The Efficacy of Pessimism in Asynchronous Q-Learning

Yuling Yan* Gen Li† Yuxin Chen† Jianqing Fan*

March 2022; Revised: March 2023

Abstract

This paper is concerned with the asynchronous form of Q-learning, which applies a stochastic approximation scheme to Markovian data samples. Motivated by the recent advances in offline reinforcement learning, we develop an algorithmic framework that incorporates the principle of pessimism into asynchronous Q-learning, which penalizes infrequently-visited state-action pairs based on suitable lower confidence bounds (LCBs). This framework leads to, among other things, improved sample efficiency and enhanced adaptivity in the presence of near-expert data. Our approach permits the observed data in some important scenarios to cover only partial state-action space, which is in stark contrast to prior theory that requires uniform coverage of all state-action pairs. When coupled with the idea of variance reduction, asynchronous Q-learning with LCB penalization achieves near-optimal sample complexity, provided that the target accuracy level is small enough. In comparison, prior works were suboptimal in terms of the dependency on the effective horizon even when i.i.d. sampling is permitted. Our results deliver the first theoretical support for the use of pessimism principle in the presence of Markovian non-i.i.d. data.

Keywords: asynchronous Q-learning, offline reinforcement learning, pessimism principle, model-free algorithms, partial coverage, variance reduction

Contents

Corresponding author: Yuxin Chen (Email: yuxinc@wharton.upenn.edu).
*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; Email: {yulingy,jqfan}@princeton.edu.
†Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA; Email: {ligen,yuxinc}@wharton.upenn.edu.

1 Introduction

The asynchronous form of Q-learning, which is a stochastic approximation paradigm that applies to Markovian non-i.i.d. samples, has found applicability in an abundance of reinforcement learning (RL) applications (Even-Dar et al., 2003; Jaakkola et al., 1994; Tsitsiklis, 1994; Watkins and Dayan, 1992). The input data takes the form of a Markovian sample trajectory induced by a policy called the *behavior policy*; in each time, asynchronous Q-learning only updates the Q-function estimate of a single state-action pair along the trajectory rather than updating all pairs at once — and hence the terminology "asynchronous" (Bertsekas and Tsitsiklis, 2003; Tsitsiklis, 1994). This classical algorithm has the virtue of being off-policy, allowing one to learn the optimal policy even when the behavior policy is suboptimal. Recent years have witnessed a resurgence of interest in understanding the performance of asynchronous Q-learning, due to a shift of attention from classical asymptotic analysis to the non-asymptotic counterpart. By and large, non-asymptotic results bear important and clear implications for the impacts of salient parameters (e.g., model capacity, horizon length) in large-dimensional RL problems.

1.1 Motivation

A central consideration in modern RL applications is data efficiency: the limited availability of data samples places increasing demands on sample-efficient RL solutions, and in turn, calls for reexamining classical algorithms like Q-learning. When it comes to asynchronous Q-learning, recent theoretical advances have led to sharpened sample complexity analyses (Li et al., 2023, 2021b; Qu and Wierman, 2020). For concreteness, consider a γ-discounted infinite-horizon Markov decision process (MDP) and a stationary behavior policy: asynchronous Q-learning provably yields ε-accuracy as soon as the sample size exceeds the order of[1] (Li

[1]Here, the higher-order term $o(\frac{1}{\varepsilon^2})$ depends also on other parameters of the MDP and of the sample trajectory (e.g., the mixing time, the discount factor, and μ_{min}).

et al., 2023)

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2} + o\left(\frac{1}{\varepsilon^2}\right) \tag{1.1}$$

modulo some log factor, where μ_{min} stands for the minimum occupancy probability of the sample trajectory over all state-action pairs. While this bound (1.1) is tight in a general sense for vanilla Q-learning, two issues immediately spring into mind.

- *Uniform coverage vs. partial coverage.* The factor $1/\mu_{\mathsf{min}}$ in (1.1) imposes a firm requirement on uniform coverage of the state-action space, namely, every state-action pair needs to be visited sufficiently often in order to guarantee reliable learning. Nevertheless, it is not uncommon for a behavior policy to provide only partial coverage of the state-action space; for instance, a behavior policy might elect to rule out several actions that are clearly underperforming. In truth, partial coverage of the state-action space results in $\mu_{\mathsf{min}} = 0$, thus making the general bound (1.1) vacuous in this case.
- *Lack of adaptivity to expert data.* The general bound (1.1) falls short of reflecting the quality of the sample trajectory (except for a general uniform coverage parameter μ_{min}). For instance, if the behavior policy is adopted by an "expert" who is already aware of which actions are (close to) optimal, then such expert data could be more informative than a general sample trajectory with the same μ_{min}. It is therefore desirable for an algorithm to adapt automatically to the quality of the data, in the hope of achieving sample size saving when expert data is available.

1.2 Main contributions

This paper seeks to make asynchronous Q-learning adaptive to near-expert data, allowing for partial coverage of the state-action space in some important scenarios. A key idea that has been recently proposed to accommodate partial coverage in the presence of near-expert data is the principle of pessimism (or conservatism) in the face of uncertainty (Jin et al., 2021; Rashidinejad et al., 2021), whose benefits have been established in the context of offline RL (or batch RL). In a nutshell, the pessimism principle penalizes the Q-function based on how infrequent a state-action pair is visited, which effectively directs the attention of an RL algorithm away from the under-covered part of the state-action space. However, it remains unclear how effective this idea of pessimism could be in the asynchronous setting when coping with Markovian data.

