\|_\infty \\
&\leq (1-\eta_t) \big\| \boldsymbol{Q}^*_{\gamma_{t-1}} - \boldsymbol{Q}^*_{\gamma_t} \big\|_\infty + \eta_t \gamma_t \|P\|_1 \big\| \boldsymbol{V}^*_{\gamma_{t-1}} - \boldsymbol{V}^*_{\gamma_t} \big\|_\infty \\
&\leq (1-\eta_t) \big\| \boldsymbol{Q}^*_{\gamma_{t-1}} - \boldsymbol{Q}^*_{\gamma_t} \big\|_\infty + \eta_t \gamma_t \big\| \boldsymbol{V}^*_{\gamma_{t-1}} - \boldsymbol{V}^*_{\gamma_t} \big\|_\infty \leq \frac{\gamma_{t-1} - \gamma_t}{1-\gamma_{t-1}},
\end{aligned} \tag{30}
$$

where the second line uses $\|\boldsymbol{AB}\|_\infty \le \|\boldsymbol{A}\|_1\|\boldsymbol{B}\|_\infty$ for matrices $\boldsymbol{A}, \boldsymbol{B}$, the third line follows from $\|\boldsymbol{P}\|_1 = 1$ since each row of $\boldsymbol{P}$ is a probability vector. For any $t > 1$, we let $\alpha = t^{-2/5} \in (0,1)$. By the triangular inequality, the switching error can be bounded as

$$\begin{aligned}\Big\|\sum_{i=1}^t \eta_i^{(t)} \boldsymbol{d}_i/\eta_i\Big\|_\infty &\le \sum_{i=1}^t \prod_{j=i}^t (1-\eta_j)\|\boldsymbol{d}_i\|_\infty \\ &= \sum_{i=1}^{\lfloor(1-\alpha)t\rfloor} \prod_{j=i}^t (1-\eta_j)\|\boldsymbol{d}_i\|_\infty + \sum_{1+\lfloor(1-\alpha)t\rfloor}^t \prod_{j=i}^t (1-\eta_j)\|\boldsymbol{d}_i\|_\infty.\end{aligned}$$

Firstly, when t obeys $\lfloor(1-\alpha)t\rfloor = \lfloor t - t^{2/5}\rfloor \ge 150$ (which is satisfied by $t \ge 160$), for any i such that $1 \le i \le \lfloor(1-\alpha)t\rfloor$, we have

$$\begin{aligned}\prod_{j=i}^t (1-\eta_j) &\le \prod_{j=\lfloor(1-\alpha)t\rfloor}^t (1-\eta_j) \le \exp\Big(-\textstyle\sum_{j=\lfloor(1-\alpha)t\rfloor}^t \eta_j\Big) \\ &\overset{\text{(i)}}{\le} \exp\bigg(-\sum_{j=\lfloor(1-\alpha)t\rfloor}^t \frac{(\log j)^3}{(10+c_1)j^{3/5}}\bigg) \overset{\text{(ii)}}{\le} \exp\bigg(-\frac{\alpha t(\log t)^3}{(10+c_1)t^{3/5}}\bigg),\end{aligned}$$

where (i) follows from (99), and (ii) follows from (100) for $j \ge \lfloor(1-\alpha)t\rfloor \ge 150$. Together with the fact that $\|\boldsymbol{d}_i\|_\infty \le 1$, the above bound implies

$$\sum_{i=1}^{\lfloor(1-\alpha)t\rfloor} \prod_{j=i}^t (1-\eta_j)\|\boldsymbol{d}_i\|_\infty \le t \cdot \exp\bigg(-\frac{\alpha t^{2/5}(\log t)^3}{10+c_1}\bigg) = \exp\big(\log t - (\log t)^3/(10+c_1)\big). \tag{31}$$

On the other hand, (30) implies

$$\|\boldsymbol{d}_i\|_\infty \le \frac{(i-1)^{-1/5} - i^{-1/5}}{(i-1)^{-1/5}} \le 1 - (1-1/i)^{1/5} \le \frac{2}{5i} \le \frac{2}{5(1-\alpha)t},$$

for all $i \ge 1 + \lfloor(1-\alpha)t\rfloor$. The second inequality above is implied by

$$1 - (1-1/i)^{1/5} = 1 - \exp\bigg(\frac{\log(1-1/i)}{5}\bigg) \le 1 - \exp\left(-2/(5i)\right) \le 2/(5i),$$

where for $i \ge 2$, we use the fact that $\log(1-x) \ge -2x$ for $x \le \log 2/2$ and $e^x \ge 1 + x$ for $x \in \mathbb{R}$. Therefore, noting that $\alpha \le 0.2$ for $t \ge 150$, we have

$$\sum_{1+\lfloor(1-\alpha)t\rfloor}^t \prod_{j=i}^t (1-\eta_j)\|\boldsymbol{d}_i\|_\infty \le \sum_{1+\lfloor(1-\alpha)t\rfloor}^t \|\boldsymbol{d}_i\|_\infty \le \frac{2\alpha t}{5(1-\alpha)t} \le \alpha/2 = \frac{1}{2t^{2/5}}. \tag{32}$$

Combining (31) and (32), once $T/\log T \ge 160$, we have

$$\begin{aligned}\Big\|\sum_{i=1}^t \eta_i^{(t)} \boldsymbol{d}_i/\eta_i\Big\|_\infty &\le \exp\big(\log t - (\log t)^3/(10+c_1)\big) + \frac{1}{2t^{2/5}} \\ &\le \exp\Big(\log T - \frac{(\log T - \log\log T)^3}{10+c_1}\Big) + \frac{(\log T)^{2/5}}{2T^{2/5}} \\ &\le \exp\Big(\log T - \frac{(\log T)^3}{8(10+c_1)}\Big) + \frac{(\log T)^{2/5}}{2T^{2/5}} \le \frac{2(\log T)^{2/5}}{T^{2/5}}\end{aligned}$$

for all $T/\log T \le t \le T$, where the third inequality follows from $\log T \ge 2\log\log T$, and the last inequality follows from $(\log T)^3/(80+8c_1) \ge \frac{7}{5}\log T$. Therefore, we complete the proof of Lemma 4.1. □

A.5 Proof of Lemma 4.2

Proof of Lemma 4.2. The ℓ_∞-norm of $\boldsymbol{\zeta}_t$ can be bounded as

$$\begin{aligned}\|\boldsymbol{\zeta}_t\|_\infty &\le \big\|\eta_0^{(t)}\boldsymbol{\Delta}_0\big\|_\infty + \sum_{i=1}^{\lfloor(1-\beta)t\rfloor}\eta_i^{(t)}\gamma_i\big\|(\boldsymbol{P}_i-\boldsymbol{P})\boldsymbol{V}_{i-1}+\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{i-1}\big\|_\infty\\ &\overset{\text{(i)}}{\le} \eta_0^{(t)} + \sum_{i=1}^{\lfloor(1-\beta)t\rfloor}\eta_i^{(t)}\gamma_i\big[\|\boldsymbol{P}_i-\boldsymbol{P}\|_1\|\boldsymbol{V}_{i-1}\|_\infty + \|\boldsymbol{P}^{\pi_{t-1}}\|_1\|\boldsymbol{\Delta}_{i-1}\|_\infty\big]\\ &\overset{\text{(ii)}}{\le} \eta_0^{(t)} + 3\sum_{i=1}^{\lfloor(1-\beta)t\rfloor}\eta_i^{(t)}\gamma_i.\end{aligned}$$

Here (i) uses $\|\boldsymbol{AB}\|_\infty \le \|\boldsymbol{A}\|_1\|\boldsymbol{B}\|_\infty$ for matrices $\boldsymbol{A},\boldsymbol{B}$, and (ii) follows from the fact that $\|\boldsymbol{P}_i-\boldsymbol{P}\|_1 \le \|\boldsymbol{P}_i\|_1+\|\boldsymbol{P}\|_1 \le 2$ and $\|\boldsymbol{P}^{\pi_{t-1}}\|_1 \le 1$ (since they are all probability matrices), and the bounded magnitudes $\|\boldsymbol{V}_{i-1}\|_\infty \le 1$ and $\|\boldsymbol{\Delta}_{i-1}\|_\infty \le 1$.

By definition of $\{\eta_t\}$, as long as $\beta < 1/2$ and $t \ge 300$,

$$\begin{aligned}\eta_0^{(t)} &= \prod_{j=1}^{t}(1-\eta_j) \le \exp\Big(-\textstyle\sum_{j=1}^{t}\eta_j\Big)\\ &\le \exp\bigg(-\sum_{j=150}^{t}\frac{(\log j)^3}{(10+c_1)j^{3/5}}\bigg) \le \exp\bigg(-\frac{(t-150)(\log t)^3}{(10+c_1)t^{3/5}}\bigg),\end{aligned} \tag{33}$$

where the second inequality follows from (99), and the last inequality follows from the monotonicity (100).

Similarly, when $\lfloor(1-\beta)t\rfloor \ge 150$ (satisfied as long as $\beta < 1/2$ and $t \ge 300$), for any $1 \le i \le \lfloor(1-\beta)t\rfloor$, it holds that

$$\begin{aligned}\eta_i^{(t)} &\le \prod_{j=1+\lfloor(1-\beta)t\rfloor}^{t}(1-\eta_j) \le \exp\Big(-\textstyle\sum_{j=1+\lfloor(1-\beta)t\rfloor}^{t}\eta_j\Big)\\ &\le \exp\bigg(-\sum_{j=1+\lfloor(1-\beta)t\rfloor}^{t}\frac{(\log j)^3}{(10+c_1)j^{3/5}}\bigg) \le \exp\bigg(-\frac{\beta t(\log t)^3}{(10+c_1)t^{3/5}}\bigg),\end{aligned} \tag{34}$$

where the last inequality follows from (100). Combining (33) and (34), once $\beta < 1/2$, we have

$$\begin{aligned}\|\boldsymbol{\zeta}_t\|_\infty &\le \exp\bigg(-\frac{(t-150)(\log t)^3}{(10+c_1)t^{3/5}}\bigg) + 3\exp\bigg(\log t - \frac{\beta t(\log t)^3}{(10+c_1)t^{3/5}}\bigg)\\ &\le \exp\bigg(-\frac{t(\log t)^3}{2(10+c_1)t^{3/5}}\bigg) + 3\exp\bigg(\log t - \frac{\beta t(\log t)^3}{(10+c_1)t^{3/5}}\bigg)\end{aligned}$$

for all $t \ge 300$. Note that when $T/\log T \ge 300$, for all t such that $T/\log T \le t \le T$,

$$t^{2/5}(\log t)^3 \ge (T/\log T)^{2/5}(\log T - \log\log T)^3 \ge (T/\log T)^{2/5}\cdot(\log T)^3/8 \ge T^{2/5}(\log T)^2/8,$$

where the second inequality follows from $\log T \ge 2\log\log T$. Meanwhile, for $T/\log T \le t \le T$,

$$\log t - \frac{\beta t(\log t)^3}{(10+c_1)t^{3/5}} \le \log T - \frac{c_2}{T^{1/5}(\log T)^2}\frac{T^{2/5}}{(\log T)^{2/5}}\frac{(\log T)^3}{8(10+c_1)} \le \log T - \frac{c_2T^{1/5}(\log T)^{3/5}}{8(10+c_1)}.$$

Therefore, for any $T/\log T \le t \le \log T$,

$$\|\boldsymbol{\zeta}_t\|_\infty \le \exp\Big(-\frac{T^{2/5}(\log T)^2}{16(10+c_1)}\Big) + \exp\Big(\log(3T) - \frac{c_2T^{1/5}(\log T)^{3/5}}{8(10+c_1)}\Big) \le \frac{2}{T}$$

if T obeys (satisfied by the conditions of Theorem 3.1)

$$T^{2/5}\log T \ge 16(10+c_1) \quad \text{and} \quad c_2T^{1/5} \ge 24(10+c_1)(\log T)^{2/5},$$

which completes the proof of Lemma 4.2. □

A.6 Proof of Lemma 4.3

Proof of Lemma 4.3. We begin with some basic notations for convenience: write

$$\boldsymbol{\xi}_t = \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \boldsymbol{z}_i, \quad \text{where} \quad \boldsymbol{z}_i = \eta_i^{(t)}\gamma_i(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}. \tag{35}$$

Then (entries of) $\{\boldsymbol{z}_i\}$ are martingale differences in the sense that

$$\mathbb{E}[\boldsymbol{z}_i \,|\, \boldsymbol{V}_{i-1}, \ldots, \boldsymbol{V}_0] = \boldsymbol{0} \quad \forall i > \lfloor(1-\beta)t\rfloor.$$

We shall obtain high-probability bound on $\|\boldsymbol{\zeta}_t\|_\infty$ via Freedman's inequality in Lemma D.6, for which we compute some basic quantities that would be of use.

Firstly, when $\lfloor(1-\beta)t\rfloor \geq 150$ and $1-\beta \geq 1/2$,

$$\begin{aligned}
\max_{\lfloor(1-\beta)t\rfloor < i \leq t} \|\boldsymbol{z}_i\|_\infty &\leq \max_{\lfloor(1-\beta)t\rfloor < i \leq t} \eta_i^{(t)}\gamma_i \|\boldsymbol{P}_i - \boldsymbol{P}\|_1 \|\boldsymbol{V}_{i-1}\|_\infty \\
&\leq \max_{\lfloor(1-\beta)t\rfloor < i \leq t} 2\eta_i \\
&\leq \frac{4(\log t)^3}{c_1 t^{3/5}} =: R,
\end{aligned} \tag{36}$$

where the second inequality follows from $\|\boldsymbol{P}_i - \boldsymbol{P}\|_1 \leq \|\boldsymbol{P}_i\|_1 + \|\boldsymbol{P}\|_1 \leq 2$ and the bounded magnitude $\|\boldsymbol{V}_{i-1}\|_\infty \leq 1$, and the third inequality follows from (103).

We denote the entrywise conditional variance of $\boldsymbol{z}_i$ as $\mathrm{Var}(\boldsymbol{z}_i \,|\, \boldsymbol{V}_{i-1}, \ldots, \boldsymbol{V}_0) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, whose j-th element is the variance of $[z_i]_j$ conditional on $\boldsymbol{V}_{i-1}, \ldots, \boldsymbol{V}_0$. By the definition of $\boldsymbol{z}_i$ in (35),

$$\begin{aligned}
\boldsymbol{W}_t &= \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \mathrm{Var}(\boldsymbol{z}_i \,|\, \boldsymbol{V}_{i-1}, \ldots, \boldsymbol{V}_0) \\
&= \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \gamma_i^2 \big(\eta_i^{(t)}\big)^2 \mathrm{Var}\left((\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1} \,\middle|\, \boldsymbol{V}_{i-1}\right) = \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \gamma_i^2 \big(\eta_i^{(t)}\big)^2 \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}).
\end{aligned}$$

It can be (entrywisely) upper bounded as

$$\begin{aligned}
\boldsymbol{W}_t &\leq \Big(\max_{\lfloor(1-\beta)t\rfloor < i \leq t} \eta_i^{(t)}\Big) \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \gamma_i^2 \eta_i^{(t)} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) \\
&\leq \Big(\max_{\lfloor(1-\beta)t\rfloor < i \leq t} \eta_i^{(t)}\Big)\Big(\sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \eta_i^{(t)}\Big)\Big(\max_{\lfloor(1-\beta)t\rfloor < i \leq t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1})\Big) \\
&\leq \frac{2(\log t)^3}{c_1 t^{3/5}} \max_{\lfloor(1-\beta)t\rfloor < i \leq t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}),
\end{aligned}$$

where the last inequality follows from (103) and the fact that $\sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \eta_i^{(t)} \leq 1$ due to (101). In addition, since $\|\boldsymbol{V}_{i-1}\|_\infty \leq 1$, a deterministic upper bound of $\boldsymbol{W}_t$ is given by

$$\|\boldsymbol{W}_t\|_\infty \leq \frac{2(\log t)^3}{c_1 t^{3/5}} =: \sigma^2. \tag{37}$$

To apply Freedman's inequality, we further choose the (smallest) positive integer K such that for some constant $c_2 \geq 1$,

$$\frac{\sigma^2}{2^K} \leq \frac{2(\log t)^3}{c_1 t^{4/5}},$$

which leads to

$$\frac{\log t}{5\log 2} \le K \le \frac{\log t}{5\log 2} + 1.$$

In view of (36) and (37) together with a union bound over all $|\mathcal{S}||\mathcal{A}|$ entries, Freedman's inequality (Lemma D.6) implies that for any fixed t obeying $\lfloor (1-\beta)t \rfloor \ge 150$ and $\beta < 1/2$,

$$\begin{aligned}|\boldsymbol{\xi}_t| &\le \frac{8(\log t)^3}{3c_1 t^{3/5}} \log \frac{2|\mathcal{S}||\mathcal{A}|\log t}{\delta \log 2}\mathbf{1} + \sqrt{8\Big(\boldsymbol{W}_t + \frac{2(\log t)^3}{c_1 t^{4/5}}\mathbf{1}\Big)\log \frac{2|\mathcal{S}||\mathcal{A}|\log t}{\delta \log 2}} \\ &\le \frac{8(\log t)^3 \log \frac{|\mathcal{S}||\mathcal{A}|t}{\delta}}{3c_1 t^{3/5}}\mathbf{1} + \sqrt{\frac{16(\log t)^3 \log \frac{|\mathcal{S}||\mathcal{A}|t}{\delta}}{c_1 t^{3/5}}\Big(\max_{\lfloor (1-\beta)t\rfloor < i \le t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + t^{-1/5}\mathbf{1}\Big)} \end{aligned} \tag{38}$$

holds with probability at least $1-\delta$, where the second line follows from the fact that $\log t/t \le \log 2/2$ for $t \ge 4$.

Finally, for any fixed t obeying $T/\log T \le t \le \log T$, we have

$$\frac{(\log t)^3 \log \frac{|\mathcal{S}||\mathcal{A}|t}{\delta}}{t^{3/5}} \le \frac{(\log T)^{\frac{18}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{3/5}}, \quad \frac{(\log t)^3 \log \frac{|\mathcal{S}||\mathcal{A}|t}{\delta}}{t^{4/5}} \le \frac{(\log T)^{\frac{19}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{4/5}}.$$

The bound in (38) thus translates to

$$|\boldsymbol{\xi}_t| \le 5\sqrt{\frac{(\log T)^{\frac{19}{5}} \log \frac{|\mathcal{A}||\mathcal{S}|T}{\delta}}{c_1 T^{3/5}}\Big(\max_{\lfloor (1-\beta)t\rfloor < i \le t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-1/5}\mathbf{1}\Big)}, \tag{39}$$

as long as t additionally satisfies

$$T^{2/5} \ge \frac{64(\log T)^4}{9c_1}\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}.$$

To see (39), note that under the conditions of Theorem 3.1, one has $T^{2/5} \ge \frac{64(\log T)^4}{9c_1}\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$, hence

$$\begin{aligned}\frac{8(\log T)^{\frac{18}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{3c_1 T^{3/5}}\mathbf{1} &\le \sqrt{\frac{(\log T)^{\frac{19}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{c_1 T^{4/5}}}\mathbf{1} \\ &\le \sqrt{\frac{(\log T)^{\frac{19}{5}} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{c_1 T^{3/5}}\Big(\max_{\lfloor (1-\beta)t\rfloor < i \le t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-1/5}\mathbf{1}\Big)},\end{aligned}$$

which completes the proof of Lemma 4.3. □

B Proof of Theorem 3.3

In this section, we provide the detailed proof of Theorem 3.3. It follows the same idea as that of Theorem 3.1, yet with different analysis on the convergence rates.

Recall that in Theorem 3.3, we set $\gamma_t = 1 - t^{-1/8}$ and

$$\eta_t = \frac{1}{1 + \frac{c_1 t^{5/8}}{(\log t)^2}}, \quad t \ge 2$$

for some constant $c_1 > 0$ in Algorithm 1. In the following, we are to set

$$\beta = \frac{c_2}{T^{1/8}(\log T)^2}$$

for the constant $c_2 > 0$ in Theorem 3.3. In this section, we still use the same notations as in the decomposition of Section 4.1.

B.1 Bound on the switching error

Following exactly the same arguments as in the proof of Lemma 4.1, the switching error can be bounded as

$$\begin{aligned}\left\| \sum_{i=1}^{t} \eta_i^{(t)} \boldsymbol{d}_i / \eta_i \right\|_\infty &\le \sum_{i=1}^{t} \prod_{j=i}^{t} (1-\eta_j) \|\boldsymbol{d}_i\|_\infty \\ &= \sum_{i=1}^{\lfloor (1-\alpha)t \rfloor} \prod_{j=i}^{t} (1-\eta_j) \|\boldsymbol{d}_i\|_\infty + \sum_{1+\lfloor (1-\alpha)t \rfloor}^{t} \prod_{j=i}^{t} (1-\eta_j) \|\boldsymbol{d}_i\|_\infty,\end{aligned}$$

where we choose $\alpha = t^{-3/8}$. Firstly, when t obeys $\lfloor (1-\alpha)t \rfloor = \lfloor t - t^{5/8} \rfloor \ge 50$ (satisfied by $t \ge 100$), for any i such that $1 \le i \le \lfloor (1-\alpha)t \rfloor$, we have

$$\begin{aligned}\prod_{j=i}^{t} (1-\eta_j) &\le \prod_{j=\lfloor (1-\alpha)t \rfloor}^{t} (1-\eta_j) \le \exp\Big(- \textstyle\sum_{j=\lfloor (1-\alpha)t \rfloor}^{t} \eta_j \Big) \\ &\overset{\text{(i)}}{\le} \exp\left(- \sum_{j=\lfloor (1-\alpha)t \rfloor}^{t} \frac{(\log j)^2}{(2+c_1) j^{5/8}} \right) \overset{\text{(ii)}}{\le} \exp\left(- \frac{\alpha t (\log t)^2}{(2+c_1) t^{5/8}} \right),\end{aligned}$$

where (i) follows from (105), and (ii) follows from (106) for $j \ge \lfloor (1-\alpha)t \rfloor \ge 50$. Together with the fact that $\|\boldsymbol{d}_i\|_\infty \le 1$, the above bound implies

$$\sum_{i=1}^{\lfloor (1-\alpha)t \rfloor} \prod_{j=i}^{t} (1-\eta_j) \|\boldsymbol{d}_i\|_\infty \le t \cdot \exp\left(- \frac{\alpha t^{3/8} (\log t)^2}{2+c_1} \right) = \exp\big(\log t - (\log t)^2/(2+c_1) \big). \tag{40}$$

On the other hand, (30) implies

$$\|\boldsymbol{d}_i\|_\infty \le \frac{(i-1)^{-1/8} - i^{-1/5}}{(i-1)^{-1/8}} \le 1 - (1-1/i)^{1/8} \le \frac{1}{4i} \le \frac{1}{4(1-\alpha)t},$$

for all $i \ge 1 + \lfloor (1-\alpha)t \rfloor$. The second inequality above is implied by

$$1 - (1-1/i)^{1/8} = 1 - \exp\left(\frac{\log(1-1/i)}{8} \right) \le 1 - \exp\left(-2/(8i)\right) \le 2/(8i),$$

where for $i \ge 2$, we use the fact that $\log(1-x) \ge -2x$ for $x \le \log 2/2$ and $e^x \ge 1 + x$ for $x \in \mathbb{R}$. Therefore, noting that $\alpha \le 0.25$ for $t \ge 150$, we have

$$\sum_{1+\lfloor (1-\alpha)t \rfloor}^{t} \prod_{j=i}^{t} (1-\eta_j) \|\boldsymbol{d}_i\|_\infty \le \sum_{1+\lfloor (1-\alpha)t \rfloor}^{t} \|\boldsymbol{d}_i\|_\infty \le \frac{\alpha t}{4(1-\alpha)t} \le \alpha/3 = \frac{1}{3t^{3/8}}. \tag{41}$$

Combining (40) and (41), once $T/\log T \ge 100$, we have

$$\begin{aligned}\left\| \sum_{i=1}^{t} \eta_i^{(t)} \boldsymbol{d}_i / \eta_i \right\|_\infty &\le \exp\big(\log t - (\log t)^2/(2+c_1) \big) + \frac{1}{3t^{3/8}} \\ &\le \exp\Big(\log T - \frac{(\log T - \log\log T)^2}{2+c_1} \Big) + \frac{(\log T)^{3/8}}{3T^{3/8}} \\ &\le \exp\Big(\log T - \frac{(\log T)^2}{4(2+c_1)} \Big) + \frac{(\log T)^{3/8}}{3T^{3/8}} \le \frac{2(\log T)^{3/8}}{T^{3/8}}\end{aligned}$$

for all $T/\log T \le t \le T$, where the third inequality follows from $\log T \ge 2 \log\log T$, and the last inequality holds as long as T obeys $(\log T)^3/(8+4c_1) \ge \frac{11}{8} \log T$.

B.2 Bounds on ζ_t

Similar to the proof of Lemma 4.2, the ℓ_∞-norm of $\boldsymbol{\zeta}_t$ can be bounded as

$$\|\boldsymbol{\zeta}_t\|_\infty \le \eta_0^{(t)} + 3\sum_{i=1}^{\lfloor(1-\beta)t\rfloor} \eta_i^{(t)}\gamma_i.$$

By definition of $\{\eta_t\}$, as long as $\beta < 1/2$ and $t \ge 100$,

$$\begin{aligned}\eta_0^{(t)} &= \prod_{j=1}^{t}(1-\eta_j) \le \exp\Big(-\textstyle\sum_{j=1}^{t}\eta_j\Big) \\ &\le \exp\bigg(-\sum_{j=50}^{t}\frac{(\log j)^2}{(2+c_1)j^{5/8}}\bigg) \le \exp\bigg(-\frac{(t-50)(\log t)^2}{(2+c_1)t^{5/8}}\bigg),\end{aligned} \tag{42}$$

where the second inequality follows from (105), and the last inequality follows from the monotonicity (106).

Similarly, when $\lfloor(1-\beta)t\rfloor \ge 50$ (which holds as long as $\beta < 1/2$ and $t \ge 100$), for any $1 \le i \le \lfloor(1-\beta)t\rfloor$, it holds that

$$\eta_i^{(t)} \le \prod_{j=1+\lfloor(1-\beta)t\rfloor}^{t}(1-\eta_j) \le \exp\Big(-\textstyle\sum_{j=1+\lfloor(1-\beta)t\rfloor}^{t}\eta_j\Big) \le \exp\Big(-\frac{\beta t(\log t)^2}{(2+c_1)t^{5/8}}\Big). \tag{43}$$

Combining (42) and (43), once $\beta < 1/2$, we have

$$\begin{aligned}\|\boldsymbol{\zeta}_t\|_\infty &\le \exp\bigg(-\frac{(t-50)(\log t)^2}{(2+c_1)t^{5/8}}\bigg) + 3\exp\bigg(\log t - \frac{\beta t(\log t)^2}{(2+c_1)t^{5/8}}\bigg) \\ &\le \exp\bigg(-\frac{t(\log t)^2}{2(2+c_1)t^{5/8}}\bigg) + 3\exp\bigg(\log t - \frac{\beta t(\log t)^3}{(2+c_1)t^{5/8}}\bigg)\end{aligned}$$

for all $t \ge 100$. Note that when $T/\log T \ge 100$, for all t such that $T/\log T \le t \le T$,

$$t^{3/8}(\log t)^2 \ge (T/\log T)^{3/8}(\log T - \log\log T)^2 \ge (T/\log T)^{3/8}\cdot(\log T)^2/4 \ge T^{3/8}(\log T)^2/4,$$

where the second inequality follows from $\log T \ge 2\log\log T$. Meanwhile, for $T/\log T \le t \le T$,

$$\log t - \frac{\beta t(\log t)^2}{(2+c_1)t^{5/8}} \le \log T - \frac{c_2}{T^{1/8}(\log T)^2}\frac{T^{3/8}}{(\log T)^{3/8}}\frac{(\log T)^2}{4(2+c_1)} \le \log T - \frac{c_2 T^{1/4}}{4(2+c_1)(\log T)^{3/8}}.$$

Therefore, for any $T/\log T \le t \le \log T$,

$$\|\boldsymbol{\zeta}_t\|_\infty \le \exp\Big(-\frac{T^{3/8}(\log T)^2}{2(2+c_1)}\Big) + \exp\Big(\log(3T) - \frac{c_2 T^{1/4}}{4(2+c_1)(\log T)^{3/8}}\Big) \le \frac{4}{T}$$

if T obeys

$$T^{3/8}\log T \ge 2(2+c_1) \quad \text{and} \quad c_2 T^{1/4} \ge 4(2+c_1)(\log T)^{11/8}.$$

B.3 Bound on ξ_t

Similar to the proof of Lemma 4.3, we write

$$\boldsymbol{\xi}_t = \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \boldsymbol{z}_i, \quad \text{where} \quad \boldsymbol{z}_i = \eta_i^{(t)}\gamma_i(\boldsymbol{P}_i - \boldsymbol{P})\boldsymbol{V}_{i-1}. \tag{44}$$

Then (entries of) $\{\boldsymbol{z}_i\}$ are martingale differences, and we shall obtain high-probability bound on $\|\boldsymbol{\xi}_t\|_\infty$ via Freedman's inequality as usual.

Firstly, parallel to (36), when $\lfloor(1-\beta)t\rfloor \geq 50$ and $1-\beta \geq 1/2$,

$$\max_{\lfloor(1-\beta)t\rfloor<i\leq t} \|\boldsymbol{z}_i\|_\infty \leq \max_{\lfloor(1-\beta)t\rfloor<i\leq t} 2\eta_i \leq \frac{4(\log t)^2}{c_1 t^{5/8}} =: R, \tag{45}$$

where the third inequality follows from (107).

We denote the entrywise conditional variance of $\boldsymbol{z}_i$ as $\mathrm{Var}(\boldsymbol{z}_i \,|\, \boldsymbol{V}_{i-1}, \ldots, \boldsymbol{V}_0) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, whose j-th element is the variance of $[z_i]_j$ conditional on $\boldsymbol{V}_{i-1}, \ldots, \boldsymbol{V}_0$. By the definition of $\boldsymbol{z}_i$ in (44),

$$\boldsymbol{W}_t = \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \mathrm{Var}(\boldsymbol{z}_i \,|\, \boldsymbol{V}_{i-1}, \ldots, \boldsymbol{V}_0) = \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \gamma_i^2 \big(\eta_i^{(t)}\big)^2 \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}).$$

Similar to the situation in Lemma 4.3 (c.f. Appendix A.6), we have the crude entrywise upper bounded

$$\begin{aligned}
\boldsymbol{W}_t &\leq \Big(\max_{\lfloor(1-\beta)t\rfloor<i\leq t} \eta_i^{(t)}\Big) \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \gamma_i^2 \eta_i^{(t)} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) \\
&\leq \frac{2(\log t)^2}{c_1 t^{5/8}} \max_{\lfloor(1-\beta)t\rfloor<i\leq t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}),
\end{aligned}$$

where the last inequality follows from (107) and the fact that $\sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \eta_i^{(t)} \leq 1$ due to (101). In addition, by the boundedness of $\|\boldsymbol{V}_{i-1}\|_\infty \leq 1$, we also have a deterministic upper bound of $\boldsymbol{W}_t$ that

$$\|\boldsymbol{W}_t\|_\infty \leq \frac{2(\log t)^2}{c_1 t^{5/8}} =: \sigma^2. \tag{46}$$

To apply Freedman's inequality, we further choose the (smallest) positive integer K such that for some constant $c_2 \geq 1$,

$$\frac{\sigma^2}{2^K} \leq \frac{2(\log t)^2}{c_1 t},$$

which leads to

$$\frac{3\log t}{8\log 2} \leq K \leq \frac{3\log t}{8\log 2} + 1.$$

In view of (45) and (46) together with a union bound over all $|\mathcal{S}||\mathcal{A}|$ entries, Freedman's inequality (Lemma D.6) implies that for any fixed t obeying $\lfloor(1-\beta)t\rfloor \geq 50$ and $\beta < 1/2$,

$$\begin{aligned}
|\boldsymbol{\xi}_t| &\leq \frac{8(\log t)^2}{3c_1 t^{5/8}} \log\frac{2|\mathcal{S}||\mathcal{A}|\log t}{\delta \log 2} \mathbf{1} + \sqrt{8\Big(\boldsymbol{W}_t + \frac{2(\log t)^2}{c_1 t}\mathbf{1}\Big) \log\frac{2|\mathcal{S}||\mathcal{A}|\log t}{\delta\log 2}} \\
&\leq \frac{8(\log t)^2 \log\frac{|\mathcal{S}||\mathcal{A}|t}{\delta}}{3c_1 t^{5/8}} \mathbf{1} + \sqrt{\frac{16(\log t)^2 \log\frac{|\mathcal{S}||\mathcal{A}|t}{\delta}}{c_1 t^{5/8}} \Big(\max_{\lfloor(1-\beta)t\rfloor<i\leq t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + t^{-3/8}\mathbf{1}\Big)}
\end{aligned} \tag{47}$$

holds with probability at least $1-\delta$, where the second line follows from the fact that $\log t/t \leq \log 2/2$ for $t \geq 4$.

Finally, for any fixed t obeying $T/\log T \leq t \leq \log T$, we have

$$\frac{(\log t)^2 \log\frac{|\mathcal{S}||\mathcal{A}|t}{\delta}}{t^{5/8}} \leq \frac{(\log T)^{\frac{21}{8}} \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{5/8}}, \quad \frac{(\log t)^2 \log\frac{|\mathcal{S}||\mathcal{A}|t}{\delta}}{t} \leq \frac{(\log T)^3 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T}.$$

The bound in (47) thus translates to

$$|\boldsymbol{\xi}_t| \leq 5\sqrt{\frac{(\log T)^3 \log\frac{|\mathcal{A}||\mathcal{S}|T}{\delta}}{c_1 T^{5/8}} \Big(\max_{\lfloor(1-\beta)t\rfloor<i\leq t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-3/8}\mathbf{1}\Big)},$$

as long as t additionally satisfies

$$T^{5/8} \geq \frac{64(\log T)^3}{9c_1} \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}.$$

B.4 Recursive bound

Combining the three bounds in preceding subsections, for sufficiently large T satisfying all the mentioned conditions and any fixed t obeying $T/\log T \le t \le T$, it holds with probability at least $1-\delta$ that

$$\begin{aligned}\boldsymbol{\Delta}_t &\le \frac{2(\log T)^{3/8}}{T^{3/8}} + \frac{4}{T} + 5\sqrt{\frac{(\log T)^3 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{c_1 T^{5/8}}\Big(\max_{\lfloor(1-\beta)t\rfloor<i\le t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-3/8}\mathbf{1}\Big)} \\ &\quad + \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \eta_i^{(t)}\gamma_i \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\Delta}_{i-1} \\ &\le \Big(1+\frac{5}{\sqrt{c_1}}\Big)\sqrt{\frac{(\log T)^3 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{5/8}}\Big(\max_{\lfloor(1-\beta)t\rfloor<i\le t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-3/8}\mathbf{1}\Big)} \\ &\quad + \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \eta_i^{(t)}\gamma_i \boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}, \end{aligned} \tag{48}$$

where the second inequality follows from the fact that

$$\frac{2(\log T)^{3/8}}{T^{3/8}} + \frac{4}{T} \le \sqrt{\frac{(\log T)^3 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{5/8}}}$$

for T so large that $T/\log T \ge 100$. Once $(1-\beta) \ge 3/4$, for any fixed t obeying $3T/(2\log T) \le t \le T$, we apply (48) with a union bound over $\{k\colon \frac{2t}{3} \le k \le t\}$ — as a result, with probability at least $1-\delta$, one has

$$\boldsymbol{\Delta}_k \le \sqrt{\boldsymbol{\varphi}_t} + \sum_{i=1+\lfloor(1-\beta)k\rfloor}^{k} \eta_i^{(k)}\gamma_i \boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1}, \qquad \text{for all } \frac{2t}{3} \le k \le t, \tag{49}$$

where we define

$$\begin{aligned}\boldsymbol{\varphi}_t &= 2\Big(1+\frac{5}{\sqrt{c_1}}\Big)^2 \cdot \frac{(\log T)^3 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{5/8}}\Big(\max_{\lfloor(1-\beta)t\rfloor<i\le t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-3/8}\mathbf{1}\Big) \\ &\ge \Big(1+\frac{5}{\sqrt{c_1}}\Big)^2 \cdot \frac{(\log T)^3 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta/T}}{T^{5/8}}\Big(\max_{\lfloor(1-\beta)t\rfloor<i\le t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-3/8}\mathbf{1}\Big). \end{aligned} \tag{50}$$

In the following, we are to follow exactly the same recipe as in the proof of Proposition 4.4 (c.f. Appendix A.3), where we correspondingly define the $\{\alpha_i^{(t)}\}$ according this set of learning rates $\{\eta_j\}$ here. We also set

$$H := T^{1/8}(\log T)^2,$$

so that $(1-\beta)^H \ge \exp(-2\beta H) \ge 2/3$ as long as $c_2 \le \log(3/2)/2$. Thus, when T is sufficiently large so that $1-\beta \ge 2/3$, in parallel with (24), we have

$$\boldsymbol{\Delta}_t \le \max_{(i_1,\ldots,i_H)\in\mathcal{I}_t}\Big\{\underbrace{\Big(\boldsymbol{I} + \sum_{h=1}^{H-1}\prod_{k=1}^{h}(\gamma_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\Big)\sqrt{\boldsymbol{\varphi}_t}}_{=:\boldsymbol{\beta}_1} + \underbrace{\prod_{h=1}^{H}(\gamma_{i_h+1}\boldsymbol{P}^{\pi_{i_h}})|\boldsymbol{\Delta}_{i_H}|}_{=:\boldsymbol{\beta}_2}\Big\}. \tag{51}$$

In the following, we are to bound $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in (51) separately. By definition of the discount factor $\{\gamma_t\}$, we have $\gamma_j \le \gamma_T$ for all $2t/3 \le j \le T$, indicating

$$|\boldsymbol{\beta}_2| \le \gamma_t^H \prod_{h=1}^{H}\boldsymbol{P}^{\pi_{i_h}}|\boldsymbol{\Delta}_H| \le \gamma_T^H \Big\|\prod_{h=1}^{H}\boldsymbol{P}^{\pi_{i_h}}\Big\|_1 \|\boldsymbol{\Delta}_H\|_\infty \overset{\text{(i)}}{\le} \gamma_T^H \overset{\text{(ii)}}{\le} \frac{1}{T}, \tag{52}$$

where (i) follows from the bounded magnitude $\|\boldsymbol{\Delta}_H\|_\infty \le 1$ and the fact that $\prod_{h=1}^H \boldsymbol{P}^{\pi_{i_h}}$ is a probability transition matrix; (ii) follows from

$$\gamma_T^H = \big(1 - T^{-1/8}\big)^{T^{1/8}(\log T)^2} \le \exp\big(-(\log T)^2\big) \le \frac{1}{T}.$$

Again, for $\boldsymbol{\beta}_1$, parallel to (26), we have the entrywise upper bound

$$\begin{aligned}|\boldsymbol{\beta}_1|^2 &\le \frac{1}{1-\gamma_T}\sum_{h=0}^{H-1}\gamma_T^h \prod_{k=1}^h \boldsymbol{P}^{\pi_{i_k}} \frac{2(1+\frac{5}{\sqrt{c_1}})^2(\log T)^3 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta/T}}{T^{5/8}}\Big(\max_{\lfloor(1-\beta)k\rfloor<i\le k}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{i-1}) + T^{-3/8}\mathbf{1}\Big)\\ &\le \frac{4(1+\frac{5}{\sqrt{c_1}})^2(\log T)^3\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/2}}\Big(\sum_{h=0}^{H-1}\gamma_T^h\prod_{k=1}^h \boldsymbol{P}^{\pi_{i_k}}\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i)+\mathbf{1}\Big),\end{aligned} \tag{53}$$

We employ the following lemma to bound the first term of (53); it is parallel to Lemma A.1, and the proof follows exactly the same arguments hence we omit here.