In order to address this issue, the current paper revisits asynchronous Q-learning in the presence of a Markovian sample trajectory generated by a behavior policy π_{b}. We focus on a γ-discounted infinite-horizon MDP with S states and A actions, and suppose that the behavior policy is stationary and satisfies a certain single-policy concentrability assumption (associated with a test distribution ρ) with coefficient $C^\star \geq 1$; informally, this means that the observed sample trajectory effectively becomes expert data as $C^\star$ approaches 1, as we shall formalize in Section 2. Our contributions are two-fold; here and below, $\widetilde{O}(\cdot)$ stands for the orderwise upper bound while hiding any logarithmic dependency.

- *Asynchronous Q-learning with LCB penalization.* We propose a variant of asynchronous Q-learning by penalizing each Q-learning iteration based on a lower confidence bound (LCB). This variant of Q-learning achieves ε-accuracy (w.r.t. a test distribution ρ) as long as the total sample size is above the order of

$$\widetilde{O}\left(\frac{SC^\star}{(1-\gamma)^5\,\varepsilon^2}\right),$$

  provided that the accuracy level ε is small enough. Given that $C^\star$ can be as small as $O(1)$ and given the trivial bound $1/\mu_{\mathsf{min}} \geq SA$ (so that (1.1) $\geq \frac{SA}{(1-\gamma)^4\varepsilon^2}$), our theory leads to sample size benefits in terms of its dependency on A when the data is near-expert.
- *Variance-reduced asynchronous Q-learning with LCB penalization.* While asynchronous Q-learning with LCB penalization allows for reduced sample complexity in the presence of near-expert data, the dependency on the effective horizon $\frac{1}{1-\gamma}$ remains suboptimal. To address this, we leverage the idea of variance

reduction (also called reference-advantage decomposition) (Li et al., 2021a; Wainwright, 2019b; Zhang et al., 2020) to further accelerate convergence of the algorithm, which in turn yields a sample complexity

$$\widetilde{O}\left(\frac{SC^{\star}}{(1-\gamma)^{3}\,\varepsilon^{2}}\right)$$

for sufficiently small accuracy level ε. The scaling $\frac{1}{(1-\gamma)^3}$ is essentially unimprovable even for the synchronous setting with independent samples (Azar et al., 2013; Rashidinejad et al., 2021). Notably, none of the prior works on offline RL were able to achieve the scaling of $\frac{SC^{\star}}{(1-\gamma)^3}$; that is, the best-known theory (Rashidinejad et al., 2021) scales as $\widetilde{O}\big(\frac{SC^{\star}}{(1-\gamma)^5\varepsilon^2}\big)$ and relies on i.i.d. sampling.

Finally, we remark that the algorithmic and theoretical frameworks put forward herein are suitable for two important scenarios in the absence of active exploration of the environment: (i) *online reinforcement learning* with a time-invariant policy (so that the data arrives on the fly with no policy evolvement), and (ii) *offline reinforcement learning,* where the data generated by the behavior policy has been pre-collected. In addition to the appealing sample complexity, model-free algorithms also enjoy the benefits of low memory and low computational complexity.

2 Models and assumptions

Basics of infinite-horizon Markov decision processes. In this paper, we consider an infinite-horizon Markov decision process, denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, r)$. Here, $\mathcal{S}$ represents the state space that contains S distinct states; $\mathcal{A}$ stands for the action space that contains A distinct actions; $\gamma \in (0,1)$ denotes the discount factor, with $\frac{1}{1-\gamma}$ representing the effective horizon; $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ stands for the probability transition kernel (with $\Delta(\mathcal{S})$ denoting the probability simplex over the set $\mathcal{S}$), such that $P(\cdot \,|\, s, a) \in \Delta(\mathcal{S})$ denotes the transition probability from state s when action a is executed; $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ indicates the deterministic reward function, such that $r(s, a)$ is the immediate reward gained in state s upon execution of action a. We assume throughout that the immediate rewards fall within the range $[0, 1]$.

Let $\Delta(\mathcal{A})$ be the probability simplex over the set $\mathcal{A}$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is an action selection rule, such that $\pi(\cdot \,|\, s) \in \Delta(\mathcal{A})$ specifies the action selection probability in state s. When π is deterministic, we often overload the notation and let $\pi(s)$ represent the action selected in state s. The value function and the Q-function of policy π are defined respectively as

$$\begin{aligned}
\forall s \in \mathcal{S}: \qquad V^{\pi}(s) &\coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s\right], \\
\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q^{\pi}(s,a) &\coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a\right],
\end{aligned}$$

where the expectation is taken over a random trajectory $(s_0, a_0, s_1, a_1, s_2, a_2, \cdots)$ induced by the MDP $\mathcal{M}$ when policy π is employed. For a given initial state distribution $\rho \in \Delta(\mathcal{S})$, we can also overload the notation of the value function to represent a certain average value function:

$$V^{\pi}(\rho) \coloneqq \mathbb{E}_{s \sim \rho}\big[V^{\pi}(s)\big].$$

Moreover, it is well known that there exists at least one *deterministic* policy, denoted by $\pi^{\star}$, that simultaneously maximizes the value function and the Q-function over all state-action pairs. Therefore, we introduce the following notation

$$V^{\star}(s) \coloneqq \max_{\pi} V^{\pi}(s), \qquad V^{\star}(\rho) \coloneqq \mathbb{E}_{s\sim\rho}\big[V^{\star}(s)\big], \qquad \text{and} \qquad Q^{\star}(s,a) \coloneqq \max_{\pi} Q^{\pi}(s,a)$$

to represent the optimal value function and the optimal Q-function. Given a test distribution $\rho \in \Delta(\mathcal{S})$ and a target accuracy level $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big)$, our aim is to develop a policy $\widehat{\pi}$ obeying

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \le \varepsilon.$$

A kind of distributions that plays an important role in our theory is the discounted state-action occupancy distribution defined as follows:

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \qquad d_{\rho}^{\pi}\left(s,a\right) \coloneqq (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t}\mathbb{P}\left(s_{t}=s,a_{t}=a\mid\pi,s_{0}\sim\rho\right), \tag{2.1}$$

$$d_{\rho}^{\pi}\left(s\right) \coloneqq (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t}\mathbb{P}\left(s_{t}=s\mid\pi,s_{0}\sim\rho\right), \tag{2.2}$$

where the trajectory $(s_0,a_0,s_1,a_1,s_2,a_2,\cdots)$ is induced by the MDP under the policy π and a given initial state distribution ρ. When π coincides with the optimal policy $\pi^\star$, we abbreviate

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \qquad d_{\rho}^{\star}(s,a)\coloneqq d_{\rho}^{\pi^\star}(s,a) \qquad \text{and} \qquad d_{\rho}^{\star}(s)\coloneqq d_{\rho}^{\pi^\star}(s)=d_{\rho}^{\pi^\star}\big(s,\pi^\star(s)\big). \tag{2.3}$$

Sampling mechanism. Suppose that the observed Markovian sample trajectory $\left\{(s_t,a_t)\right\}_{t\geq 0}$ is obtained by executing a behavior policy π_{b} in the MDP $\mathcal{M}$. We say that the total sample size is T if the algorithm employs T state-action pairs of this trajectory, i.e., $\left\{(s_t,a_t)\right\}_{0\leq t\leq T}$. Assume that $\mu_{\mathsf{b}}(s,a)$ is the stationary distribution of the this Markov chain generated by π_{b}, with the minimum state-action occupancy probability defined to be

$$\mu_{\mathsf{min}}\coloneqq\min_{s\in\mathcal{S},\,a\in\mathcal{A}}\mu_{\mathsf{b}}(s,a).$$

We impose the following assumptions on π_{b} throughout this paper.

Assumption 1. *The behavior policy π_{b} is stationary, and the Markov chain induced by π_{b} is uniformly ergodic.*

Remark 1. In words, uniform ergodicity says that for any initial state-action pair, the total-variation distance between the distribution of (s_t,a_t) and the stationary distribution of the chain decays geometrically in t; see Paulin (2015, Definition 1.1) for a precise definition of uniform ergodicity.

Furthermore, for a given test distribution or initial state distribution $\rho\in\Delta(\mathcal{S})$, we adopt the following concept as introduced in Rashidinejad et al. (2021).

Assumption 2 (Single-policy concentrability)**.** *Suppose that there exists some constant $C^\star\geq 1$ such that*

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \qquad \frac{d_{\rho}^{\star}\left(s,a\right)}{\mu_{\mathsf{b}}\left(s,a\right)}\leq C^{\star}, \tag{2.4}$$

where we define $0/0=0$ by convention. Throughout this paper, $C^\star\geq 1$ is called the single-policy concentrability coefficient.

In some sense, the single-policy concentrability coefficient measures the closeness between the stationary distribution of the observed data and a certain occupancy distribution induced by the optimal policy. In particular, if we take $\rho=\mu^\star$ to be the stationary *state distribution* of the MDP under the deterministic policy $\pi^\star$, then it can be easily verified that $d_{\mu^\star}^{\star}(s,a)=\mu^\star(s)\,\mathbb{1}\{\pi^\star(s)=a\}$, allowing us to rewrite (2.4) w.r.t. the density ratio of two stationary distributions as follows:

$$\forall s\in\mathcal{S}: \qquad \frac{\mu^{\star}\left(s\right)}{\mu_{\mathsf{b}}\big(s,\pi^{\star}(s)\big)}\leq C^{\star}. \tag{2.5}$$

In this paper, the sample data is said to be near-expert if $C^\star=O(1)$, as in this case the empirical distribution of the sample data is not far away from what is induced by the optimal policy. Note that the introduction of the single-policy concentrability coefficient $C^\star$ is solely for the purpose of theoretical analysis, and the algorithms proposed herein do not require prior knowledge of $C^\star$ at all. It is also worth noting that $C^\star$ (cf. (2.4)) is a function of the test distribution ρ as well, although we suppress this dependency in the notation $C^\star$ for the sake of conciseness.

Another important quantity that affects the performance of our model-free algorithms is the mixing time associated with the sample trajectory. To be precise, for any $0 < \delta < 1$, the mixing time of the Markov chain induced by the MDP $\mathcal{M}$ under behavior policy π_{b} is defined as

$$t_{\mathsf{mix}}(\delta) := \min \left\{ t : \max_{s_0 \in \mathcal{S}, a_0 \in \mathcal{A}} d_{\mathsf{TV}}\left(P^t\left(\cdot \mid s_0, a_0\right), \mu_{\mathsf{b}}\right) \leq \delta \right\}.$$

Here, $P^t(\cdot \mid s_0, a_0)$ stands for the distribution of (s_t, a_t) (i.e., the state-action pair in the t-th step of the trajectory) when the chain is initialized to (s_0, a_0), whereas $d_{\mathsf{TV}}(\mu, \nu)$ is the total-variation distance between two distributions μ and ν over a discrete space $\mathcal{X}$ (Tsybakov and Zaiats, 2009), namely,

$$d_{\mathsf{TV}}(\mu, \nu) = \frac{1}{2} \sum_{x \in \mathcal{X}} \left| \mu(x) - \nu(x) \right| = \sup_{B \subseteq \mathcal{X}} \left| \mu(B) - \nu(B) \right|.$$

In particular, we shall abbreviate

$$t_{\mathsf{mix}} := t_{\mathsf{mix}}(1/4),$$

following the convention in prior works like Paulin (2015). Clearly, this important quantity measures how long it takes for a Markov chain to decorrelate itself from the initial state.