Lemma B.1. *Suppose $T \ge 100$, then it holds for all $T/\log T \le t \le T$ that*

$$\sum_{h=0}^{H-1}\gamma_T^h\prod_{k=1}^h \boldsymbol{P}^{\pi_{i_k}}\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) \le \Big(7 + 72t_{mix}(\log T)^{1/8} + 8T^{1/8}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1}. \tag{54}$$

Invoking Lemma B.1, we obtain an upper bound

$$|\boldsymbol{\beta}_1|^2 \le \frac{4(1+\frac{5}{\sqrt{c_1}})^2(\log T)^3\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/2}}\Big(8 + 72t_{\mathrm{mix}}(\log T)^{1/8} + 8T^{1/8}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1}. \tag{55}$$

Recalling (51), by the upper bound (52) on $|\boldsymbol{\beta}_2|$ and (55) on $|\boldsymbol{\beta}_1|$, for any fixed t such that $3T/(2\log T) \le t \le T$,

$$\begin{aligned}\boldsymbol{\Delta}_t &\le \frac{1}{T} + \sqrt{\frac{4(1+\frac{5}{\sqrt{c_1}})^2(\log T)^3\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/2}}\Big(8 + 72t_{\mathrm{mix}}(\log T)^{1/8} + 8T^{1/8}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1}}\\ &\le 4\Big(1+\frac{5}{\sqrt{c_1}}\Big)\sqrt{\frac{(\log T)^4\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/2}}\Big(8(\log T)^{-1/8} + 72t_{\mathrm{mix}} + 8(T/\log T)^{1/8}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1}}\end{aligned} \tag{56}$$

holds with probability at least $1-\delta$.

We now further take a union bound for (56) over $\{t\colon 3T/(2\log T) \le t \le T\}$, which leads to the simultaneous $(1-\delta)$ high-probability bound

$$\boldsymbol{\Delta}_t \le 16\Big(1+\frac{5}{\sqrt{c_1}}\Big)\sqrt{\frac{(\log T)^4\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/2}}\Big((\log T)^{-1/8} + 9t_{\mathrm{mix}} + (T/\log T)^{1/8}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\Big)\mathbf{1}}$$

for all t such that $3T/(2\log T) \le t \le T$, since $\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta/T} \le 2\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$.

B.5 Solving the recursive bound

We solve the recursive bounds to obtain the final high-probability bound. For any positive integer k, we define

$$u_k = \max\Big\{\|\boldsymbol{\Delta}_i\|_\infty\colon 2^k\frac{3T}{2\log T} \le i \le T\Big\}.$$

A naive upper bound is $u_k \le 1$ for all k. Furthermore, by (15) and the definition of u_k, with probability at least $1-\delta$, it holds simultaneously for all $k \ge 1$ that

$$u_{k+1} \le \sqrt{\frac{c_3(\log T)^4\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/2}}\Big((\log T)^{-1/8} + 9t_{\mathrm{mix}} + (T/\log T)^{1/8}u_k\Big)}$$

for some absolute constant $c_3 > 0$. If there exists some k such that $2^{k+1} \le 2\log T/3$ and

$$u_k \le ((\log T)^{-1/8} + 9t_{\text{mix}})(T/\log T)^{-1/8}, \tag{57}$$

then $2^{k+1}\frac{3T}{2\log T} \le T$, and

$$\|\mathbf{\Delta}_T\|_\infty \le u_{k+1} \le \sqrt{\frac{c_3(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/2}}(2(\log T)^{-1/8} + 18t_{\text{mix}})}.$$

On the other hand, suppose $u_j > ((\log T)^{-1/8}+9t_{\text{mix}})(T/\log T)^{-1/8}$ for any $1 \le j \le k$ for which $2^{k+1} \le 2\log T/3$. Then

$$u_{j+1} \le \sqrt{\frac{2c_3(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{3/8}}}u_j, \quad \text{for all } 1 \le j \le k,$$

which implies

$$\log u_{j+1} \le \frac{1}{2}\log u_j + \frac{1}{2}\log\theta, \quad \text{where } \theta = 2c_3 \cdot T^{-3/8}(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}.$$

Recursively applying this relation for $1 \le j < k$ yields

$$2^k \log u_k \le \log u_0 + \sum_{j=0}^{k-1} 2^j \log\theta \le (2^k - 1)\log\theta,$$

where we used the fact that $u_0 \le 1$. Hence

$$\begin{aligned} u_k \le \theta^{1-1/2^k} &= \Big(2c_3 \cdot T^{-3/8}(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)^{1-1/2^k} \\ &= \Big(2c_3 \cdot (\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)^{1-1/2^k} \cdot T^{-3/8} \cdot T^{\frac{3}{8\cdot 2^k}} \\ &\le 2c_3 \cdot \frac{(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{3/8}} \cdot \exp\big(3\log T \cdot 2^{-k}/8\big). \end{aligned}$$

Now we can set $2^k \ge 3\log T/16$, so that $2^k 3T/(2\log T) \le T$, and the above upper bound translates to

$$\|\mathbf{\Delta}_T\|_\infty \le u_k \le 16c_3 \cdot \frac{(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{3/8}}.$$

Combining the two cases above, we arrive at

$$\|\mathbf{\Delta}_T\|_\infty \le \Big(c_4\sqrt{(\log T)^{-1/8} + 9t_{\text{mix}}} + 16c_3T^{-1/8}\Big)\frac{(\log T)^4 \log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/4}}$$

with probability at least $1-\delta$ for T obeying the conditions of Proposition 4.4.

We now take a moment to collect all the conditions we impose on T, which are

$$\begin{aligned} &(\log T)^3/(8+4c_1) \ge \frac{11}{8}\log T, \quad T/\log T \ge 100, \quad T^{3/8}\log T \ge 2(2+c_1), \\ &c_2T^{1/4} \ge 4(2+c_1)(\log T)^{11/8}, \quad T^{5/8} \ge \frac{64(\log T)^3}{9c_1}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}. \end{aligned}$$

They can be further simplified to

$$\begin{aligned} &(\log T)^2 \ge 11(2+c_1)/2, \quad T/\log T \ge 100, \\ &c_2T^{1/4} \ge 4(2+c_1)(\log T)^{11/8}, \quad T^{5/8} \ge \frac{64(\log T)^3}{9c_1}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}. \end{aligned}$$

B.6 From estimation to policy learning

Finally, we are to leverage the performance difference lemma for discounted reward MDP to obtain the bounds on average reward performance.

Invoking Lemma D.9 with $\gamma = \gamma_T$, $\pi' = \pi^*_{\gamma_T}$ and $\pi = \pi_T$, the greedy policy obtained in the T-th iteration, we know that

$$V^{\pi_T}_{\gamma_T}(s) - V^*_{\gamma_T}(s) \geq -T^{1/8} \cdot \sup_{s' \in \mathcal{S}} \big| Q^*_{\gamma_T}(s', \pi_T(s')) - V^*_{\gamma_T}(s') \big|.$$

Since π_T is the greedy policy with respect to Q_T, for any $s' \in \mathcal{S}$, we ahve

$$\begin{aligned} Q^*_{\gamma_T}(s', \pi_T(s')) &\geq Q_T(s', \pi_T(s')) - \|\boldsymbol{\Delta}_T\|_\infty \\ &\geq Q_T(s', \pi^*_{\gamma_T}(s')) - \|\boldsymbol{\Delta}_T\|_\infty \\ &\geq Q^*_{\gamma_T}(s', \pi^*_{\gamma_T}(s')) - 2\|\boldsymbol{\Delta}_T\|_\infty \geq V^*_{\gamma_T}(s') - 2\|\boldsymbol{\Delta}_T\|_\infty, \end{aligned}$$

and also $Q^*_{\gamma_T}(s', \pi_T(s')) \leq \max_{a \in \mathcal{A}} Q^*_{\gamma_T}(s', a) = V^*_{\gamma_T}(s')$. Therefore, we have $0 \geq V^{\pi_T}_{\gamma_T}(s) - V^*_{\gamma_T}(s) \geq -2\|\boldsymbol{\Delta}_T\|_\infty$. Further invoking Lemma D.1, and recalling that π^* is the optimal policy for average reward, we have

$$\begin{aligned} V^{\pi_T}(s) &\geq V^{\pi_T}_{\gamma_T}(s) - (1-\gamma_T) t_{\mathrm{mix}} \\ &\geq V^*_{\gamma_T}(s) - 2\|\boldsymbol{\Delta}_T\|_\infty - 3(1-\gamma_T) t_{\mathrm{mix}} \\ &\geq V^{\pi^*}_{\gamma_T}(s) - 2\|\boldsymbol{\Delta}_T\|_\infty - 3(1-\gamma_T) t_{\mathrm{mix}} \\ &\geq V^{\pi^*}(s) - 2\|\boldsymbol{\Delta}_T\|_\infty - 6(1-\gamma_T) t_{\mathrm{mix}}, \end{aligned}$$

where the first line follows from Lemma D.1 for π_T, the second line uses the previous result, the third line uses the optimality of $\pi^*_{\gamma_T}$ for γ_T-discounted reward, and the last line uses Lemma D.1 for π^*. Therefore, on the event that $\|\boldsymbol{\Delta}_T\|_\infty \leq \big(c_4\sqrt{(\log T)^{-1/8} + 3t_{\mathrm{mix}}} + 16c_3 T^{-1/8}\big) \frac{(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/4}}$, we have

$$V^*(s) - V^{\pi_T}(s) \leq 2\big(c_4\sqrt{(\log T)^{-1/8} + 3t_{\mathrm{mix}}} + 16c_3 T^{-1/8}\big) \frac{(\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T^{1/8}} + 6T^{-1/8} t_{\mathrm{mix}}.$$

Since $t_{\mathrm{mix}} \geq 1$, the above bound translates to

$$V^*(s) - V^{\pi_T}(s) \leq \frac{c t_{\mathrm{mix}}}{T^{1/8}} (\log T)^4 \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$$

for some absolute constant $c > 0$ that only depends on c_1, c_2. This completes the proof of Theorem 3.3.

C Technical proofs regarding Theorem 5.2

In this section, we provide detailed proofs for results regarding Algorithm 2. Appendix C.1 provides a proof sketch for Theorem 5.2, while the remaining of this section provides detailed proofs for supportive results.

C.1 Sketch of analysis for Algorithm 2

In this section, we provide a sketch of analysis for the estimation error of Algorithm 2, which forms the basis for proving Theorem 5.2. The general approach is similar to Section 4.1 with slightly different bounds. Let

$$\boldsymbol{\delta}_t = \mathbf{q}_t - \mathbf{q}^*_{\alpha_t},$$

for $t \geq 0$, which is the estimation error of Algorithm 2 in the t-th iteration.

Our updating rule (17) in the t-th iteration satisfies

$$\mathbf{q}_t = (1-\theta_t)\mathbf{q}_{t-1} + \theta_t\big[\boldsymbol{r} + \alpha_t \boldsymbol{P}_t \mathbf{v}_{t-1}\big].$$

Employing the Bellman equation $\mathbf{q}^*_{\alpha_t} = \boldsymbol{r} + \alpha_t \boldsymbol{P}\mathbf{v}^*_{\alpha_t}$, we obtain the decomposition

$$\begin{aligned}\boldsymbol{\delta}_t &= \mathbf{q}_t - \mathbf{q}^*_{\alpha_t} = (1-\theta_t)\mathbf{q}_{t-1} + \theta_t\big[\boldsymbol{r} + \alpha_t \boldsymbol{P}_t \mathbf{v}_{t-1}\big] - \mathbf{q}^*_{\alpha_t} \\ &= (1-\theta_t)\boldsymbol{\delta}_{t-1} + (1-\theta_t)(\mathbf{q}^*_{\alpha_{t-1}} - \mathbf{q}^*_{\alpha_t}) + \theta_t\big[\boldsymbol{r} + \alpha_t \boldsymbol{P}_t\mathbf{v}_{t-1} - \mathbf{q}^*_{\alpha_t}\big] \\ &= (1-\theta_t)\boldsymbol{\delta}_{t-1} + (1-\theta_t)(\mathbf{q}^*_{\alpha_{t-1}} - \mathbf{q}^*_{\alpha_t}) + \theta_t\big[\alpha_t \boldsymbol{P}_t\mathbf{v}_{t-1} - \alpha_t\boldsymbol{P}\mathbf{v}^*_{\alpha_t}\big] \\ &= (1-\theta_t)\boldsymbol{\delta}_{t-1} + \theta_t\alpha_t\big[\boldsymbol{P}_t\mathbf{v}_{t-1} - \boldsymbol{P}\mathbf{v}^*_{\alpha_{t-1}}\big] + (1-\theta_t)(\mathbf{q}^*_{\alpha_{t-1}} - \mathbf{q}^*_{\alpha_t}) + \theta_t\boldsymbol{P}(\mathbf{q}^*_{\alpha_{t-1}} - \mathbf{q}^*_{\alpha_t}) \end{aligned} \tag{58}$$

Similar as the arguments in the analysis of Algorithm 1 (see Section 4), we have

$$\boldsymbol{P}_t\mathbf{v}_{t-1} - \boldsymbol{P}\mathbf{v}^*_{\alpha_{t-1}} = (\boldsymbol{P}_t - \boldsymbol{P})\mathbf{v}_{t-1} + \boldsymbol{P}(\mathbf{v}_{t-1} - \mathbf{v}^*_{\alpha_{t-1}}),$$

where the term $\boldsymbol{P}(\mathbf{v}_{t-1} - \mathbf{v}^*_{\alpha_{t-1}})$ is linked to $\boldsymbol{\delta}_{t-1}$ via

$$\boldsymbol{P}(\mathbf{v}_{t-1} - \mathbf{v}^*_{\alpha_{t-1}}) \le \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\delta}_{t-1}, \tag{59a}$$

$$\boldsymbol{P}(\mathbf{v}_{t-1} - \mathbf{v}^*_{\alpha_{t-1}}) \ge \boldsymbol{P}^{\pi^*_{\alpha_{t-1}}}\boldsymbol{\delta}_{t-1}. \tag{59b}$$

Here with some abuse of notations, we define π_t as the greedy policy with respect to q_t. Plugging (59a) and (59b) back into (58) leads to the recursive relation

$$\boldsymbol{\delta}_t \le (1-\theta_t)\boldsymbol{\delta}_t + \theta_t\alpha_t\big[\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\mathbf{v}_{t-1}\big] + \bar{\boldsymbol{d}}_t; \tag{60a}$$

$$\boldsymbol{\delta}_t \ge (1-\theta_t)\boldsymbol{\delta}_t + \theta_t\alpha_t\big[\boldsymbol{P}^{\pi^*_{\alpha_{t-1}}}\boldsymbol{\delta}_{t-1} + (\boldsymbol{P}_t - \boldsymbol{P})\mathbf{v}_{t-1}\big] + \bar{\boldsymbol{d}}_t, \tag{60b}$$

where we define the switching error in the t-th iteration as

$$\bar{\boldsymbol{d}}_t = (1-\theta_t)(\mathbf{q}^*_{\alpha_{t-1}} - \mathbf{q}^*_{\alpha_t}) + \theta_t\boldsymbol{P}(\mathbf{q}^*_{\alpha_{t-1}} - \mathbf{q}^*_{\alpha_t}).$$

Applying (60a) and (60b) recursively, we arrive at

$$\begin{aligned}\boldsymbol{\delta}_t &\le \theta_0^{(t)}\boldsymbol{\delta}_0 + \sum_{i=1}^t \theta_i^{(t)}\boldsymbol{d}_i/\theta_i + \sum_{i=1}^t \theta_i^{(t)}\alpha_i\big[(\boldsymbol{P}_i - \boldsymbol{P})\mathbf{v}_{i-1} + \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\delta}_{i-1}\big], \\ \boldsymbol{\delta}_t &\ge \theta_0^{(t)}\boldsymbol{\delta}_0 + \sum_{i=1}^t \theta_i^{(t)}\boldsymbol{d}_i/\theta_i + \sum_{i=1}^t \theta_i^{(t)}\alpha_i\big[(\boldsymbol{P}_i - \boldsymbol{P})\mathbf{v}_{i-1} + \boldsymbol{P}^{\pi^*_{\alpha_{i-1}}}\boldsymbol{\delta}_{i-1}\big].\end{aligned} \tag{61}$$

where we define $\theta_t^{(t)} = 1$,

$$\theta_0^{(t)} = \textstyle\prod_{j=1}^t(1-\theta_j), \quad \text{and} \quad \theta_i^{(t)} = \theta_i \cdot \textstyle\prod_{j=i+1}^t(1-\theta_j), \quad \forall i \ge 1.$$

Now we set (with some abuse of notation)

$$\beta = \frac{c_2}{T^{1/3}(\log T)^2}$$

for some constant $c_2 > 0$. The upper bound of (61) can be decomposed as

$$\begin{aligned}\boldsymbol{\delta}_t \le\ & \underbrace{\theta_0^{(t)}\boldsymbol{\delta}_0 + \sum_{i=1}^{\lfloor(1-\beta)t\rfloor} \theta_i^{(t)}\alpha_i\big[(\boldsymbol{P}_i - \boldsymbol{P})\mathbf{v}_{i-1} + \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{\delta}_{i-1}\big]}_{=:\boldsymbol{\zeta}_t} \\ &+ \underbrace{\sum_{i=1+\lfloor(1-\beta)t\rfloor}^t \theta_i^{(t)}\alpha_i(\boldsymbol{P}_i - \boldsymbol{P})\mathbf{v}_{i-1}}_{=:\boldsymbol{\xi}_t} + \sum_{i=1+\lfloor(1-\beta)t\rfloor}^t \theta_i^{(t)}\alpha_i\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\delta}_{i-1} + \underbrace{\sum_{i=1}^t \theta_i^{(t)}\bar{\boldsymbol{d}}_i/\theta_i}_{=:\boldsymbol{\omega}_t}\end{aligned}$$

The three terms, similar to before, are bounded as follows.