Remark 2. Another simpler sampling mechanism studied in prior literature (e.g., Rashidinejad et al. (2021)) is i.i.d. sampling, under which the observed sample trajectory takes the form of $\{(s_t, a_t, s'_t)\}_{1 \leq t \leq T}$ with

$$(s_t, a_t) \sim \mu_{\mathsf{b}} \qquad \text{and} \qquad s'_t \sim P(\cdot \mid s_t, a_t), \qquad 1 \leq t \leq T$$

independently generated. It is worth mentioning that the theorems and analysis in the current paper automatically apply to i.i.d. sampling by taking $t_{\mathsf{mix}} = 1$. Clearly, the Markovian sample trajectory studied herein is in general more challenging to cope with, due to the complicated Markovian dependency.

3 Asynchronous Q-learning with LCB penalization

In this section, we describe how to incorporate the pessimism principle into classical asynchronous Q-learning, accompanied by our theoretical performance guarantees.

3.1 Algorithm

We introduce the key algorithmic ingredients of our first algorithm: asynchronous Q-learning with LCB penalization. The complete details can be found in Algorithm 1.

Asynchronous Q-learning. Let us begin by reviewing the basics of asynchronous Q-learning, which maintains iterates $\{Q_t\}$ as the Q-function estimates. In each iteration t, the algorithm takes action $a_{t-1} \sim \pi_{\mathsf{b}}(\cdot \mid s_{t-1})$, observes the next state $s_t \sim P(\cdot \mid s_{t-1}, a_{t-1})$, and then updates its Q-function estimate w.r.t. a single state-action pair (s_{t-1}, a_{t-1}) as follows

$$\begin{aligned} Q_t\left(s_{t-1}, a_{t-1}\right) &= \left(1 - \eta_n\right) Q_{t-1}\left(s_{t-1}, a_{t-1}\right) + \eta_n \Big\{ r\left(s_{t-1}, a_{t-1}\right) + \gamma V_{t-1}\left(s_t\right) \Big\}, \\ Q_t\left(s, a\right) &= Q_{t-1}\left(s, a\right), \qquad \forall (s, a) \neq (s_{t-1}, a_{t-1}). \end{aligned}$$

Here, n represents the number of visits to (s_{t-1}, a_{t-1}) prior to the t-th iteration, $0 < \eta_n < 1$ stands for the learning rate, and the value function estimate is defined to be $V_{t-1}(s) := \max_{a \in \mathcal{A}} Q_{t-1}(s, a)$.

The pessimism principle and LCB penalization. In order to accommodate under-coverage of the state-action space in the presence of near-expert data, a key idea is to penalize the Q-function of those state-action pairs that are rarely visited (i.e., the ones that are not favored by the "expert"), so as to downplay their influence on the Q-estimates. Specifically, in the t-th iteration, we modify the Q-learning update by inserting a penalty term b_n:

$$Q_t\left(s_{t-1}, a_{t-1}\right) = \left(1 - \eta_n\right) Q_{t-1}\left(s_{t-1}, a_{t-1}\right) + \eta_n \Big\{ r\left(s_{t-1}, a_{t-1}\right) + \gamma V_{t-1}\left(s_t\right) - b_n \Big\}, \tag{3.1a}$$

Algorithm 1: Asynchronous Q learning with LCB penalization.

Input: number of iterations T, initial state s.

Initialize: $Q_0(s,a)=0$, $V_0(s)=0$, $n_0(s,a)=0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, $H=\lceil\frac{4}{1-\gamma}\log\frac{ST}{\delta}\rceil$.

for $t=1$ **to** T **do**

Draw $a_{t-1}\sim\pi_{\mathsf{b}}(\cdot\,|\,s_{t-1})$, and observe $s_t\sim P(\cdot\,|\,s_{t-1},a_{t-1})$.

Let $n_t(s_{t-1},a_{t-1})=n_{t-1}(s_{t-1},a_{t-1})+1$; and $n_t(s,a)=n_{t-1}(s,a)$, $\forall(s,a)\neq(s_{t-1},a_{t-1})$.

Set $n\leftarrow n_t(s,a)$, and take $\eta_n=(H+1)/(H+n)$.

Update

$$Q_t(s_{t-1},a_{t-1})=(1-\eta_n)\,Q_{t-1}(s_{t-1},a_{t-1})+\eta_n\Big\{r(s_{t-1},a_{t-1})+\gamma V_{t-1}(s_t)-b_n\Big\}$$

and $Q_t(s,a)=Q_{t-1}(s,a)$ for all $(s,a)\neq(s_{t-1},a_{t-1})$, where

$$b_n=C_{\mathsf{b}}\sqrt{\frac{H\log(ST/\delta)}{n(1-\gamma)^2}}$$

for some sufficiently large constant $C_{\mathsf{b}}>0$.

Update

$$V_t(s_{t-1})=\max\Big\{\max_{a\in\mathcal{A}}Q_t(s_{t-1},a)\,,\ V_{t-1}(s_{t-1})\Big\},$$

and $V_t(s)=V_{t-1}(s)$ for all $s\neq s_{t-1}$.