Lemma C.1. *For all sufficiently large t such that $t \ge \max\{e^{2+c_1}, 100\}$, it holds that*

$$\Big\|\sum_{i=1}^t \theta_i^{(t)}\boldsymbol{d}_i/\theta_i\Big\|_\infty \le 3t^{-2/9}.$$

Proof of Lemma C.1. See Appendix C.5 for a detailed proof. □

Lemma C.2. *Suppose T is sufficiently large such that $c_2 \log T \geq 5(2+c_1)(\log T)^{1/3} + \log\log T$, $c_1 T^3 \geq 3(\log T)^5$ and $(T/\log T)^{1/3} \geq 4(2+c_1)$. Then*

$$\left\| \theta_0^{(t)} \boldsymbol{\delta}_0 + \sum_{i=1}^{\lfloor (1-\beta)t \rfloor} \theta_i^{(t)} \theta_i^{(t)} \alpha_i (\boldsymbol{P}_i - \boldsymbol{P}) \mathbf{v}_{i-1} \right\|_\infty \leq 2/t$$

for any t such that $T/\log T \leq t \leq T$.

Proof of Lemma C.2. See Appendix C.6 for a detailed proof. □

We utilize the Freedman's inequality to control the term $\boldsymbol{\xi}_t$.

Lemma C.3. *When $\beta < 1/2$ and $T/\log T \geq \max\{e^{c_1}, 100\}$, for any fixed t such that $T/\log T \leq t \leq T$,*

$$|\boldsymbol{\xi}_t| \leq \sqrt{\frac{c_3 (\log T)^2 (\log \frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}} \Big(\max_{1+\lfloor (1-\beta)t \rfloor \leq i \leq t} \mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1}) + 2(\log T)^6 \Big)}$$

holds with probability at least $1-\delta$, where $c_3 = 16/c_1 + 144/c_1^2$ is an absolute constant.

Proof of Lemma C.3. See Appendix C.7 for a detailed proof. □

Based on the above three lemmas, we have the following recursive relation, which form the base of proving Theorem 5.2.

Proposition C.4. *Suppose T is sufficiently large such that*

$$\begin{aligned} & T/\log T \geq \max\{e^{2+c_1}, 100, 64(2+c_1)^3\}, \\ & c_3 (\log T)^5 \geq 5, \quad c_1 T^3 \geq 3(\log T)^5, \\ & 2c_3 T \geq (1+3/c_1)^2 (\log T)^2 \\ & c_2 \log T \geq 5(2+c_1)(\log T)^{1/3} + \log\log T \end{aligned}$$

for some constant $c_2 > 0$ and $c_3 = 16/c_1 + 144/c_1^2$. Then with probability at least $1-\delta$,

$$\|\boldsymbol{\delta}_t\|_\infty \leq \sqrt{\frac{4c_3 (\log T)^2 (\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}} \Big(1 + c_4 (\log T)^2 \max_{\lfloor t/2 \rfloor \leq i < t} \|\boldsymbol{\delta}_i\|_\infty \Big)}$$

holds simultaneously for all $3T/(2\log T) \leq t \leq T$, where we define the constant $c_4 = 2(1+3/c_1)$.

Proof of Proposition C.4. See Appendix C.4 for a detailed proof. □

C.2 Proof of Theorem 5.2

Proof of Theorem 5.2. We solve the recursive bound of Propositions C.4 that

$$\|\boldsymbol{\delta}_t\|_\infty \leq \sqrt{\frac{4c_3 (\log T)^2 (\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}} \left(1 + c_4 (\log T)^2 \max_{\lfloor t/2 \rfloor \leq i < t} \|\boldsymbol{\delta}_i\|_\infty \right)} \tag{62}$$

with probability at least $1-\delta$ for all $3T/(2\log T) \leq t \leq T$. For any positive integer k, we define

$$w_k = \max \Big\{ \|\boldsymbol{\delta}_i\|_\infty \colon 2^k \frac{3T}{2\log T} \leq i \leq T \Big\}.$$

By (62) and the definition of w_k, with probability at least $1-\delta$, it holds simultaneously for all $k \geq 1$ that

$$u_{k+1} \leq \sqrt{\frac{4c_3 (\log T)^2 (\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}} \Big(1 + c_4 (\log T)^2 u_k \Big)}.$$

If there exists some k such that $u_k \le 1$ and $2^{k+1} \le 2\log T/3$, then

$$\|\boldsymbol{\delta}_T\|_\infty \le u_{k+1} \le \sqrt{\frac{4c_3(\log T)^2(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}}\big(1+c_4(\log T)^2\big)} \le \frac{c_5(\log T)^2(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{1/9}}$$

for some sufficiently large constant $c_5 > 0$. Otherwise, suppose $u_j > 1$ for any $1 \le j \le k$ where $k \le \lfloor \log_2(2\log T/3)\rfloor - 1$, then

$$u_{j+1} \le \sqrt{\frac{4c_3(1+c_4)(\log T)^4(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}}u_j}, \quad \text{for all } 1 \le j \le k,$$

which implies

$$\log u_{j+1} \le \frac{1}{2}\log u_j + \frac{1}{2}\log\theta, \quad \text{where } \theta = \frac{4c_3(1+c_4)(\log T)^4(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}}.$$

Recursively applying this relation for $1 \le j < k$ yields

$$2^k \log u_k \le \log u_0 + \sum_{j=0}^{k-1} 2^j \log\theta \le \log T + (2^k-1)\log\theta,$$

where we used the fact that $u_0 \le \max_{1\le i\le T}\{\|\mathbf{v}^*_{\alpha_i}\|_\infty + \|\mathbf{v}_i\|_\infty\} \le 3(\log T)^2 T^{1/3}/c_1 + T^{1/9}$, as long as T satisfies $3(\log T)^2T^{1/3}/c_1 + T^{1/9} \le T$. Hence

$$u_k \le \theta^{1-1/2^k}(\log T)^{1/2^k} \le \frac{4c_3(1+c_4)(\log T)^2(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}} \cdot T^{2^{-k+1}},$$

as long as $T^{7/9} \ge 4c_3(1+c_4)(\log T)^4(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2$. Now we can set $k = \log_2(T/6)$, so that $2^k \cdot 3T/\log T < T$ and $T^{2^{-k+1}} \le e^{12}$, hence

$$\|\boldsymbol{\delta}\|_T \le \frac{c_6(\log T)^4(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}}$$

for the constant $c_6 = 4c_3(1+c_4)$. Combining the two cases above, for any fixed T obeying the given conditions,

$$\|\boldsymbol{\delta}\|_T \le \frac{c_5(\log T)^4(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{1/9}}$$

holds with probability at least $1-\delta$ for some constant $c_5 > 0$. Finally, by Lemma 5.1, we have

$$\begin{aligned}\big\|\mathbf{q}_T - \mathbf{q}^* - J^*/(1-\alpha_T)\mathbf{1}\big\|_\infty &\le \|\boldsymbol{\delta}_T\|_\infty + \big\|\mathbf{q}^*_{\alpha_T} - \mathbf{q}^* - J^*/(1-\alpha_T)\mathbf{1}\big| \\ &\le \frac{B + c_5(\log T)^4(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{1/9}}.\end{aligned}$$

Put it another way, given any $\varepsilon > 0$, as long as T satisfies

$$T^{1/9} \ge \frac{B + c_5(\log T)^4(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{\varepsilon},$$

we have $\big\|\mathbf{q}_T - \mathbf{q}^* - J^*/(1-\alpha_T)\mathbf{1}\big\|_\infty \le \varepsilon$. On the same event of probability at least $1-\delta$, we also have

$$\begin{aligned}\big\|\mathbf{v}_T - \mathbf{v}^* - J^*/(1-\alpha_T)\mathbf{1}\big\|_\infty &\le \|\mathbf{v}_T - \mathbf{v}^*_{\alpha_T}\|_\infty + \big\|\mathbf{v}^*_{\alpha_T} - \mathbf{v}^* - J^*/(1-\alpha_T)\mathbf{1}\big| \\ &\le \frac{B + c_5(\log T)^4(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{1/9}}.\end{aligned}$$

We thus have $\big\|\mathbf{v}_T - \mathbf{v}^* - J^*/(1-\alpha_T)\mathbf{1}\big\|_\infty \le \varepsilon$ as well. Still by Lemma 5.1, we also obtain the form of $\mathbf{v}^*$ and $\mathbf{q}^*$, which by standard MDP theory (Puterman, 2014) satisfies the Bellman equation (1). Therefore, we complete the proof of Theorem 5.2. □

C.3 Proof of Corollary 5.3

Proof of Corollary 5.3. Recall that $\pi_T(s) = \operatorname{argmax}_{a\in\mathcal{A}} q_T(s,a)$ for all $s \in \mathcal{S}$ is the greedy policy, and π^* is the optimal policy for average reward. Then under the conditions of Theorem 5.2, we have

$$\big\|q_T(s,a) - q^{\pi^*}(s,a) - J^*/(1-\alpha_T)\big\|_\infty \le \varepsilon, \tag{63}$$

where $\varepsilon = \frac{B+c(\log T)^4(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{1/9}}$, since the function q^* defined in Theorem 5.2 equals q^{π^*}. Assume without loss of generality that π^* and π_T are both deterministic policies. Invoking Lemma D.8 with $\pi' = \pi_T$ and $\pi = \pi^*$, we know that

$$\begin{aligned} J^{\pi_T} - J^{\pi^*} &= \mathbb{E}_{s\sim d_{\pi'}}\Big[\sum_{a\in\mathcal{A}}\big(\pi_T(a\,|\,s) - \pi^*(a\,|\,s)\big)q^*(s,a)\Big] \\ &\ge -\max_{s\in\mathcal{S}}\Big|q^*\big(s,\pi_T(s)\big) - q^*\big(s,\pi^*(s)\big)\Big|. \end{aligned}$$

Here by the optimality of π^* and (63), we have

$$\begin{aligned} q^*\big(s,\pi^*(s)\big) &\ge q^*\big(s,\pi_T(s)\big) \\ &\ge q_T\big(s,\pi_T(s)\big) - J^*/(1-\alpha_T) - \varepsilon \\ &\ge q_T\big(s,\pi^*(s)\big) - J^*/(1-\alpha_T) - \varepsilon \\ &\ge q^*\big(s,\pi^*(s)\big) - 2\varepsilon, \end{aligned}$$

where the second and fourth lines uses the uniform bound (63), and the third line uses the property of greedy policy π_T. This leads to

$$J^{\pi_T} - J^{\pi^*} \ge -2\varepsilon$$

as desired. □

C.4 Proof of Proposition C.4

Proof of Proposition C.4. Putting Lemmas C.1, C.2 and C.3 together, we arrive at

$$\begin{aligned} \boldsymbol{\delta}_t \le \big(3t^{-2/9} + 2/t\big)\mathbf{1} &+ \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t}\theta_i^{(t)}\alpha_i\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\delta}_{i-1} \\ &+ \sqrt{\frac{c_3(\log T)^2(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Big(\max_{1+\lfloor(1-\beta)t\rfloor\le i\le t}\operatorname{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1}) + 2(\log T)^6\mathbf{1}\Big)} \end{aligned}$$

with probability at least $1-\delta$ for any fixed t within $T/\log T \le t \le T$, as long as T satisfies

$$\begin{aligned} T/\log T \ge \max\{e^{2+c_1}, 100\}, \quad (T/\log T)^{1/3} \ge 4(2+c_1), \\ c_2\log T \ge 5(2+c_1)(\log T)^{1/3} + \log\log T, \quad c_1T^3 \ge 3(\log T)^5. \end{aligned}$$

Further simplifying this inequality leads to

$$\begin{aligned} \boldsymbol{\delta}_t \le &\sqrt{\frac{2c_3(\log T)^2(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Big(\max_{1+\lfloor(1-\beta)t\rfloor\le i\le t}\operatorname{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1}) + 2T^{2/9}\mathbf{1}\Big)} \\ &+ \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t}\theta_i^{(t)}\alpha_i\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\delta}_{i-1}, \end{aligned} \tag{64}$$

as long as T additionally satisfy $(\log T)^{50/9} \ge \frac{9}{2c_3}$. Once $(1-\beta) \ge 3/4$, for any fixed t obeying $3T/(2\log T) \le t \le T$, we apply (64) with a union bound over $\{k\colon \frac{2t}{3} \le k \le t\}$ — as a result, with probability at least $1-\delta$,

$$\boldsymbol{\delta}_k \le \sqrt{\boldsymbol{\phi}_t} + \sum_{i=1+\lfloor(1-\beta)k\rfloor}^{k}\theta_i^{(k)}\alpha_i\boldsymbol{P}^{\pi_{i-1}}\boldsymbol{\delta}_{i-1}, \qquad \text{for all } \frac{2t}{3} \le k \le t, \tag{65}$$

where we define

$$\boldsymbol{\phi}_t = \frac{2c_3(\log T)^2(\log \frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Big(\max_{1+\lfloor(1-\beta)k\rfloor\le i\le k}\mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1}) + 2T^{2/9}\mathbf{1}\Big). \tag{66}$$

Following exactly the same arguments as in the proof of Proposition 4.4, we define

$$\lambda_i^{(t)} := \frac{\theta_{i+1}^{(t)}}{\sum_{j=\lfloor(1-\beta)t\rfloor}^{t-1}\theta_{j+1}^{(t)}}, \quad \lfloor(1-\beta)t\rfloor \le i \le t-1, \tag{67}$$

and arrive at the decomposition

$$\begin{aligned}
\boldsymbol{\delta}_t &\le \sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\Big(\lambda_{i_1}^{(t)}\sqrt{\boldsymbol{\phi}_t} + \theta_{i_1+1}^{(t)}\alpha_{i_1+1}\boldsymbol{P}^{\pi_{i_1}}\boldsymbol{\delta}_{i_1}\Big)\\
&\le \sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\Big[\lambda_{i_1}^{(t)}\sqrt{\boldsymbol{\phi}_t} + \theta_{i_1+1}^{(t)}\alpha_{i_1+1}\boldsymbol{P}^{\pi_{i_1}}\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\Big(\lambda_{i_2}^{(i_1)}\sqrt{\boldsymbol{\phi}_t} + \theta_{i_2+1}^{(i_1)}\alpha_{i_2+1}\boldsymbol{P}^{\pi_{i_2}}\boldsymbol{\delta}_{i_2}\Big)\Big]\\
&\le \sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\lambda_{i_1}^{(t)}\sqrt{\boldsymbol{\phi}_t} + \sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\lambda_{i_1}^{(t)}\lambda_{i_2}^{(i_1)}\big(\alpha_{i_1+1}\boldsymbol{P}^{\pi_{i_1}}\big)\sqrt{\boldsymbol{\phi}_t}\\
&\quad + \sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\theta_{i_1+1}^{(t)}\theta_{i_2+1}^{(i_1)}\prod_{k=1}^{2}(\alpha_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\boldsymbol{\delta}_{i_2}\Big)\\
&\le \sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\lambda_{i_1}^{(t)}\lambda_{i_2}^{(i_1)}\{\boldsymbol{I} + \alpha_{i_1+1}\boldsymbol{P}^{\pi_{i_1}}\}\sqrt{\boldsymbol{\phi}_t}\\
&\quad + \sum_{i_1=\lfloor(1-\beta)t\rfloor}^{t-1}\sum_{i_2=\lfloor(1-\beta)i_1\rfloor}^{i_1-1}\theta_{i_1+1}^{(t)}\theta_{i_2+1}^{(i_1)}\prod_{k=1}^{2}(\alpha_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\boldsymbol{\delta}_{i_2}.
\end{aligned} \tag{68}$$

To further apply such recursion, we define

$$G = T^{1/9}\log T, \quad \text{and} \quad \lambda_{\{i_k\}_{k=1}^G} := \lambda_{i_1}^{(t)}\lambda_{i_2}^{(i_1)}\ldots\lambda_{i_G}^{(i_{G-1})} \ge 0$$

for any $t > i_1 > i_2 > \cdots > i_H$, which satisfies

$$\alpha_{\{i_k\}_{k=1}^G} \ge \theta_{i_1+1}^{(t)}\theta_{i_2+1}^{(i_1)}\cdots\theta_{i_G+1}^{(i_{G-1})}.$$

We define the index set

$$\mathcal{I}_t = \big\{(i_1,\ldots,i_G)\colon \lfloor(1-\beta)t\rfloor \le i_1 \le t-1,\ \lfloor(1-\beta)i_{j-1}\rfloor \le i_j \le i_{j-1}-1,\ \forall 1\le j<G\big\},$$

which satisfies

$$\sum_{(i_1,\ldots,i_G)\in\mathcal{I}_t}\lambda_{\{i_k\}_{k=1}^G} = 1.$$

Note that once $T^{2/9}\log T \ge 6$ (implied by a stronger condition $T^{2/9} \ge (\log T)^5$ imposed before), we have

$$(1-\beta)^G = (1-c_2T^{-1/3}(\log T)^{-2})^{T^{1/9}\log T} \ge \exp\big(-2c_2T^{-1/3}(\log T)^{-2}\cdot T^{1/9}\log T\big) \ge \frac{2}{3},$$

which implies

$$i_1 > i_2 > \cdots > i_G \ge 2t/3, \quad \text{for all } (i_1,\ldots,i_G)\in\mathcal{I}_t.$$

Recursively invoking (68), we obtain

$$\begin{aligned}\boldsymbol{\delta}_t &\le \sum_{(i_1,\ldots,i_G)\in\mathcal{I}_t} \lambda_{\{i_k\}_{k=1}^H}\left\{\Big(\boldsymbol{I}+\sum_{h=1}^{G-1}\prod_{k=1}^{h}(\alpha_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\Big)\sqrt{\boldsymbol{\phi}_t}+\prod_{h=1}^{G}(\alpha_{i_h+1}\boldsymbol{P}^{\pi_{i_h}})|\boldsymbol{\delta}_{i_G}|\right\}\\ &\le \max_{(i_1,\ldots,i_G)\in\mathcal{I}_t}\Bigg\{\underbrace{\Big(\boldsymbol{I}+\sum_{h=1}^{G-1}\prod_{k=1}^{G}(\alpha_{i_k+1}\boldsymbol{P}^{\pi_{i_k}})\Big)\sqrt{\boldsymbol{\phi}_t}}_{=:\boldsymbol{\beta}_1}+\underbrace{\prod_{h=1}^{G}(\alpha_{i_h+1}\boldsymbol{P}^{\pi_{i_h}})|\boldsymbol{\delta}_{i_G}|}_{=:\boldsymbol{\beta}_2}\Bigg\}. \end{aligned} \tag{69}$$

We now treat the two terms $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ separately. The easier part is

$$\begin{aligned}\prod_{h=1}^{G}(\alpha_{i_h+1}\boldsymbol{P}^{\pi_{i_h}})|\boldsymbol{\delta}_{i_G}| &\le \prod_{h=1}^{G}\alpha_{i_h+1}\cdot\Big\|\prod_{h=1}^{G}\boldsymbol{P}^{\pi_{i_h}}\Big\|_1\cdot\|\boldsymbol{\delta}_{i_G}\|_\infty\mathbf{1}\\ &\le \alpha_T^G\cdot\max_{2t/3\le j\le t-1}\|\boldsymbol{\delta}_j\|_\infty\mathbf{1}\\ &\le (1-T^{-1/9})^{T^{1/9}\log T}\cdot(T^{1/9}+3T^{1/3}(\log T)^3/c_1)\le(1+3/c_1)T^{-2/3}(\log T)^3,\end{aligned}$$

where the second inequality uses the fact that $\prod_{h=1}^{G}\boldsymbol{P}^{\pi_{i_h}}$ is a probability matrix, and the third inequality uses $\|\boldsymbol{\delta}_j\|_\infty\le\|\mathbf{q}_j\|_\infty+\|\mathbf{q}^*_{\alpha_j}\|_\infty\le 3j^{1/3}(\log j)^2/c_1+j^{1/9}$ by Lemma E.2. On the other hand, following exactly the same argument as 26, we have

$$\begin{aligned}|\boldsymbol{\beta}_1|^2&\le\sum_{h=0}^{G-1}\alpha_T^h\cdot\sum_{h=0}^{G-1}\alpha_T^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\boldsymbol{\phi}_t\\ &\le\frac{1}{1-\alpha_T}\sum_{h=0}^{G-1}\alpha_T^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\frac{2c_3(\log T)^2(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Big(\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1})+2T^{2/9}\mathbf{1}\Big)\\ &\le\frac{2c_3(\log T)^2(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Bigg(2T^{4/9}\mathbf{1}+\sum_{h=0}^{G-1}\alpha_T^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_i)\Bigg).\end{aligned} \tag{70}$$

Applying Lemma D.7, we know that

$$\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_i)\le\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}^*_{\alpha_i})+2(1+3/c_1)(\log T)^2T^{1/3}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\delta}_i\|_\infty,$$

which, by the boundedness that $\|\mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}^*_{\alpha_i})\|_\infty\le\|\mathbf{v}^*_{\alpha_i}\|^2_\infty\le\frac{1}{(1-\alpha_i)^2}\le T^{2/9}$, is further bounded as

$$\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_i)\le T^{2/9}+2(1+3/c_1)(\log T)^2T^{1/3}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\delta}_i\|_\infty.$$

Plug back into (70), we have

$$\begin{aligned}|\boldsymbol{\beta}_1|^2&\le\frac{2c_3(\log T)^2(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Bigg(2T^{4/9}\mathbf{1}+\frac{T^{2/9}}{1-\alpha_T}\mathbf{1}+\frac{2(1+3/c_1)(\log T)^2T^{1/3}}{1-\alpha_T}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\delta}_i\|_\infty\Bigg)\\ &\le\frac{2c_3(\log T)^2(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Bigg(3T^{4/9}\mathbf{1}+c_4(\log T)^2T^{4/9}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\delta}_i\|_\infty\Bigg),\end{aligned} \tag{71}$$

where we define the constant $c_4=2(1+3/c_1)>0$. Therefore, combining with (69), we have

$$\begin{aligned}\boldsymbol{\delta}_t&\le(1+3/c_1)T^{-2/3}(\log T)^3\\ &\quad+\sqrt{\frac{2c_3(\log T)^2(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Bigg(3T^{4/9}\mathbf{1}+c_4(\log T)^2T^{4/9}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\delta}_i\|_\infty\Bigg)}\\ &\le\sqrt{\frac{4c_3(\log T)^2(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Bigg(4T^{4/9}\mathbf{1}+c_4(\log T)^2T^{4/9}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\delta}_i\|_\infty\Bigg)}\end{aligned} \tag{72}$$

$$= \sqrt{\frac{4c_3(\log T)^2(\log \frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/9}}\Big(\mathbf{1} + c_4(\log T)^2 \max_{\lfloor t/2\rfloor \le i<t} \|\boldsymbol{\delta}_i\|_\infty\Big)}$$

that holds with probability at least $1-\delta$ for any fixed t obeying $3T/(2\log T) \le t \le T$, as long as T satisfies $T^{10/9}(\log \frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2 \ge \frac{(1+3/c_1)^2}{2c_3}(\log T)^4$ to ensure (72). Taking a union bound over such t and noting that $\log \frac{T|\mathcal{S}||\mathcal{A}|}{\delta/T} \le 2\log \frac{T|\mathcal{S}||\mathcal{A}|}{\delta}$, we know that

$$\boldsymbol{\delta}_t \le \sqrt{\frac{4c_3(\log T)^2(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}}\Big(1 + c_4(\log T)^2 \max_{\lfloor t/2\rfloor \le i<t} \|\boldsymbol{\delta}_i\|_\infty\Big)}$$

holds simultaneously for all $3T/(2\log T) \le t \le T$ with probability at least $1-\delta/2$. On the other hand, following exactly the same arguments, starting from the lower bound in (61) leads to

$$\boldsymbol{\delta}_t \ge -\sqrt{\frac{4c_3(\log T)^2(\log \frac{T|\mathcal{S}||\mathcal{A}|}{2\delta})^2}{T^{2/9}}\Big(1 + c_4(\log T)^2 \max_{\lfloor t/2\rfloor \le i<t} \|\boldsymbol{\delta}_i\|_\infty\Big)}$$

simultaneously for all $3T/(2\log T) \le t \le T$ with probability at least $1-\delta/2$. Finally, we summarize all the conditions as

$$\begin{aligned} T/\log T \ge \max\{e^{2+c_1}, 100\}, \quad (T/\log T)^{1/3} \ge 4(2+c_1), \\ c_2 \log T \ge 5(2+c_1)(\log T)^{1/3} + \log\log T, \quad c_1 T^3 \ge 3(\log T)^5, \\ (\log T)^{50/9} \ge \frac{9}{2c_3}, \quad T^{10/9}(\log \frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2 \ge \frac{(1+3/c_1)^2}{2c_3}(\log T)^4, \end{aligned}$$

which can be simplified to

$$\begin{aligned} T/\log T \ge \max\{e^{2+c_1}, 100, 64(2+c_1)^3\}, \quad c_3(\log T)^5 \ge 5, \quad c_1 T^3 \ge 3(\log T)^5, \\ c_2 \log T \ge 5(2+c_1)(\log T)^{1/3} + \log\log T, \quad 2c_3 T \ge (1+3/c_1)^2(\log T)^2. \end{aligned}$$

We thus complete the proof of Proposition C.4 by taking a union bound. □

C.5 Proof of Lemma C.1

Proof of Lemma C.1. Firstly, Lemma D.3 implies that each term $\boldsymbol{d}_i$ can be bounded as

$$\begin{aligned} \|\boldsymbol{d}_i\|_\infty &= \big\|(1-\theta_i)(\mathbf{q}^*_{\alpha_{i-1}} - \mathbf{q}^*_{\alpha_i}) + \theta_i \boldsymbol{P}(\mathbf{q}^*_{\alpha_{i-1}} - \mathbf{q}^*_{\alpha_i})\big\|_\infty \\ &\le \|\mathbf{q}^*_{\alpha_{i-1}} - \mathbf{q}^*_{\alpha_i}\|_\infty \le \frac{1}{1-\alpha_i} - \frac{1}{1-\alpha_{i-1}} = i^{1/9} - (i-1)^{1/9} \le (i-1)^{-8/9}/9 \le i^{-8/9}, \end{aligned}$$

where the second last inequality follows from the fact that $x^{1/9} - y^{1/9} = (x-y)z^{-8/9}/9$ for some z lying between x and y.

For any fixed $t \ge 1$, let $\lambda = t^{-1/3}$, then the switching error can be decomposed as

$$\sum_{i=1}^{t} \theta_i^{(t)} \boldsymbol{d}_i/\theta_i = \sum_{i=1}^{\lfloor(1-\lambda)t\rfloor} \theta_i^{(t)} \boldsymbol{d}_i/\theta_i + \sum_{\lfloor(1-\lambda)t\rfloor+1}^{t} \theta_i^{(t)} \boldsymbol{d}_i/\theta_i, \tag{73}$$

where the first summation can be bounded as

$$\sum_{i=1}^{\lfloor(1-\lambda)t\rfloor} \theta_i^{(t)} \boldsymbol{d}_i/\theta_i \le \sum_{i=1}^{\lfloor(1-\lambda)t\rfloor} \theta_i^{(t)} \|\boldsymbol{d}_i\|_\infty/\theta_i \le \sum_{i=1}^{\lfloor(1-\lambda)t\rfloor} \prod_{j=i+1}^{t} (1-\theta_j) \cdot i^{-8/9}, \tag{74}$$

where invoking (109) we know that once $(1-\lambda)t = t - t^{2/3} \ge 50$,

$$\prod_{j=i+1}^{t} (1-\theta_j) \le \exp\Big(-\textstyle\sum_{j=\lfloor(1-\lambda)t\rfloor+1}^{t} \theta_j\Big) \le \exp\Big(-\lambda t \cdot c'(\log t)^2 t^{-2/3}\Big) \le t^{-1}$$

as long as $\log t \geq 1/c' = 2 + c_1$. When $\lambda < 1/2$, the second summation is bounded as

$$\sum_{\lfloor(1-\lambda)t\rfloor+1}^{t} \theta_i^{(t)} \boldsymbol{d}_i/\theta_i \leq \sum_{\lfloor(1-\lambda)t\rfloor+1}^{t} \|\boldsymbol{d}_i\|_\infty \leq \sum_{\lfloor(1-\lambda)t\rfloor+1}^{t} i^{-8/9} \leq \lambda t \cdot (t/2)^{-8/9} \leq 2t^{-2/9}.$$

Putting them together, we arrive at

$$\Big\| \sum_{i=1}^{t} \theta_i^{(t)} \boldsymbol{d}_i/\theta_i \Big\|_\infty \leq 3t^{-2/9}$$

as long as $\log t \geq 2 + c_1$ and $t - t^{2/3} > 50$ (which is satisfied when $t \geq 100$). □

C.6 Proof of Lemma C.2

Proof of Lemma C.2. By the initialization, $\|\boldsymbol{\delta}_0\| \leq 20$. Once $t \geq 50$, by the monotonicity in (109),

$$\begin{aligned}\|\theta_0^{(t)}\boldsymbol{\delta}_0\|_\infty &\leq \prod_{j=1}^{t}(1-\theta_j)\|\boldsymbol{\delta}_0\|_\infty \leq 20\exp\Big(-\textstyle\sum_{j=1}^{t}\theta_j\Big) \leq 20\exp\Big(-\textstyle\sum_{j=50}^{t}\theta_j\Big) \\ &\leq 20\exp\Big(-\textstyle\sum_{j=50}^{t}(\log t)^2 t^{-2/3}/(2+c_1)\Big) \leq 20\exp\Big(-(\log t)^2(t-50)t^{-2/3}/(2+c_1)\Big) \leq t^{-1}\end{aligned}$$

as long as $t^{1/3}\log t \geq 4(2+c_1)$. By the boundedness of $\|\mathbf{v}_i\|_\infty \leq 3(\log i)^2 i^{1/3}/c_1$, invoking (110) we know that when $\lfloor(1-\beta)t\rfloor \geq 50$,

$$\begin{aligned}\Big\| \sum_{i=1}^{\lfloor(1-\beta)t\rfloor} \theta_i^{(t)}\theta_i^{(t)}\alpha_i(\boldsymbol{P}_i - \boldsymbol{P})\mathbf{v}_{i-1} \Big\|_\infty &\leq \sum_{i=1}^{\lfloor(1-\beta)t\rfloor} \theta_i^{(t)}\|\mathbf{v}_{i-1}\|_\infty \\ &\leq t \cdot \max_{0\leq i\leq t}\|\mathbf{v}_i\|_\infty \cdot \exp\Big(-\beta t \cdot \frac{(\log t)^2}{(2+c_2)t^{2/3}}\Big) \\ &\leq 3(\log t)^2 t^{4/3}/c_1 \cdot \exp\Big(-\beta t \cdot \frac{(\log t)^2}{(2+c_2)t^{2/3}}\Big)\end{aligned}$$

Here for any t such that $T/\log T \leq t \leq T$, we have

$$\beta t \cdot \frac{(\log t)^2}{(2+c_1)t^{2/3}} = \frac{c_2 t^{1/3}(\log t)^2}{(2+c_1)T^{1/3}} \geq \frac{c_2(\log t)^2}{(2+c_1)(\log T)^{1/3}} \geq 5\log t,$$

as long as $c_2 \log T \geq 5(2+c_1)(\log T)^{1/3} + \log\log T$, which leads to

$$\Big\| \sum_{i=1}^{\lfloor(1-\beta)t\rfloor} \theta_i^{(t)}\theta_i^{(t)}\alpha_i(\boldsymbol{P}_i - \boldsymbol{P})\mathbf{v}_{i-1} \Big\|_\infty \leq t^{-1}$$

for any t such that $T/\log T \leq t \leq T$, if T satisfies $c_1 T^3 \geq 3(\log T)^5$. We thus complete the proof of Lemma C.2. □

C.7 Proof of Lemma C.3

Proof of Lemma C.3. We write

$$\boldsymbol{x}_i := \theta_i^{(t)}\alpha_i(\boldsymbol{P}_i - \boldsymbol{P})\mathbf{v}_{i-1},$$

so that $\boldsymbol{\xi}_t = \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \boldsymbol{x}_i$, where $\{\boldsymbol{x}_i\}_{i=1+\lfloor(1-\beta)t\rfloor}^{t}$ is a martingale difference squence with the coarse deterministic bound

$$\|\boldsymbol{x}_i\|_\infty \leq \theta_i^{(t)}\alpha_i\|\mathbf{v}_{i-1}\|_\infty \leq \theta_i\|\mathbf{v}_{i-1}\|_\infty \leq \frac{(\log i)^2}{c_1 i^{2/3}} \cdot 3(\log i)^2 i^{1/3}/c_1 \leq 3c_1^{-2}(\log t)^4 t^{-1/3} =: R.$$

as long as $\beta < 1/2$. Furthermore, we define the sum of conditional variances as

$$\begin{aligned}
\boldsymbol{W}_t &= \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \mathrm{Var}(\boldsymbol{x}_i \,|\, \mathbf{v}_1, \ldots, \mathbf{v}_{i-1}) \\
&= \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \big(\theta_i^{(t)}\big)^2 \mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1}) \\
&\le \sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \theta_i^{(t)} \cdot \Big(\max_{1+\lfloor(1-\beta)t\rfloor\le i\le t} \theta_i^{(t)}\Big) \cdot \Big(\max_{1+\lfloor(1-\beta)t\rfloor\le i\le t} \mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1})\Big) \\
&\le \frac{(\log t)^2}{c_1(1-\beta)^{2/3}t^{2/3}} \cdot \max_{1+\lfloor(1-\beta)t\rfloor\le i\le t} \mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1}) \le \frac{2(\log t)^2}{c_1 t^{2/3}} \cdot \max_{1+\lfloor(1-\beta)t\rfloor\le i\le t} \mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1})
\end{aligned}$$

as long as $\beta < 1/2$, where the last line follows from the fact that $\sum_{i=1+\lfloor(1-\beta)t\rfloor}^{t} \theta_i^{(t)} \le \sum_{i=0}^{t} \theta_i^{(t)} = 1$. Since $\|\mathbf{v}_i\|_\infty \le 3(\log i)^2 i^{1/3}/c_1$, we have a coarse deterministic upper bound that

$$\|\boldsymbol{W}_t\|_\infty \le \frac{18(\log t)^2}{c_1^3 t^{2/3}} \cdot (\log t)^4 t^{2/3} = 18(\log t)^6/c_1^3 =: \sigma^2.$$

We now choose the positive integer K such that

$$\frac{(\log t)^2}{c_1 t} \le \frac{\sigma^2}{2^K} \le \frac{2(\log t)^2}{c_1 t}, \quad K \le \frac{\log \frac{2c_1^2 t}{(\log t)^4}}{\log 2} \le 2\log\frac{2t}{(\log t)^2} \le t/2,$$

as long as $t \ge \max\{e^{c_1}, 100\}$. Applying the Freedman's inequality (c.f. Lemma D.6) with a union bound over all $|\mathcal{S}||\mathcal{A}|$ entries, we know that with probability at least $1-\delta$, for any fixed t such that $T/\log T \le t \le T$,

$$\begin{aligned}
|\boldsymbol{\xi}_t| &\le \sqrt{8\max\Big\{\boldsymbol{W}_t, \frac{\sigma^2}{2^K}\Big\}\log\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}} + \frac{4R}{3}\log\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta} \\
&\le \sqrt{8\Big(\frac{2(\log t)^2}{c_1 t^{2/3}} \cdot \max_{1+\lfloor(1-\beta)t\rfloor\le i\le t} \mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1}) + \frac{\sigma^2}{2^K}\Big)\log\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}} + \frac{4R}{3}\log\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta} \\
&\le \sqrt{\frac{16(\log t)^2\log(t|\mathcal{S}||\mathcal{A}|/\delta)}{c_1 t^{2/3}}\Big(\max_{1+\lfloor(1-\beta)t\rfloor\le i\le t} \mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1}) + \frac{1}{t^{1/3}}\Big)} + \frac{12(\log t)^4}{c_1^2 t^{1/3}}\log\frac{t|\mathcal{S}||\mathcal{A}|}{\delta} \\
&\le \sqrt{\frac{c_3(\log T)^2(\log\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})^2}{T^{2/3}}\Big(\max_{1+\lfloor(1-\beta)t\rfloor\le i\le t} \mathrm{Var}_{\boldsymbol{P}}(\mathbf{v}_{i-1}) + 2(\log T)^6\Big)}
\end{aligned}$$

where $c_3 = 32/c_1 + 144/c_1^4$. This completes the proof of Lemma C.3. □

D Supporting lemmas

D.1 Relations of value functions

We quote the following lemma adapted from Jin and Sidford (2021, Lemma 2). It relates the average reward J^π and the rescaled discounted ones Q_γ^π, V_γ^π of any policy π. The proof the lemma is similar to that of Jin and Sidford (2021) and is omitted here.

Lemma D.1 (Lemma 2 of Jin and Sidford (2021))**.** *For any policy $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ and any discount factor $\gamma \in (0,1]$, it holds that $\|J^\pi\mathbf{1} - \boldsymbol{Q}_\gamma^\pi\|_\infty \le 3(1-\gamma)t_{mix}$ and $\|J^\pi\mathbf{1} - \boldsymbol{V}_\gamma^\pi\|_\infty \le (1-\gamma)t_{mix}$.*

The following lemma is adapted from Dong et al. (2021) and De Farias and Van Roy (2006), showing that our framework is applicable when the finite mixing time condition is replaced by the reward averaging time τ (see Remark 2.2). The proof the lemma is similar to that of De Farias and Van Roy (2006) and is omitted here.

Lemma D.2 (Lemma 2 of Dong et al. (2021), Theorem 4.1 of De Farias and Van Roy (2006))**.** *For any policy $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ and any discount factor $\gamma \in (0,1]$, it holds that $\|J^\pi \mathbf{1} - \boldsymbol{Q}_\gamma^\pi\|_\infty \le (1-\gamma)\tau_\pi$ and $\|J^\pi \mathbf{1} - \boldsymbol{V}_\gamma^\pi\|_\infty \le (1-\gamma)\tau_\pi$*

The following lemma bounds the difference between rescaled optimal value and Q-functions with two discount factors.