Output: $\widehat{\pi}$ such that $\widehat{\pi}(s)=\arg\max_{a\in\mathcal{A}}Q_T(s,a)$ for all $s\in\mathcal{S}$.

$$Q_t(s,a)=Q_{t-1}(s,a)\,,\qquad\forall(s,a)\neq(s_{t-1},a_{t-1}),\tag{3.1b}$$

where the penalty term b_n is chosen to be some lower-confidence bound (LCB) determined by the Hoeffding concentration inequality. More precisely, we shall set

$$b_n=C_{\mathsf{b}}\sqrt{\frac{H\log(ST/\delta)}{n(1-\gamma)^2}}\tag{3.2}$$

throughout this paper, where we take $H=\lceil\frac{4}{1-\gamma}\log\frac{ST}{\delta}\rceil$ — so that b_n is on the order of $\widetilde{O}\big(\sqrt{\frac{1}{(1-\gamma)^3n}}\,\big)$ — and recall that n is the number of visits to (s_{t-1},a_{t-1}) prior to time t. The rationale behind this specific choice will be made clear in the analysis.

Monotonicity of value function estimates. In addition to the above pessimism principle, another consideration is to ensure that the value function estimate V_t always improves upon (or at least, is no worse than) the previous estimate. Towards this end, we take

$$V_t(s_{t-1})=\max\Big\{\max_{a\in\mathcal{A}}Q_t(s_{t-1},a)\,,\ V_{t-1}(s_{t-1})\Big\},$$
$$V_t(s)=V_{t-1}(s)\qquad\text{for all } s\neq s_{t-1},$$

which yields monotonically non-decreasing value function estimates $\{V_t\}_{t\geq0}$. This simple modification facilitates analysis while ensuring that $V_t(s)$ is always non-negative (as long as we initialize $V_t(s)\geq0$ for all $s\in\mathcal{S}$).

Computational and memory complexities. The whole algorithm, as summarized in Algorithm 1 has low runtime $O(T)$ and low memory complexity $O(\min\{T,SA\})$ (note that if a state-action pair is never visited, we do not need to record/update any quantity related to it).

3.2 Theoretical guarantees

Equipped with LCB penalization, asynchronous Q-learning is capable of achieving appealing sample efficiency, even though the observed sample trajectory might not provide full coverage of the state-action space. This is stated in the following theorem, whose proof is postponed to Section B.

Theorem 1. *Suppose that Assumptions 1 and 2 hold, and recall that T is the total number of samples. With probability exceeding $1-\delta$, the policy $\widehat{\pi}$ returned by Algorithm 1 satisfies*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \lesssim \sqrt{\frac{C^{\star} S \iota^2}{T(1-\gamma)^5}} + \frac{C^{\star} S t_{\mathsf{mix}} \iota}{T(1-\gamma)^2} + \frac{C^{\star} t_{\mathsf{mix}} \iota^2}{T(1-\gamma)^3}, \tag{3.3}$$

where $\iota := \log(ST/\delta)$.

By taking the right-hand side of (3.3) to be bounded above by ε, we immediately see that Algorithm 1 achieves ε-accuracy with high probability, as long as the total sample size T exceeds

$$\widetilde{O}\left(\frac{SC^{\star}}{(1-\gamma)^5 \varepsilon^2} + \frac{\big(S + \frac{1}{1-\gamma}\big) t_{\mathsf{mix}} C^{\star}}{(1-\gamma)^2 \varepsilon}\right). \tag{3.4}$$

This also means that the sample complexity of Algorithm 1 scales as

$$\widetilde{O}\left(\frac{SC^{\star}}{(1-\gamma)^5 \varepsilon^2}\right) \tag{3.5}$$

for any target accuracy level $0 < \varepsilon \leq \frac{S}{\big(S+\frac{1}{1-\gamma}\big)(1-\gamma)^3 t_{\mathsf{mix}}}$. When we have near-expert data (so that $C^{\star} = O(1)$), the sample complexity can be as low as

$$\widetilde{O}\left(\frac{S}{(1-\gamma)^5 \varepsilon^2}\right).$$

In comparison, the general bound (1.1) developed in the previous literature requires at least $\frac{SA}{(1-\gamma)^4 \varepsilon^2}$ samples (since $1/\mu_{\mathsf{min}} \geq SA$) regardless of what behavior policy is employed. As a result, the proposed algorithm enjoys enhanced adaptivity to near-expert data, particularly in the presence of large action space and/or partial coverage.

It is worth noting, however, that the bound (3.5) exhibits a dependency $\frac{1}{(1-\gamma)^5}$ on the effective horizon as opposed to $\frac{1}{(1-\gamma)^4}$, due to the adoption of the Hoeffding-style penalty (3.2). This is potentially improvable by designing more careful Bernstein-style penalty (akin to Jin et al. (2018, Section 3)). Nevertheless, we do not pursue this for two reasons: (a) the Hoeffding-style penalty streamlines analysis; (b) the Bernstein-style penalty alone is insufficient to yield optimal sample complexity, and we shall put forward another algorithm momentarily to achieve sample optimality.

4 Variance-reduced asynchronous Q-learning with LCB penalization

As we have alluded to previously, the algorithm presented in Section 3 falls short of achieving optimal dependency on the effective horizon. To address this issue, a plausible idea is to leverage the variance reduction technique — originally introduced in finite-sum stochastic optimization (Johnson and Zhang, 2013) and imported to online RL recently (Zhang et al., 2020) — to further accelerate convergence of the algorithm. This section is devoted to the development of a new variant of asynchronous Q-learning that incorporates both pessimism and variance reduction.

4.1 Algorithm

We start by describing the key ideas of a variance-reduced variant of Algorithm 1. This algorithm enjoys the same computational cost (i.e., $O(T)$) and memory complexity (i.e., $O(SA)$) as Algorithm 1, with full details are summarized in Algorithm 2 (in conjunction with Algorithms 3 and 4).

Algorithm 2: Variance-reduced asynchronous Q-learning with LCB penalization.