Lemma D.3. *For any γ_1, γ_2 such that $0 < \gamma_1 < \gamma_2 < 1$, it holds that*

$$\mathbf{0} \le \frac{\boldsymbol{Q}_{\gamma_2}^*}{1-\gamma_2} - \frac{\boldsymbol{Q}_{\gamma_1}^*}{1-\gamma_1} \le \left(\frac{1}{1-\gamma_2} - \frac{1}{1-\gamma_1}\right)\mathbf{1}; \tag{75a}$$

$$\mathbf{0} \le \frac{\boldsymbol{V}_{\gamma_2}^*}{1-\gamma_2} - \frac{\boldsymbol{V}_{\gamma_1}^*}{1-\gamma_1} \le \left(\frac{1}{1-\gamma_2} - \frac{1}{1-\gamma_1}\right)\mathbf{1}, \tag{75b}$$

where $\mathbf{0}$ (resp. $\mathbf{1}$) is a vector or matrix with all entries equal to 0 (resp. 1).

Proof of Lemma D.3. We first show (75a). The upper bound follows directly from Lemma 17 of Dong et al. (2021) with a rescaling. For any policy π, we note that

$$\frac{\boldsymbol{Q}_{\gamma_2}^\pi}{1-\gamma_2} = \mathbb{E}_\pi\left[\sum_{k=1}^\infty \gamma_2^{t-1} r(s_k,a_k) \,\middle|\, s_1 = s\right] \ge \mathbb{E}_\pi\left[\sum_{k=1}^\infty \gamma_1^{t-1} r(s_k,a_k) \,\middle|\, s_1 = s\right] = \frac{\boldsymbol{Q}_{\gamma_1}^\pi}{1-\gamma_1},$$

which follows from $\gamma_2 > \gamma_1$ and the non-negativeness of the reward function r. Applying the above relation to $\pi_{\gamma_1}^*$, we obtain

$$\frac{\boldsymbol{Q}_{\gamma_2}^*}{1-\gamma_2} \ge \frac{\boldsymbol{Q}_{\gamma_2}^{\pi_{\gamma_1}^*}}{1-\gamma_2} \ge \frac{\boldsymbol{Q}_{\gamma_1}^{\pi_{\gamma_1}^*}}{1-\gamma_1} = \frac{\boldsymbol{Q}_{\gamma_1}^*}{1-\gamma_1},$$

which completes the proof of lower bound in (75a). Further noting $V_\gamma^*(s) = \max_{a' \in \mathcal{A}} Q_\gamma^*(s,a')$ for any $s \in \mathcal{S}$, the fact that maximum is a contraction map leads to (75b) and completes the proof of Lemma D.3. □

Lemma D.4. *For any discount factors γ_1, γ_2 such that $0 < \gamma_1 < \gamma_2 < 1$, it holds that*

$$\|\boldsymbol{Q}_{\gamma_1}^* - \boldsymbol{Q}_{\gamma_2}^*\|_\infty \le \frac{\gamma_2 - \gamma_1}{1-\gamma_2} \quad \text{and} \quad \|\boldsymbol{V}_{\gamma_1}^* - \boldsymbol{V}_{\gamma_2}^*\|_\infty \le \frac{\gamma_2 - \gamma_1}{1-\gamma_2}. \tag{76}$$

Proof of Lemma D.4. The boundedness of reward function $\mathbf{0} \le \boldsymbol{r} \le \mathbf{1}$ implies $\mathbf{0} \le \boldsymbol{Q}_{\gamma_i}^* \le \mathbf{1}$ for $i = 1, 2$. For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, applying the lower bound in Lemma D.3 yields

$$Q_{\gamma_2}^*(s,a) \ge \frac{1-\gamma_2}{1-\gamma_1} Q_{\gamma_1}^*(s,a) = Q_{\gamma_1}^*(s,a) + \frac{\gamma_1 - \gamma_2}{1-\gamma_1} Q_{\gamma_1}^*(s,a) \ge Q_{\gamma_1}^*(s,a) - \frac{\gamma_2 - \gamma_1}{1-\gamma_1},$$

where the second inequality follows from $Q_{\gamma_1}^*(s,a) \le 1$. On the other hand, the upper bound in Lemma D.3 together with $Q_{\gamma_1}^*(s,a) \ge 0$ implies

$$Q_{\gamma_2}^*(s,a) \le \frac{1-\gamma_2}{1-\gamma_1} Q_{\gamma_1}^*(s,a) + \frac{\gamma_2 - \gamma_1}{1-\gamma_1} \le Q_{\gamma_1}^*(s,a) + \frac{\gamma_2 - \gamma_1}{1-\gamma_1},$$

which leads to $\|\boldsymbol{Q}_{\gamma_1}^* - \boldsymbol{Q}_{\gamma_2}^*\|_\infty \le \frac{\gamma_2 - \gamma_1}{1-\gamma_2}$. The second inequality in (76) follows from the same arguments. □

Lemma D.5. *For any $i \ge 1$, it holds that*

$$\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) - \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_{\gamma_i}^*) \le 4\|\boldsymbol{\Delta}_i\|_\infty. \tag{77}$$

Proof of Lemma D.5. We start by bounding $\boldsymbol{V}_i - \boldsymbol{V}_{\gamma_i}^*$ for all $i \ge 1$. Recalling that $\boldsymbol{\Delta}_i = \boldsymbol{Q}_i - \boldsymbol{Q}_{\gamma_i}^*$, we have

$$\boldsymbol{V}_i - \boldsymbol{V}_{\gamma_i}^* = \boldsymbol{P}^{\pi_i}\boldsymbol{Q}_i - \boldsymbol{P}^{\pi_{\gamma_i}^*}\boldsymbol{Q}_{\gamma_i}^* \overset{\text{(i)}}{\le} \boldsymbol{P}^{\pi_i}\boldsymbol{Q}_i - \boldsymbol{P}^{\pi_i}\boldsymbol{Q}_{\gamma_i}^* \le \|\boldsymbol{P}^{\pi_i}\|_1 \|\boldsymbol{Q}_i - \boldsymbol{Q}_{\gamma_i}^*\|_\infty \overset{\text{(ii)}}{\le} \|\boldsymbol{\Delta}_i\|_\infty,$$

where (i) follows from the optimality of $\pi^*_{\gamma_i}$ with respect to $\boldsymbol{Q}^*_{\gamma_i}$, and (ii) holds since $\boldsymbol{P}^{\pi_i}$ is a probability transition matrix. Similarly,

$$\boldsymbol{V}_i - \boldsymbol{V}^*_{\gamma_i} \geq \boldsymbol{P}^{\pi^*_{\gamma_i}}\boldsymbol{Q}_i - \boldsymbol{P}^{\pi^*_{\gamma_i}}\boldsymbol{Q}^*_{\gamma_i} \geq -\|\boldsymbol{P}^{\pi^*_{\gamma_i}}\|_1\|\boldsymbol{Q}_i - \boldsymbol{Q}^*_{\gamma_i}\|_\infty \geq -\|\boldsymbol{\Delta}_i\|_\infty,$$

which leads to

$$\|\boldsymbol{V}_i - \boldsymbol{V}^*_{\gamma_i}\|_\infty \leq \|\boldsymbol{\Delta}_i\|_\infty \tag{78}$$

for all $i \geq 1$. By the definition of entrywise variance, we have

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) - \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_i}) &= \big(\boldsymbol{P}(\boldsymbol{V}_i \circ \boldsymbol{V}_i) - (\boldsymbol{P}\boldsymbol{V}_i)\circ(\boldsymbol{P}\boldsymbol{V}_i)\big) - \big(\boldsymbol{P}(\boldsymbol{V}^*_{\gamma_i} \circ \boldsymbol{V}^*_{\gamma_i}) - (\boldsymbol{P}\boldsymbol{V}^*_{\gamma_i})\circ(\boldsymbol{P}\boldsymbol{V}^*_{\gamma_i})\big)\\ &= \boldsymbol{P}(\boldsymbol{V}_i \circ \boldsymbol{V}_i - \boldsymbol{V}^*_{\gamma_i} \circ \boldsymbol{V}^*_{\gamma_i}) + (\boldsymbol{P}\boldsymbol{V}^*_{\gamma_i})\circ(\boldsymbol{P}\boldsymbol{V}^*_{\gamma_i}) - (\boldsymbol{P}\boldsymbol{V}_i)\circ(\boldsymbol{P}\boldsymbol{V}_i)\\ &= \boldsymbol{P}\big((\boldsymbol{V}_i - \boldsymbol{V}^*_{\gamma_i})\circ(\boldsymbol{V}_i + \boldsymbol{V}^*_{\gamma_i})\big) - (\boldsymbol{P}\boldsymbol{V}_i - \boldsymbol{P}\boldsymbol{V}^*_{\gamma_i})\circ(\boldsymbol{P}\boldsymbol{V}_i + \boldsymbol{P}\boldsymbol{V}^*_{\gamma_i}).\end{aligned} \tag{79}$$

Furthermore, we have

$$\begin{aligned}\big\|\boldsymbol{P}\big((\boldsymbol{V}_i - \boldsymbol{V}^*_{\gamma_i})\circ(\boldsymbol{V}_i + \boldsymbol{V}^*_{\gamma_i})\big)\big\|_\infty &\leq \|\boldsymbol{P}\|_1\big\|(\boldsymbol{V}_i - \boldsymbol{V}^*_{\gamma_i})\circ(\boldsymbol{V}_i + \boldsymbol{V}^*_{\gamma_i})\big\|_\infty\\ &\leq \|\boldsymbol{P}\|_1\big\|\boldsymbol{V}_i - \boldsymbol{V}^*_{\gamma_i}\big\|_\infty\|\boldsymbol{V}_i + \boldsymbol{V}^*_{\gamma_i}\big\|_\infty \leq 2\|\boldsymbol{\Delta}_i\|_\infty,\end{aligned}$$

where the last inequality follows from $\|\boldsymbol{P}\|_1 = 1$ and $\|\boldsymbol{V}_i + \boldsymbol{V}^*_{\gamma_i}\big\|_\infty \leq \|\boldsymbol{V}_i\|_\infty + \|\boldsymbol{V}^*_{\gamma_i}\|_\infty \leq 2$ by the bounded magnitude, and (78). Meanwhile, similar arguments yield

$$\begin{aligned}\big\|(\boldsymbol{P}\boldsymbol{V}_i - \boldsymbol{P}\boldsymbol{V}^*_{\gamma_i})\circ(\boldsymbol{P}\boldsymbol{V}_i + \boldsymbol{P}\boldsymbol{V}^*_{\gamma_i})\big\|_\infty &\leq \|\boldsymbol{P}\boldsymbol{V}_i - \boldsymbol{P}\boldsymbol{V}^*_{\gamma_i}\|_\infty\|\boldsymbol{P}\boldsymbol{V}_i + \boldsymbol{P}\boldsymbol{V}^*_{\gamma_i}\|_\infty\\ &\leq \|\boldsymbol{P}\|_1\|\boldsymbol{V}_i - \boldsymbol{V}^*_{\gamma_i}\|_\infty\|\boldsymbol{P}\|_1\|\boldsymbol{V}_i + \boldsymbol{V}^*_{\gamma_i}\|_\infty \leq 2\|\boldsymbol{\Delta}_i\|_\infty.\end{aligned}$$

Plugging the two bounds back in (79) completes the proof of (77). □

D.2 Proof of Lemma 5.1

Proof of Lemma 5.1. By Feinberg and Shwartz (2012, Theorem 8.1), there exists an optimal policy π^* (namely, a Blackwell optimal policy) for the average reward and some $\alpha^* \in (0,1)$, such that for all $\alpha \in (0,\alpha^*)$, π^* is also the optimal policy for α-discounted reward. Since π^* is the optimal policy for average reward, its long-term average reward equals J^*. We define $\mathbf{v}^* \in \mathbb{R}^{|\mathcal{S}|}$ by

$$v^*(s) = \mathbb{E}_{\pi^*}\Big[\sum_{k=0}^{\infty}(r_k - J^*)\,\Big|\, s_0 = s\Big], \quad \forall s \in \mathcal{S},$$

namely, the bias function of π^*. We also define $\boldsymbol{P}^* \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ as the transition matrix induced by π^*, and $\boldsymbol{P}^{**} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{P}^*)^{t-1}$ as the long-term average transition matrix. Furthermore, let

$$H_P = (\boldsymbol{I} - \boldsymbol{P}^* + \boldsymbol{P}^{**})^{-1}(\boldsymbol{I} - \boldsymbol{P}^*) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}.$$

By Puterman (2014, Theorem 8.2.3), letting $\rho = (1-\alpha)/\alpha$, for α that is sufficiently close to 1, or equivalently for ρ that is sufficiently close to 0, we have the Laurent series expansion

$$\mathbf{v}^{\pi^*}_\alpha = (1+\rho)\Big[\rho^{-1}J^*\mathbf{1} + \mathbf{v}^* + \sum_{n=1}^{\infty}\rho^n\mathbf{y}_n\Big],$$

where J^* is the long-term average reward of π^*, $\mathbf{y}_n = (-1)^n H_P^{n+1}\mathbf{f}$, where $\mathbf{f} \in \mathbb{R}^{|\mathcal{S}|}$ is the vectorization of $f(s) = r(s,\pi^*(s))$ that also satisfies $\mathbf{v}^* = H_P\mathbf{f}$. Thus, recalling the Blackwell optimality of π^* and taking α sufficiently close to 1, we have

$$\mathbf{v}^*_\alpha = \mathbf{v}^{\pi^*}_\alpha = (1+\rho)\Big[\rho^{-1}J^*\mathbf{1} + \mathbf{v}^* + \sum_{n=1}^{\infty}\rho^n\mathbf{y}_n\Big].$$

It is clear to see $\lim_{\rho\to 0}\sum_{n=0}^{\infty}\rho^n\|\mathbf{y}_{n+1}\|_\infty<\infty$ from the definition of $\mathbf{y}_n$. Thus, since $\rho=(1-\alpha)/\alpha$,

$$\begin{aligned}\sum_{n=1}^{\infty}\rho^n\mathbf{y}_n &= \alpha\frac{\mathbf{v}_\alpha^*-\frac{J^*}{1-\alpha}\mathbf{1}-\frac{\mathbf{v}^*}{\alpha}}{1-\alpha}\\ &= \alpha\frac{\mathbf{v}_\alpha^*-\frac{J^*}{1-\alpha}\mathbf{1}-\mathbf{v}^*}{1-\alpha}-\mathbf{v}^*,\end{aligned}$$

which means

$$\lim_{\alpha\to 1}\frac{\left\|\mathbf{v}_\alpha^*-\frac{J^*}{1-\alpha}\mathbf{1}-\mathbf{v}^*\right\|_\infty}{1-\alpha}<\infty.$$

Therefore, there exists a constant $B_1>0$ which only depends on the underlying MDP that

$$\sup_{\alpha\in(0,1)}\frac{\left\|\mathbf{v}_\alpha^*-\frac{J^*}{1-\alpha}\mathbf{1}-\mathbf{v}^*\right\|_\infty}{1-\alpha}\le B_1. \tag{80}$$

We now define the function

$$q^*(s,a)=r(s,a)-J^*+\mathbb{E}_{s'\sim P(\cdot\,|\,s,a)}\left[v^*(s')\right]$$

for all state-action pairs $(s,a)\in\mathcal{S}\times\mathcal{A}$, so that (v^*,q^*) satisfies the Bellman equation (1) by standard MDP theory (Puterman, 2014). The equivalent vectorized representation is

$$\mathbf{q}^*=\boldsymbol{r}-J^*\mathbf{1}+\boldsymbol{P}^{\pi^*}\mathbf{v}^*.$$

Then the Bellman equation for optimal discounted value functions implies

$$\begin{aligned}\mathbf{q}_\alpha^* &= \boldsymbol{r}+\alpha\boldsymbol{P}^{\pi_\alpha^*}\mathbf{v}_\alpha^*\\ &\ge \boldsymbol{r}+\alpha\boldsymbol{P}^{\pi^*}\mathbf{v}_\alpha^*\\ &\ge \boldsymbol{r}+\alpha\boldsymbol{P}^{\pi^*}\big(\tfrac{J^*}{1-\alpha}\mathbf{1}+\mathbf{v}^*-(1-\alpha)B_1\mathbf{1}\big)\\ &\ge \boldsymbol{r}+\boldsymbol{P}^{\pi^*}\mathbf{v}^*+\big(\tfrac{\alpha}{1-\alpha}J^*-(1-\alpha)\|\mathbf{v}^*\|_\infty-\alpha(1-\alpha)B_1\big)\mathbf{1}\\ &= \mathbf{q}^*+J^*\mathbf{1}+\big(\tfrac{\alpha}{1-\alpha}J^*-(1-\alpha)\|\mathbf{v}^*\|_\infty-\alpha(1-\alpha)B_1\big)\mathbf{1}\\ &= \mathbf{q}^*+\big(\tfrac{J^*}{1-\alpha}-(1-\alpha)\|\mathbf{v}^*\|_\infty-\alpha(1-\alpha)B_1\big)\mathbf{1},\end{aligned}$$

where the second line follows from the optimality of π_α^* with respect to $\mathbf{v}_\alpha^*$, the third line follows from (80), and the fifth line follows from the Bellman equation (1) for (q^*,v^*). Similarly, we have

$$\begin{aligned}\mathbf{q}_\alpha^* &= \boldsymbol{r}+\alpha\boldsymbol{P}^{\pi_\alpha^*}\mathbf{v}_\alpha^*\\ &\le \boldsymbol{r}+\alpha\boldsymbol{P}^{\pi_\alpha^*}\big(\tfrac{J^*}{1-\alpha}\mathbf{1}+\mathbf{v}^*+(1-\alpha)B_1\mathbf{1}\big)\\ &\le \boldsymbol{r}+\alpha\boldsymbol{P}^{\pi^*}\mathbf{v}^*+\big(\tfrac{\alpha}{1-\alpha}J^*+\alpha(1-\alpha)B_1\big)\mathbf{1}\\ &\le \mathbf{q}^*+J^*\mathbf{1}+\big(\tfrac{\alpha}{1-\alpha}J^*+(1-\alpha)\|\mathbf{v}^*\|_\infty+\alpha(1-\alpha)B_1\big)\mathbf{1}\\ &= \mathbf{q}^*+\big(\tfrac{J^*}{1-\alpha}-(1-\alpha)\|\mathbf{v}^*\|_\infty-\alpha(1-\alpha)B_1\big)\mathbf{1}.\end{aligned}$$

Combining the above two inequalities, we know that there exists some constant $B_2>0$ such that

$$\sup_{\alpha\in(0,1)}\frac{\left\|\mathbf{q}_\alpha^*-\frac{J^*}{1-\alpha}\mathbf{1}-\mathbf{q}^*\right\|_\infty}{1-\alpha}\le B_2.$$

Taking $B=\max\{B_1,B_2\}$ completes the proof of Lemma 5.1. □

D.3 Freedman's inequality

The following lemma is adapted by Li et al. (2021) from Freedman's inequality Freedman (1975).

Lemma D.6 (Theorem 5 of Li et al. (2021)). *Suppose $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

$$|X_k| \le R \quad \text{and} \quad \mathbb{E}\big[X_k \,\big|\, \{X_j\}_{j<k}\big] = 0, \quad \forall\, k \ge 1.$$

Define

$$W_n := \sum_{k=1}^n \mathbb{E}_{k-1}[X_k^2],$$

where we write $\mathbb{E}_{k-1}$ for the expectation conditional on $\{X_j\}_{j<k}$. Then for any given $\sigma^2 \ge 0$, one has

$$\mathbb{P}\big(|Y_n| \ge y \text{ and } W_n \le \sigma^2\big) \le 2\exp\bigg(-\frac{y^2/2}{\sigma^2 + Ry/3}\bigg).$$

In addition, suppose $W_n \le \sigma^2$ holds deterministically. For any positive integer K, with probability at least $1-\delta$ one has

$$|Y_n| \le \sqrt{8\max\Big\{W_n, \frac{\sigma^2}{2^K}\Big\}\log\frac{2K}{\delta}} + \frac{4R}{3}\log\frac{2K}{\delta}. \tag{81}$$

D.4 Proof of Lemma A.1

Proof of Lemma A.1. Employing (77) of Lemma D.5, we have the entrywise bound

$$\begin{aligned}
&\sum_{h=0}^{H-1}\gamma_T^h \prod_{k=1}^h \boldsymbol{P}^{\pi_{i_k}} \max_{\lfloor t/2\rfloor \le i < t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) \\
&\le \sum_{h=0}^{H-1}\gamma_T^h \prod_{k=1}^h \boldsymbol{P}^{\pi_{i_k}} \max_{\lfloor t/2\rfloor \le i < t} \big\{\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_i}) + \|\boldsymbol{\Delta}_i\|_\infty \mathbf{1}\big\} \\
&\le \sum_{h=0}^{H-1}\gamma_T^h \prod_{k=1}^h \boldsymbol{P}^{\pi_{i_k}} \max_{\lfloor t/2\rfloor \le i < t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_i}) + 4\sum_{h=0}^{H-1}\gamma_T^h \prod_{k=1}^h \boldsymbol{P}^{\pi_{i_k}} \max_{\lfloor t/2\rfloor \le i < t} \|\boldsymbol{\Delta}_i\|_\infty \mathbf{1} \\
&= \sum_{h=0}^{H-1}\gamma_T^h \prod_{k=1}^h \boldsymbol{P}^{\pi_{i_k}} \max_{\lfloor t/2\rfloor \le i < t} \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_i}) + \frac{4}{1-\gamma_T} \max_{\lfloor t/2\rfloor \le i < t} \|\boldsymbol{\Delta}_i\|_\infty \mathbf{1},
\end{aligned} \tag{82}$$

where the third line follows from $\max_i(\boldsymbol{a}_i + \boldsymbol{b}_i) \le \max_i \boldsymbol{a}_i + \max_i \boldsymbol{b}_i$, and the last line follows from the fact that $\prod_{k=1}^h \boldsymbol{P}^{\pi_{i_k}}\mathbf{1} = \mathbf{1}$ since the product is still a probability transition matrix. Following the same arguments as the way we bound (79), it holds for all $i \ge 1$ that

$$\big\|\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_i}) - \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_t})\big\|_\infty \le 4\big\|\boldsymbol{Q}^*_{\gamma_i} - \boldsymbol{Q}^*_{\gamma_t}\|_\infty. \tag{83}$$

On the other hand,

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_t}) &= \boldsymbol{P}(\boldsymbol{V}^*_{\gamma_t} \circ \boldsymbol{V}^*_{\gamma_t}) - (\boldsymbol{P}\boldsymbol{V}^*_{\gamma_t}) \circ (\boldsymbol{P}\boldsymbol{V}^*_{\gamma_t}) \\
&= \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^*_{\gamma_t} \circ \boldsymbol{Q}^*_{\gamma_t}) + \boldsymbol{P}(\boldsymbol{V}^*_{\gamma_t} \circ \boldsymbol{V}^*_{\gamma_t}) - \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^*_{\gamma_t} \circ \boldsymbol{Q}^*_{\gamma_t}) - (\boldsymbol{P}\boldsymbol{V}^*_{\gamma_t}) \circ (\boldsymbol{P}\boldsymbol{V}^*_{\gamma_t}) \\
&= \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^*_{\gamma_t} \circ \boldsymbol{Q}^*_{\gamma_t}) + \big(\boldsymbol{P}^{\pi^*_{\gamma_t}}(\boldsymbol{Q}^*_{\gamma_t} \circ \boldsymbol{Q}^*_{\gamma_t}) - \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^*_{\gamma_t} \circ \boldsymbol{Q}^*_{\gamma_t})\big) \\
&\quad - \frac{1}{\gamma_t^2}\big(\boldsymbol{Q}^*_{\gamma_t} - (1-\gamma_t)\boldsymbol{r}\big) \circ \big(\boldsymbol{Q}^*_{\gamma_t} - (1-\gamma_t)\boldsymbol{r}\big),
\end{aligned} \tag{84}$$

where the last equality follows from the Bellman equation $\boldsymbol{Q}^*_\gamma = (1-\gamma)\boldsymbol{r} + \gamma\boldsymbol{P}\boldsymbol{V}^*_\gamma$ for all $\gamma \in (0,1)$. Using similar arguments as Li et al. (2021), the second term in (84) can be bounded as

$$\big\|\boldsymbol{P}^{\pi^*_{\gamma_t}}(\boldsymbol{Q}^*_{\gamma_t} \circ \boldsymbol{Q}^*_{\gamma_t}) - \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^*_{\gamma_t} \circ \boldsymbol{Q}^*_{\gamma_t})\big\|_\infty \tag{85}$$

$$
\begin{aligned}
&= \big\|\boldsymbol{P}\boldsymbol{\Pi}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{Q}^*_{\gamma_t}) - \boldsymbol{P}\boldsymbol{\Pi}^{\pi^*_{\gamma_t}}(\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{Q}^*_{\gamma_t})\big\|_\infty && (86)\\
&\le \|\boldsymbol{P}\|_1\big\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{Q}^*_{\gamma_t}) - \boldsymbol{P}\boldsymbol{\Pi}^{\pi^*_{\gamma_t}}(\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{Q}^*_{\gamma_t})\big\|_\infty && (87)\\
&\overset{\text{(i)}}{\le} \big\|(\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{\Pi}^{\pi^*_{\gamma_t}}\boldsymbol{Q}^*_{\gamma_t})\circ(\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}^*_{\gamma_t} + \boldsymbol{\Pi}^{\pi^*_{\gamma_t}}\boldsymbol{Q}^*_{\gamma_t})\big\|_\infty && (88)\\
&\le \big\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{V}^*_{\gamma_t}\big\|_\infty\big\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}^*_{\gamma_t} + \boldsymbol{\Pi}^{\pi^*_{\gamma_t}}\boldsymbol{Q}^*_{\gamma_t}\big\|_\infty && (89)\\
&\overset{\text{(ii)}}{\le} 2\big\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}_{i_{h+1}}\big\|_\infty + 2\big\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}_{i_{h+1}} - \boldsymbol{V}^*_{\gamma_t}\big\|_\infty && (90)\\
&\overset{\text{(iii)}}{\le} 2\big\|\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}^*_{i_{h+1}}\big\|_\infty + 2\big\|\boldsymbol{V}_{i_{h+1}} - \boldsymbol{V}^*_{\gamma_t}\big\|_\infty && (91)\\
&\overset{\text{(iv)}}{\le} 4\max_{\lfloor t/2\rfloor\le i<t}\big\|\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}_i\big\|_\infty && (92)\\
&\overset{\text{(v)}}{\le} 4\max_{\lfloor t/2\rfloor\le i<t}\big\|\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}^*_{\gamma_i}\big\|_\infty + 4\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty, && (93)
\end{aligned}
$$

where (i) follows from the fact that $\|\boldsymbol{P}\|_1 = 1$, (ii) is due to the boundedness of $\|\boldsymbol{Q}^*_{\gamma_t}\| \le 1$, (iii) follows from $\|\boldsymbol{\Pi}^{\pi_{i_{h+1}}}\|_1 = 1$ and $\boldsymbol{V}_{i_{h+1}} = \boldsymbol{\Pi}^{\pi_{i_{h+1}}}\boldsymbol{Q}_{i_{h+1}}$, (iv) follows from $\big\|\boldsymbol{V}_{i_{h+1}} - \boldsymbol{V}^*_{\gamma_t}\big\|_\infty \le \big\|\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}^*_{i_{h+1}}\big\|_\infty$, and (v) follows from the property of entrywise maximum. We thus obtain the entrywise upper bound (uniform over all $\lfloor t/2\rfloor \le i < t$)

$$
\begin{aligned}
\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_i}) &\le \mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_t}) + 4\max_{\lfloor t/2\rfloor\le i<t}\big\|\boldsymbol{Q}^*_{\gamma_i} - \boldsymbol{Q}^*_{\gamma_t}\|_\infty\mathbf{1}\\
&\le \boldsymbol{P}^{\pi_{i_{h+1}}}(\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{Q}^*_{\gamma_t}) + 8\max_{\lfloor t/2\rfloor\le i<t}\big\|\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}^*_{\gamma_i}\big\|_\infty\mathbf{1} + 4\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\mathbf{1}\\
&\quad - \frac{1}{\gamma_t^2}\big(\boldsymbol{Q}^*_{\gamma_t} - (1-\gamma_t)\boldsymbol{r}\big)\circ\big(\boldsymbol{Q}^*_{\gamma_t} - (1-\gamma_t)\boldsymbol{r}\big)\\
&\le \frac{1}{\gamma_t}\big(\gamma_t\boldsymbol{P}^{\pi_{i_{h+1}}} - \boldsymbol{I}\big)(\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{Q}^*_{\gamma_t}) + 8\max_{\lfloor t/2\rfloor\le i<t}\big\|\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}^*_{\gamma_i}\big\|_\infty\mathbf{1} + 4\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\mathbf{1}\\
&\quad + \frac{2(1-\gamma_t)}{\gamma_t^2}\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{r} - \frac{(1-\gamma_t)^2}{\gamma_t^2}\boldsymbol{r}\circ\boldsymbol{r}\\
&\le \frac{1}{\gamma_T}\big(\gamma_T\boldsymbol{P}^{\pi_{i_{h+1}}} - \boldsymbol{I}\big)(\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{Q}^*_{\gamma_t}) + 8\max_{\lfloor t/2\rfloor\le i<t}\big\|\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}^*_{\gamma_i}\big\|_\infty\mathbf{1} + 4\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\mathbf{1}\\
&\quad + \frac{2(1-\gamma_t)}{\gamma_t^2}\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{r} - \frac{(1-\gamma_t)^2}{\gamma_t^2}\boldsymbol{r}\circ\boldsymbol{r}, \qquad (94)
\end{aligned}
$$

where the first inequality follows from (83) and the second inequality follows from (84) and (93), and the last inequality follows from the monotonicity of γ_t. Applying (94) to (82) yields the telescoping sum

$$
\begin{aligned}
&\sum_{h=0}^{H-1}\gamma_T^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i)\\
&\le \sum_{h=0}^{H-1}\gamma_T^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\max_{\lfloor t/2\rfloor\le i<t}\mathrm{Var}_{\boldsymbol{P}}(\boldsymbol{V}^*_{\gamma_i}) + \frac{4}{1-\gamma_T}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\mathbf{1}\\
&\le \frac{1}{\gamma_T}\sum_{h=0}^{H-1}\gamma_T^h\prod_{k=1}^{h}\boldsymbol{P}^{\pi_{i_k}}\big(\gamma_T\boldsymbol{P}^{\pi_{i_{h+1}}} - \boldsymbol{I}\big)(\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{Q}^*_{\gamma_t})\\
&\quad + \Big(\frac{8}{1-\gamma_t}\max_{\lfloor t/2\rfloor\le i<t}\big\|\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}^*_{\gamma_i}\big\|_\infty + \frac{8}{1-\gamma_t}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty + \frac{2(1-\gamma_t)}{\gamma_t^2(1-\gamma_T)}\|\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{r}\|_\infty\Big)\mathbf{1}\\
&= \frac{1}{\gamma_T}\Big(\gamma_T^H\prod_{k=1}^{H}\boldsymbol{P}^{\pi_{i_{h+1}}} - \boldsymbol{I}\Big)(\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{Q}^*_{\gamma_t})\\
&\quad + \Big(\frac{8}{1-\gamma_T}\max_{\lfloor t/2\rfloor\le i<t}\big\|\boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}^*_{\gamma_i}\big\|_\infty + \frac{8}{1-\gamma_T}\max_{\lfloor t/2\rfloor\le i<t}\|\boldsymbol{\Delta}_i\|_\infty + \frac{2(1-\gamma_t)}{\gamma_t^2(1-\gamma_T)}\|\boldsymbol{Q}^*_{\gamma_t}\circ\boldsymbol{r}\|_\infty\Big)\mathbf{1},
\end{aligned}
$$

where the first inequality follows from (82), and the second inequality uses $\prod_{k=1}^{h} \boldsymbol{P}^{\pi_{i_k}} \mathbf{1} = \mathbf{1}$ as the product is a probability transition matrix. As the rows of the probability transition matrix $\prod_{k=1}^{H} \boldsymbol{P}^{\pi_{i_{h+1}}}$ all sum up to one, we have

$$\Big\| \frac{1}{\gamma_T} \Big(\gamma_T^H \prod_{k=1}^{H} \boldsymbol{P}^{\pi_{i_{h+1}}} - \boldsymbol{I} \Big) (\boldsymbol{Q}^*_{\gamma_t} \circ \boldsymbol{Q}^*_{\gamma_t}) \Big\|_\infty \le \frac{2}{\gamma_T} \|\boldsymbol{Q}^*_{\gamma_t}\|_\infty^2 \le \frac{2}{\gamma_T} \le 4$$

by the boundedness of $\|\boldsymbol{Q}^*_\gamma\|_\infty \le 1$ for all $\gamma \in (0,1)$, as well as the fact that $\gamma_t \ge 1/2$ for $t \ge 160$. On the other hand, for any $\lfloor t/2 \rfloor \le i < t$, invoking Lemma D.1, we obtain

$$\begin{aligned} \big\| \boldsymbol{Q}^*_{\gamma_t} - \boldsymbol{Q}^*_{\gamma_i} \big\|_\infty &\le \big\| \boldsymbol{Q}^*_{\gamma_t} - J^* \mathbf{1} \big\|_\infty + \big\| J^* \mathbf{1} - \boldsymbol{Q}^*_{\gamma_i} \big\|_\infty \\ &\le 3(1-\gamma_t) t_{\mathrm{mix}} + 3(1-\gamma_i) t_{\mathrm{mix}} \le 3\big(1 + 2^{1/5}\big) t^{-1/5} t_{\mathrm{mix}} \le 9 T^{-1/5} (\log T)^{1/5} t_{\mathrm{mix}} \end{aligned}$$

for $T/\log T \le t \le \log T$. In addition, for $T \ge 160$, we also have

$$\frac{2(1-\gamma_t)}{\gamma_t^2 (1-\gamma_T)} \|\boldsymbol{Q}^*_{\gamma_t} \circ \boldsymbol{r}\|_\infty \le 3 \|\boldsymbol{Q}^*_{\gamma_t}\|_\infty \|\boldsymbol{r}\|_\infty \le 3,$$

which, together with the above pieces, leads to (27) and completes the proof of Lemma A.1. □

Lemma D.7. *For any $i \ge 1$, it holds that*

$$\big\| \operatorname{Var}_{\boldsymbol{P}}(\mathbf{v}_i) - \operatorname{Var}_{\boldsymbol{P}}(\mathbf{v}^*_{\alpha_i}) \big\|_\infty \le 3 (\log i)^2 i^{1/3} \|\boldsymbol{\delta}_i\|_\infty. \tag{95}$$

Proof of Lemma D.7. We first bound $\mathbf{v}_i - \mathbf{v}^*_{\alpha_i}$ for all $i \ge 1$. Recalling that $\boldsymbol{\delta}_i = \mathbf{q}_i - \mathbf{q}^*_{\alpha_i}$, we have

$$\mathbf{v}_i - \mathbf{v}^*_{\alpha_i} = \boldsymbol{P}^{\pi_i} \mathbf{q}_i - \boldsymbol{P}^{\pi^*_{\alpha_i}} \mathbf{q}^*_{\alpha_i} \overset{\text{(i)}}{\le} \boldsymbol{P}^{\pi_i} \mathbf{q}_i - \boldsymbol{P}^{\pi_i} \mathbf{q}^*_{\alpha_i} \le \|\boldsymbol{P}^{\pi_i}\|_1 \|\mathbf{q}_i - \mathbf{q}^*_{\alpha_i}\|_\infty \overset{\text{(ii)}}{\le} \|\boldsymbol{\delta}_i\|_\infty,$$

where (i) follows from the optimality of $\pi^*_{\alpha_i}$ with respect to $\mathbf{q}^*_{\alpha_i}$, and (ii) holds since $\boldsymbol{P}^{\pi_i}$ is a probability transition matrix. Similarly,

$$\mathbf{v}_i - \mathbf{v}^*_{\alpha_i} \ge \boldsymbol{P}^{\pi^*_{\alpha_i}} \mathbf{q}_i - \boldsymbol{P}^{\pi^*_{\alpha_i}} \mathbf{q}^*_{\alpha_i} \ge -\|\boldsymbol{P}^{\pi^*_{\alpha_i}}\|_1 \|\mathbf{q}_i - \mathbf{q}^*_{\alpha_i}\|_\infty \ge -\|\boldsymbol{\delta}_i\|_\infty,$$

which leads to

$$\|\mathbf{v}_i - \mathbf{v}^*_{\alpha_i}\|_\infty \le \|\boldsymbol{\delta}_i\|_\infty \tag{96}$$

for all $i \ge 1$. By the definition of entrywise variance, we have

$$\begin{aligned} \operatorname{Var}_{\boldsymbol{P}}(\mathbf{v}_i) - \operatorname{Var}_{\boldsymbol{P}}(\mathbf{v}^*_{\alpha_i}) &= \big(\boldsymbol{P}(\mathbf{v}_i \circ \mathbf{v}_i) - (\boldsymbol{P}\mathbf{v}_i) \circ (\boldsymbol{P}\mathbf{v}_i) \big) - \big(\boldsymbol{P}(\mathbf{v}^*_{\alpha_i} \circ \mathbf{v}^*_{\alpha_i}) - (\boldsymbol{P}\mathbf{v}^*_{\alpha_i}) \circ (\boldsymbol{P}\mathbf{v}^*_{\alpha_i}) \big) \\ &= \boldsymbol{P}(\mathbf{v}_i \circ \mathbf{v}_i - \mathbf{v}^*_{\alpha_i} \circ \mathbf{v}^*_{\alpha_i}) + (\boldsymbol{P}\mathbf{v}^*_{\alpha_i}) \circ (\boldsymbol{P}\mathbf{v}^*_{\alpha_i}) - (\boldsymbol{P}\mathbf{v}_i) \circ (\boldsymbol{P}\mathbf{v}_i) \\ &= \boldsymbol{P}\big((\mathbf{v}_i - \mathbf{v}^*_{\alpha_i}) \circ (\mathbf{v}_i + \mathbf{v}^*_{\alpha_i}) \big) - (\boldsymbol{P}\mathbf{v}_i - \boldsymbol{P}\mathbf{v}^*_{\alpha_i}) \circ (\boldsymbol{P}\mathbf{v}_i + \boldsymbol{P}\mathbf{v}^*_{\alpha_i}). \end{aligned} \tag{97}$$