Input: number of iterations T, initial state s.
Initialize: $\overline{V}(s) = 0$ for all $s \in \mathcal{S}$, $K = \lfloor \log_4(3T/4) \rfloor$.
for $k = 1$ **to** K **do**
  Set $T_k^{\mathsf{ref}} = 4^{k-1}$ and $T_k = 3 \times 4^{k-1}$.
  Call function $\mathsf{Empirical\text{-}transition}(T_k^{\mathsf{ref}}, \overline{V}, s)$ (cf. Algorithm 3) and return $(\widetilde{P}, b^{\mathsf{ref}}, s_1)$.
  Call function $\mathsf{VR\text{-}Q\text{-}epoch}(T_k, \overline{V}, \widetilde{P}, b^{\mathsf{ref}}, s_1)$ (cf. Algorithm 4) and return (Q, V, s_2).
  Update the reference $\overline{V} = V$, and set the initial state in the next epoch as $s = s_2$.
Output: $\widehat{\pi}$ such that $\widehat{\pi}(s) = \arg\max_{a \in \mathcal{A}} Q(s,a)$ for all $s \in \mathcal{S}$.

Algorithm 3: $\mathsf{Empirical\text{-}transition}(T^{\mathsf{ref}}, \overline{V}, s_0^{\mathsf{ref}})$

Input: number of samples T^{ref}, reference $\overline{V}$, initial state s_0^{ref}.
Initialize: $n^{\mathsf{ref}}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\iota = \log \frac{ST}{\delta}$.
for $t = 1$ **to** T^{ref} **do**
  Draw $a_{t-1}^{\mathsf{ref}} \sim \pi_{\mathsf{b}}(\cdot \,|\, s_{t-1}^{\mathsf{ref}})$, and observe $s_t^{\mathsf{ref}} \sim P(\cdot \,|\, s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}})$.
  Let $n^{\mathsf{ref}}(s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}}) \leftarrow n^{\mathsf{ref}}(s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}}) + 1$. Set $n \leftarrow n^{\mathsf{ref}}(s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}})$.
  Update

$$
\begin{aligned}
\widetilde{P}\big(s_t^{\mathsf{ref}} \,|\, s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}}\big) &\leftarrow \frac{(n-1)\widetilde{P}\big(s_t^{\mathsf{ref}} \,|\, s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}}\big) + 1}{n}, \\
\mu^{\mathsf{ref}}(s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}}) &\leftarrow \frac{(n-1)\mu^{\mathsf{ref}}(s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}}) + \overline{V}\left(s_t^{\mathsf{ref}}\right)}{n}, \\
\sigma^{\mathsf{ref}}(s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}}) &\leftarrow \frac{(n-1)\sigma^{\mathsf{ref}}(s_{t-1}^{\mathsf{ref}}, a_{t-1}^{\mathsf{ref}}) + \overline{V}^2\left(s_t^{\mathsf{ref}}\right)}{n}.
\end{aligned}
$$

Compute the penalty term: for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, take

$$
b^{\mathsf{ref}}(s,a) = C_{\mathsf{b}}\left(\sqrt{\frac{\sigma^{\mathsf{ref}}(s,a) - \left[\mu^{\mathsf{ref}}(s,a)\right]^2}{n^{\mathsf{ref}}(s,a)}\iota} + \frac{\iota^{3/4}}{(1-\gamma)\left[n^{\mathsf{ref}}(s,a)\right]^{3/4}} + \frac{\iota}{(1-\gamma)\, n^{\mathsf{ref}}(s,a)}\right)
$$

for some sufficiently large constant $C_{\mathsf{b}} > 0$.
Output: empirical probability transition $\widetilde{P}$, penalty b^{ref}, last state $s_{T^{\mathsf{ref}}}$.

Variance reduction. Suppose for the moment that we have access to a "reference" value function estimate $\overline{V}$ that is hopefully not far away from the true optimal value $V^\star$. Let us employ a batch of samples — more concretely, a total number of T^{ref} consecutive samples $\{(s_i^{\mathsf{ref}}, a_i^{\mathsf{ref}}, s_{i+1}^{\mathsf{ref}}) : 0 \le i < T^{\mathsf{ref}}\}$ — to compute an empirical estimate $\widetilde{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ of the probability transition kernel P. We can then incorporate variance reduction into the update rule (3.1) of Algorithm 1 as follows:

$$
\begin{aligned}
&Q_t(s_{t-1}, a_{t-1}) = (1-\eta_n)\, Q_{t-1}(s_{t-1}, a_{t-1}) + \\
&\quad \eta_n \left\{ r(s_{t-1}, a_{t-1}) + \gamma V_{t-1}(s_t) - \gamma \overline{V}(s_t) + \gamma \big\langle \widetilde{P}(\cdot \,|\, s_{t-1}, a_{t-1}), \overline{V} \big\rangle - b_n(s_{t-1}, a_{t-1}) \right\}.
\end{aligned} \tag{4.1}
$$

Here, the penalty term $b_n(s_{t-1}, a_{t-1})$ is set to be a certain data-driven lower confidence bound tailored to this variance-reduced update rule. In particular, this penalty term is chosen to track the uncertainty of both the "advantage term" $V_{t-1}(s_t) - \overline{V}(s_t)$ and the "reference term" $\big\langle \widetilde{P}(\cdot \,|\, s_{t-1}, a_{t-1}), \overline{V} \big\rangle$, inspired by the reference-advantage decomposition introduced in Zhang et al. (2020); see Algorithm 4 for a precise description. As can be anticipated, if $\overline{V}$ is a more accurate estimate of $V^\star$ than V_{t-1} (i.e., $\overline{V} \approx V^\star$ and $\|\overline{V} - V^\star\|_\infty \ll \|V_{t-1} - V^\star\|_\infty$), then the main stochastic term $\overline{V}(s_t)$ (or $\overline{V}(s_t) - V^\star(s_t)$) in (4.1) is expected to

Algorithm 4: VR-Q-epoch$(T, \overline{V}, \widetilde{P}, b^{\mathsf{ref}}, s_0)$

Input: number of iterations T, reference $\overline{V}$, transition kernel $\widetilde{P}$, penalty b^{ref}, initial state s_0.
Initialize: $Q_0(s,a) = 0$, $V_0(s) = 0$, $n_0(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\iota = \log \frac{ST}{\delta}$, $H = \lceil \frac{4\iota}{1-\gamma} \rceil$.
for $t = 1$ **to** T **do**
  Draw $a_{t-1} \sim \pi_{\mathsf{b}}(\cdot \,|\, s_{t-1})$, and observe $s_t \sim P(\cdot \,|\, s_{t-1}, a_{t-1})$.
  Let $n_t(s_{t-1}, a_{t-1}) \leftarrow n_{t-1}(s_{t-1}, a_{t-1}) + 1$; and $n_t(s,a) \leftarrow n_{t-1}(s,a)$ for all $(s,a) \neq (s_{t-1}, a_{t-1})$.
  Set $n \leftarrow n_t(s,a)$, and take $\eta_n = (H+1)/(H+n)$.
  Set $\mu_n^{\mathsf{adv}}(s,a) = \mu_{n-1}^{\mathsf{adv}}(s,a)$ and $\sigma_n^{\mathsf{adv}}(s,a) = \sigma_{n-1}^{\mathsf{adv}}(s,a)$ for all $(s,a) \neq (s_{t-1}, a_{t-1})$; update

$$\begin{aligned}
\mu_n^{\mathsf{adv}}(s_{t-1}, a_{t-1}) &= (1-\eta_n)\,\mu_{n-1}^{\mathsf{adv}}(s_{t-1}, a_{t-1}) + \eta_n \left[V_{t-1}(s_t) - \overline{V}(s_t)\right], \\
\sigma_n^{\mathsf{adv}}(s_{t-1}, a_{t-1}) &= (1-\eta_n)\,\sigma_{n-1}^{\mathsf{adv}}(s_{t-1}, a_{t-1}) + \eta_n \left[V_{t-1}(s_t) - \overline{V}(s_t)\right]^2.
\end{aligned}$$

  Compute $\mathsf{sd}_n^{\mathsf{adv}}(s_{t-1}, a_{t-1}) = \sigma_n^{\mathsf{adv}}(s_{t-1}, a_{t-1}) - \left[\mu_n^{\mathsf{adv}}(s_{t-1}, a_{t-1})\right]^2$.
  Update

$$\begin{aligned}
Q_t(s_{t-1}, a_{t-1}) &= (1-\eta_n)\, Q_{t-1}(s_{t-1}, a_{t-1}) \\
&\quad + \eta_n \left\{ r(s_{t-1}, a_{t-1}) + \gamma V_{t-1}(s_t) - \gamma \overline{V}(s_t) + \gamma \big\langle \widetilde{P}(\cdot \,|\, s_{t-1}, a_{t-1}), \overline{V} \big\rangle - b_n \right\}.
\end{aligned}$$

  and $Q_t(s,a) = Q_{t-1}(s,a)$ for all $(s,a) \neq (s_{t-1}, a_{t-1})$, where $b_t = b^{\mathsf{ref}}(s_{t-1}, a_{t-1}) + b^{\mathsf{adv}}$ and

$$b^{\mathsf{adv}} = C_{\mathsf{b}} \left(\sqrt{\frac{H\iota}{n} \frac{\mathsf{sd}_n^{\mathsf{adv}}(s_{t-1}, a_{t-1}) - (1-\eta_n)\mathsf{sd}_{n-1}^{\mathsf{adv}}(s_{t-1}, a_{t-1})}{\eta_n}} + \frac{H^{3/4}\iota^{3/4}}{n^{3/4}(1-\gamma)} + \frac{H\iota}{n(1-\gamma)} \right)$$

  for some sufficiently large constant $C_{\mathsf{b}} > 0$.
  Update

$$V_t(s_{t-1}) = \max\left\{ \max_{a \in \mathcal{A}} Q_t(s_{t-1}, a),\ V_{t-1}(s_{t-1}) \right\},$$

  and $V_t(s) = V_{t-1}(s)$ for all $s \neq s_{t-1}$.
Output: Q-function estimate Q_T, value function estimate V_T, last state s_T.

be much less volatile than the counterpart $V_{t-1}(s_t)$ in (3.1), thus resulting in substantial variance reduction and hence accelerated convergence. It remains to develop a plausible approach that produces such reliable "reference" value function estimates.

An epoch-based paradigm. The proposed algorithm proceeds in an epoch-based manner ($K = \lfloor \log_4(3T/4) \rfloor$ epochs in total). In the k-th epoch, we use the value function estimate at the end of the previous epoch as the reference function estimate $\overline{V}$; the number of samples used to construct the empirical transition kernel and the number of samples employed to run the updates (4.1) are denoted respectively by T_k^{ref} and T_k, both of which are chosen to grow exponentially with the epoch number k (more specifically, we shall choose $T_k^{\mathsf{ref}} = 4^{k-1}$ and $T_k = 3 \cdot 4^{k-1}$). Such choices allow one to ensure that: (i) the estimation error keeps improving over epochs; and (ii) the samples used in the latest epoch always account for roughly 3/4 of the total sample size used so far, thus mitigating inefficient use of samples despite the lack of sample reuse.