Furthermore, we have

$$\begin{aligned} \big\| \boldsymbol{P}\big((\mathbf{v}_i - \mathbf{v}^*_{\alpha_i}) \circ (\mathbf{v}_i + \mathbf{v}^*_{\alpha_i}) \big) \big\|_\infty &\le \|\boldsymbol{P}\|_1 \big\| (\mathbf{v}_i - \mathbf{v}^*_{\alpha_i}) \circ (\mathbf{v}_i + \mathbf{v}^*_{\alpha_i}) \big\|_\infty \\ &\le \|\boldsymbol{P}\|_1 \big\| \mathbf{v}_i - \mathbf{v}^*_{\alpha_i} \big\|_\infty \big\| \mathbf{v}_i + \mathbf{v}^*_{\alpha_i} \big\|_\infty \\ &\le \big(\|\mathbf{v}_i\|_\infty + \|\mathbf{v}^*_{\alpha_i}\|_\infty \big) \|\boldsymbol{\delta}_i\|_\infty, \end{aligned}$$

where the last inequality follows from $\|\boldsymbol{P}\|_1 = 1$, the triangular inequality, and (96). Meanwhile, similar arguments yield

$$\begin{aligned} \big\| (\boldsymbol{P}\mathbf{v}_i - \boldsymbol{P}\mathbf{v}^*_{\alpha_i}) \circ (\boldsymbol{P}\mathbf{v}_i + \boldsymbol{P}\mathbf{v}^*_{\alpha_i}) \big\|_\infty &\le \|\boldsymbol{P}\mathbf{v}_i - \boldsymbol{P}\mathbf{v}^*_{\alpha_i}\|_\infty \|\boldsymbol{P}\mathbf{v}_i + \boldsymbol{P}\mathbf{v}^*_{\alpha_i}\|_\infty \\ &\le \|\boldsymbol{P}\|_1 \|\mathbf{v}_i - \mathbf{v}^*_{\alpha_i}\|_\infty \|\boldsymbol{P}\|_1 \|\mathbf{v}_i + \mathbf{v}^*_{\alpha_i}\|_\infty \le \big(\|\mathbf{v}_i\|_\infty + \|\mathbf{v}^*_{\alpha_i}\|_\infty \big) \|\boldsymbol{\delta}_i\|_\infty. \end{aligned}$$

Plugging the two bounds back in (97) yields

$$\begin{aligned} \big\| \operatorname{Var}_{\boldsymbol{P}}(\mathbf{v}_i) - \operatorname{Var}_{\boldsymbol{P}}(\mathbf{v}^*_{\alpha_i}) \big\|_\infty &\le 2\big(\|\mathbf{v}_i\|_\infty + \|\mathbf{v}^*_{\alpha_i}\|_\infty \big) \|\boldsymbol{\delta}_i\|_\infty \\ &\le 2\big(i^{1/9} + 3(\log i)^2 i^{1/3} / c_1 \big) \|\boldsymbol{\delta}_i\|_\infty \le 2(1 + 3/c_1)(\log i)^2 i^{1/3} \|\boldsymbol{\delta}_i\|_\infty. \end{aligned}$$

□

D.5 Performance difference lemmas

The following performance difference lemma for average reward is adapted from Even-Dar et al. (2009, Lemma 4.1).

Lemma D.8 (Performance difference lemma)**.** *For any policy π whose long-term average reward is J^π, we define its bias function as*

$$q^\pi(s,a) = \mathbb{E}_\pi\Big[\sum_{k=1}^{\infty}\big(r(s_k,a_k) - J^\pi\big)\,\Big|\, s_1 = s, a_1 = a\Big],$$

Then for any two policies π and π', the difference between their long-term average reward is

$$J^{\pi'} - J^\pi = \mathbb{E}_{s\sim d_{\pi'}}\Big[\sum_{a\in\mathcal{A}}\big(\pi'(a\,|\,s) - \pi(a\,|\,s)\big)q^\pi(s,a)\Big],$$

where d_π is the long-term average visit probability of s under policy π.

We quote without proof the performance difference lemma for discounted reward (Kakade and Langford, 2002).

Lemma D.9 (Performance difference lemma)**.** *Recall that Q_γ^π and V_γ^π is the (scaled) Q- and value functions for any policy π and dicount factor γ. For any policies π and π', it holds that*

$$V_\gamma^\pi(s) - V_\gamma^{\pi'}(s) = \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d^\pi}\mathbb{E}_{a'\sim\pi(\cdot\,|\,s')}\big[Q_\gamma^{\pi'}(s',a') - V_\gamma^{\pi'}(s')\big], \quad \textit{for all } s\in\mathcal{S},$$

where $d^\pi(s') = (1-\gamma)\sum_{t=1}^\infty \mathbb{P}_\pi(S_t = s'\,|\,s_0 = s)$.

E Useful facts

In this section, we collect some useful facts that are useful for the technical proofs.

E.1 Magnitude of estimates

Lemma E.1. *Suppose the initialization of Algorithm 1 satisfies $\mathbf{0}\le \boldsymbol{Q}_0 \le \mathbf{1}$ (and thus $\mathbf{0}\le \boldsymbol{V}_0\le \mathbf{1}$), then $\mathbf{0}\le \boldsymbol{Q}_t\le \mathbf{1}$ and $\|\boldsymbol{\Delta}_t\|_\infty \le 1$ for all $t\ge 1$.*

Proof of Lemma E.1. Note that if $\mathbf{0}\le \boldsymbol{Q}_{t-1}\le \mathbf{1}$ and $\mathbf{0}\le \boldsymbol{V}_{t-1}\le \mathbf{1}$, then by the updating rule (6), since $\mathbf{0}\le \boldsymbol{r}\le \mathbf{1}$ and each row of $\boldsymbol{P}_t$ adds up to 1, we have $\boldsymbol{Q}_t\ge \mathbf{0}$ and

$$\boldsymbol{Q}_t \le (1-\eta_t)\mathbf{1} + \eta_t\big[(1-\gamma_t)\mathbf{1} + \gamma_t\mathbf{1}\big] = \mathbf{1}.$$

The desired result then follows from an induction argument. As a consequence, we have $\|\boldsymbol{\Delta}_t\|_\infty\le 1$ by the boundedness of optimal value functions that $\mathbf{0}\le \boldsymbol{Q}^*_{\gamma_t}\le \mathbf{1}$ for all $t\ge 1$. □

Lemma E.2. *With $\boldsymbol{q}_0 = \mathbf{0}$ and $\theta_t = (1+\frac{c_1 t^{2/3}}{(\log t)^2})^{-1}$ for $t\ge 2$ and $\theta_1 = 0$, Algorithm 2 obeys $\mathbf{0}\le \mathbf{q}_t \le 3(\log t)^2 t^{1/3}/c_1\mathbf{1}$ and $\mathbf{0}\le \mathbf{v}_t\le 3(\log t)^2 t^{1/3}/c_1\mathbf{1}$ for all $t\ge 2$.*

Proof of Lemma E.2. Noting the induction that

$$\begin{aligned}\|\mathbf{q}_t\|_\infty &\le (1-\theta_t)\|\mathbf{q}_{t-1}\|_\infty + \theta_t\big(\|\boldsymbol{r}\|_\infty + \alpha_t\|\mathbf{q}_{t-1}\|_\infty\big)\\ &\le (1-\theta_t+\theta_t\alpha_t)\|\mathbf{q}_{t-1}\|_\infty + \theta_t \le \|\mathbf{q}_{t-1}\|_\infty + \theta_t,\end{aligned}$$

we have

$$\|\mathbf{q}_t\|_\infty \le \sum_{i=1}^t \theta_i \le (\log t)^2\sum_{i=1}^t i^{-2/3}/c_1 \le 3(\log t)^2 t^{1/3}/c_1,$$

where the second inequality follows from (108). On the other hand, as $\mathbf{0}\le \boldsymbol{r}\le \mathbf{1}$, we have

$$\mathbf{q}_t \ge (1-\theta)\mathbf{q}_{t-1} + \theta_t\alpha_t\mathbf{q}_{t-1},$$

which, by induction, leads to $\mathbf{q}_t\ge \mathbf{0}$ for all $t\ge 1$. The definition of $\mathbf{v}_t$ implies $\mathbf{0}\le \mathbf{v}_t\le 3(\log t)^2 t^{1/3}/c_1\mathbf{1}$ for all $t\ge 2$, which complestes the proof of Lemma E.2. □

E.2 Learning rates for estimating optimal value in Theorem 3.1

In this part, we collect some useful facts about the learning rates speficied for Theorem 3.1 that would be used repeatedly. Recall that we define the learning rates as $\eta_t = (1 + \frac{c_1 t^{3/5}}{(\log t)^3})^{-1}$, $\forall\ t \geq 2$ for some constant $c_1 > 0$.

Bound on η_j. Firstly, we have the naive upper bound

$$\eta_j = \frac{1}{1 + \frac{c_1 j^{3/5}}{(\log j)^3}} \leq \frac{(\log j)^3}{c_1 j^{3/5}}, \qquad \text{for all } j \geq 2. \tag{98}$$

One can also check that $j^{3/5}/(\log j)^3 \geq 0.1$ for all $j \geq 1$ with the convention that $1/0 = \infty$, which leads to the lower bounds

$$\eta_j = \frac{1}{1 + \frac{c_1 j^{3/5}}{(\log j)^3}} \geq \frac{1}{\frac{10 j^{3/5}}{(\log j)^3} + \frac{c_1 j^{3/5}}{(\log j)^3}} \geq \frac{(\log j)^3}{(10 + c_1) j^{3/5}}, \qquad \text{for all } j \geq 2. \tag{99}$$

Furthermore, it can be easily checked that $(\log j)^3 j^{-3/5}$ is decreasing in j for $j \geq 150$. Hence

$$\eta_j \geq \eta_t \ \text{ and } \ \frac{(\log j)^3}{j^{3/5}} \geq \frac{(\log t)^3}{t^{3/5}}, \qquad \text{for all } t > 150 \ \text{ and } \ 150 \leq j \leq t. \tag{100}$$

Compound learning rates. Some associated quantities in the analysis include $\eta_t^{(t)} = \eta_t$,

$$\eta_0^{(t)} = \textstyle\prod_{j=1}^t (1 - \eta_j), \quad \text{and} \quad \eta_i^{(t)} = \eta_i \cdot \prod_{j=i+1}^t (1 - \eta_j), \quad \forall 1 \leq i < t.$$

The sum of $\eta_i^{(t)}$ over i obeys

$$\sum_{i=0}^t \eta_i^{(t)} = \prod_{j=1}^t (1 - \eta_j) + \eta_1 \prod_{j=2}^t (1 - \eta_j) + \cdots + \eta_{t-1}(1 - \eta_t) + \eta_t = 1. \tag{101}$$

Moreover, we consider a generic $\beta \in (0, 1)$ and bound the compound learning rates for $i \leq \lfloor (1-\beta)t \rfloor$ and $i > \lfloor (1-\beta)t \rfloor$ separately.

- When $\beta < 1/2$, for any $t \geq 300$ (so that $\lfloor (1-\beta)t \rfloor \geq 150$) and any $i \leq \lfloor (1-\beta)t \rfloor$,

$$\begin{aligned} \eta_i^{(t)} &\leq \prod_{j=i+1}^t (1 - \eta_j) \leq \exp\Big(- \textstyle\sum_{j=\lfloor (1-\beta)t \rfloor}^t \eta_j \Big) \\ &\leq \exp\Big(- \textstyle\sum_{j=\lfloor (1-\beta)t \rfloor}^t \eta_t \Big) \leq \exp\Big(- \beta t \cdot \frac{(\log t)^3}{c_1 t^{3/5}} \Big), \end{aligned} \tag{102}$$

  where the third inequality follows from (100), and the fourth inequality follows from (98).

- When $\beta < 1/2$, for any $t \geq 300$ and any $i > \lfloor (1-\beta)t \rfloor \geq t/2 \geq 150$,

$$\eta_i^{(t)} \leq \eta_i \leq \frac{(\log i)^3}{c_1 i^{3/5}} \leq \frac{2(\log t)^3}{c_1 t^{3/5}} \tag{103}$$

  following (98) and (100) and the fact that $\log i \leq \log t$ and $i^{-3/5} \leq 2t^{-3/5}$ for any $i \geq t/2$.

E.3 Learning rates for optimal policy in Theorem 3.3

We now present some useful facts about the learning rates that are particular to Theorem 3.3. Recall that in Theorem 3.3, we set $\gamma_t = 1 - t^{-1/8}$ and

$$\eta_t = \frac{1}{1 + \frac{c_1 t^{5/8}}{(\log t)^2}}, \quad t \geq 2$$

for some constant $c_1 > 0$ in Algorithm 1.

Bound on η_j. Firstly, we have the naive upper bound

$$\eta_j = \frac{1}{1+\frac{c_1 j^{5/8}}{(\log j)^2}} \le \frac{(\log j)^2}{c_1 j^{5/8}}, \qquad \text{for all } j \ge 2. \tag{104}$$

One can check that $j^{5/8}/(\log j)^2 \ge 0.5$ for all $j \ge 1$ with the convention that $1/0 = \infty$, which leads to the lower bounds

$$\eta_j = \frac{1}{1+\frac{c_1 j^{5/8}}{(\log j)^2}} \ge \frac{1}{\frac{2j^{5/8}}{(\log j)^2}+\frac{c_1 j^{5/8}}{(\log j)^2}} \ge \frac{(\log j)^2}{(2+c_1) j^{5/8}}, \qquad \text{for all } j \ge 2. \tag{105}$$

Furthermore, it can be easily checked that $(\log j)^2 j^{-5/8}$ is decreasing in j for $j \ge 50$. Hence

$$\eta_j \ge \eta_t \ \text{ and } \ \frac{(\log j)^2}{j^{5/8}} \ge \frac{(\log t)^2}{t^{5/8}}, \qquad \text{for all } t > 50 \ \text{ and } \ 50 \le j \le t. \tag{106}$$

Compound learning rates. Some associated quantities in the analysis include $\eta_t^{(t)} = \eta_t$,

$$\eta_0^{(t)} = \textstyle\prod_{j=1}^t (1-\eta_j), \quad \text{and} \quad \eta_i^{(t)} = \eta_i \cdot \prod_{j=i+1}^t (1-\eta_j), \quad \forall 1 \le i < t.$$

The sum of $\eta_i^{(t)}$ over i obeys

$$\sum_{i=0}^t \eta_i^{(t)} = \prod_{j=1}^t (1-\eta_j) + \eta_1 \prod_{j=2}^t (1-\eta_j) + \cdots + \eta_{t-1}(1-\eta_t) + \eta_t = 1.$$

Moreover, we consider a generic $\beta \in (0,1)$ and bound the compound learning rates for $i \le \lfloor (1-\beta)t \rfloor$ and $i > \lfloor (1-\beta)t \rfloor$ separately.

- When $\beta < 1/2$, for any $t \ge 100$ (so that $\lfloor (1-\beta)t \rfloor \ge 50$) and any $i \le \lfloor (1-\beta)t \rfloor$,

$$\begin{aligned} \eta_i^{(t)} \le \prod_{j=i+1}^t (1-\eta_j) &\le \exp\Big(- \textstyle\sum_{j=\lfloor (1-\beta)t \rfloor}^t \eta_j \Big) \\ &\le \exp\Big(- \textstyle\sum_{j=\lfloor (1-\beta)t \rfloor}^t \eta_t \Big) \le \exp\Big(- \beta t \cdot \frac{(\log t)^2}{c_1 t^{5/8}} \Big), \end{aligned}$$

  where the third inequality follows from (100), and the fourth inequality follows from (98).

- When $\beta < 1/2$, for any $t \ge 100$ and any $i > \lfloor (1-\beta)t \rfloor \ge t/2 \ge 50$,

$$\eta_i^{(t)} \le \eta_i \le \frac{(\log i)^3}{c_1 i^{5/8}} \le \frac{2(\log t)^2}{c_1 t^{5/8}} \tag{107}$$

  following (104) and (106) and the fact that $\log i \le \log t$ and $i^{-5/8} \le 2t^{-5/8}$ for any $i \ge t/2$.

E.4 Learning rates for bias function in Theorem 5.2

We now proceed to some basic facts on the learning rates of Algorithm 2. Recall that we define the learning rates as

$$\theta_t = \frac{1}{1+\frac{c_1 t^{2/3}}{(\log t)^2}}, \quad \forall\, t \ge 2.$$

Bound on θ_j. Firstly, we have the naive upper bound

$$\theta_j = \frac{1}{1+\frac{c_1 j^{2/3}}{(\log j)^2}} \le \frac{(\log j)^2}{c_1 j^{2/3}}, \qquad \text{for all } j \ge 1. \tag{108}$$

Also, one can check that $j^{2/3}/(\log j)^2 \ge 1/2$, which leads to

$$\theta_j = \frac{1}{1+\frac{c_1 j^{2/3}}{(\log j)^2}} \ge \frac{(\log j)^2}{(2+c_1) j^{2/3}}, \qquad \text{for all } j \ge 1.$$

Furthermore, it can be easily checked that $(\log j)^2 j^{-2/3}$ is decreasing in j for $j \ge 30$. Hence

$$\theta_j \ge \theta_t \ \text{ and } \ \frac{(\log j)^2}{j^{2/3}} \ge \frac{(\log t)^2}{t^{2/3}}, \qquad \text{for all } t > 50 \ \text{ and } \ 50 \le j \le t. \tag{109}$$

Compound learning rates. Some associated quantities in the analysis include $\theta_t^{(t)} = 1$,

$$\theta_0^{(t)} = \textstyle\prod_{j=1}^t (1-\theta_j), \quad \text{and} \quad \theta_i^{(t)} = \theta_i \cdot \prod_{j=i+1}^t (1-\theta_j), \quad \forall 1 \le i < t.$$

The sum of $\theta_i^{(t)}$ over i obeys

$$\sum_{i=0}^t \theta_i^{(t)} = \prod_{j=1}^t (1-\theta_j) + \theta_1 \prod_{j=2}^t (1-\theta_j) + \cdots + \theta_{t-1}(1-\theta_t) + \theta_t = 1.$$

Moreover, we also define

$$\beta = \frac{c_2}{T^{1/3}\log T},$$

and bound the compound learning rates for $i \le \lfloor (1-\beta)t \rfloor$ and $i > \lfloor (1-\beta)t \rfloor$ separately.

- For any t such that $\lfloor (1-\beta)t \rfloor \ge 50$ (so that $t \ge 50$) and any $i \le \lfloor (1-\beta)t \rfloor$,

$$\begin{aligned} \theta_i^{(t)} &\le \prod_{j=i+1}^t (1-\theta_j) \le \exp\Big(-\textstyle\sum_{j=\lfloor (1-\beta)t \rfloor}^t \theta_j \Big) \\ &\le \exp\Big(-\textstyle\sum_{j=\lfloor (1-\beta)t \rfloor}^t \theta_t \Big) \le \exp\Big(-\beta t \cdot \frac{(\log t)^2}{(2+c_1)t^{2/3}} \Big), \end{aligned} \tag{110}$$

  where the third inequality follows from (109), and the fourth inequality follows from (108).

- For any t such that $\lfloor (1-\beta)t \rfloor \ge 50$ (so that $t \ge 50$) and any $i > \lfloor (1-\beta)t \rfloor$,

$$\theta_i^{(t)} \le \theta_i \le \frac{c'(\log i)^3}{i^{3/5}} \le \frac{c'(\log t)^3}{t^{3/5}}$$

  following (108) and (109).