4.2 Theoretical guarantees

Armed with the variance reduction idea, we are able to further improve the sample complexity in terms of the dependency on $\frac{1}{1-\gamma}$, as stated below. The proof can be found in Appendix D.

Theorem 2. *Suppose that Assumptions 1 and 2 hold, and recall that T is the total number of samples. Assume that $1/2 \leq \gamma < 1$. Then with probability exceeding $1-\delta$, the policy $\widehat{\pi}$ returned by Algorithm 2* *satisfies*

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) \lesssim \sqrt{\frac{SC^\star\iota}{T\left(1-\gamma\right)^3}} + \frac{SC^\star\iota^4}{T\left(1-\gamma\right)^4} + \frac{St_{\mathsf{mix}}C^\star\iota}{T\left(1-\gamma\right)^2} + \frac{t_{\mathsf{mix}}C^\star\iota^2}{T\left(1-\gamma\right)^3},$$

where $\iota := \log\frac{ST}{\delta}$.

Theorem 2 asserts that the sample size needed for Algorithm 2 to achieve ε-accuracy is at most

$$\widetilde{O}\left(\frac{SC^\star}{\left(1-\gamma\right)^3\varepsilon^2} + \frac{SC^\star}{\left(1-\gamma\right)^4\varepsilon} + \frac{SC^\star t_{\mathsf{mix}}}{\left(1-\gamma\right)^2\varepsilon} + \frac{t_{\mathsf{mix}}C^\star}{\left(1-\gamma\right)^3\varepsilon}\right). \tag{4.2}$$

In particular, if the accuracy level $\varepsilon \leq \min\left\{1-\gamma, \frac{S}{t_{\mathsf{mix}}}, \frac{1}{(1-\gamma)t_{\mathsf{mix}}}\right\}$, then the sample complexity of Algorithm 2 simplifies to

$$\widetilde{O}\left(\frac{SC^\star}{\left(1-\gamma\right)^3\varepsilon^2}\right). \tag{4.3}$$

This bound is essentially unimprovable; in fact, even for the simpler i.i.d. sampling mechanism described in Remark 2 (which can be viewed as a sample trajectory with $t_{\mathsf{mix}} = 1$), a lower bound has been established by Rashidinejad et al. (2021) that coincides with (4.3) when $C^\star = O(1)$. All this confirms the efficacy of the pessimism principle in conjunction with variance reduction when running model-free algorithms.

Let us take a moment to compare our setting and results with several offline RL papers that are most relevant to our work. Xie et al. (2021) (resp. the concurrent paper Shi et al. (2022)) proposed a model-based algorithm called PEVI-ADV (resp. a model-free algorithm called LCB-Q-ADV) to learn the optimal policy from a collection of independent episodes of offline samples, both of which achieve optimal sample complexity $\widetilde{O}(\varepsilon^{-2}H^4SC^\star)$ for finite-horizon MDPs with *nonstationary* transition kernels. In comparison, the current paper focuses on asynchronous Q-learning in a *stationary* discounted infinite-horizon MDP with data taking the form of a Markovian sample trajectory induced by a behavior policy, which is drastically different from and technically more challenging (e.g. i.i.d. data vs. Markovian non-i.i.d. data) than Xie et al. (2021) and the concurrent work Shi et al. (2022). In addition, the problem studied in Rashidinejad et al. (2021) can be viewed as a special case of our paper, given the i.i.d. nature of the sampling mechanism assumed therein. Rashidinejad et al. (2021) derived a sample complexity upper bound $\widetilde{O}\big(\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}\big)$, which is suboptimal by a factor of $(1-\gamma)^{-2}$ compared to our result in Theorem 2. Another concurrent paper Li et al. (2022) studied the model-based approach (i.e. the VI-LCB algorithm), which contrasts sharply with the model-free algorithms considered herein. It is als noteworthy that the results in Li et al. (2022) fall short of handling Markovian non-i.i.d. data.

In addition, the last two terms in our sample complexity bound (4.2) rely on the mixing time. To examine the necessity of the term $\frac{SC^\star t_{\mathsf{mix}}}{(1-\gamma)^2\varepsilon}$ in (4.2), we proceed to develop a minimax lower bound as follows; the proof is postponed to Appendix F.

Theorem 3. *Consider any* $S \geq 16$, $\gamma \in [1/2, 1)$, *and* $t_{\mathsf{mix}} \geq \frac{10}{1-\gamma}$. *Define the following set*

$$\mathscr{M} := \left\{ \{\mathcal{M}, \rho, \pi_{\mathsf{b}}\} \,\middle|\, |\mathcal{S}| = S, |\mathcal{A}| = 2, \rho \in \Delta\left(\mathcal{S}\right), \pi_{\mathsf{b}} : \mathcal{S} \to \Delta(\mathcal{A}), C^\star \leq 3, \right.$$
$$\left. \text{the mixing time of the Markov Chain induced by } \pi_{\mathsf{b}} \text{ and the MDP } \mathcal{M} \text{ is at most } 2t_{\mathsf{mix}} \right\}.$$

There exists some universal constants $c_1, c_2 > 0$ *such that, for any* $\varepsilon \in (0, \frac{c_1}{1-\gamma}]$, *if the sample size obeys*

$$T \leq c_2 \frac{St_{\mathsf{mix}}}{\left(1-\gamma\right)\varepsilon},$$

then with probability at least 0.5, *one has*

$$\inf_{\widehat{\pi}} \sup_{\mathscr{M}} \left\{V^\star\left(\rho\right) - V^{\widehat{\pi}}(\rho)\right\} \geq \varepsilon.$$

Here, the infimum is over all policy estimator $\widehat{\pi}$ *for the optimal policy based on the observed sample trajectory.*

In a nutshell, Theorem 3 reveals that when $S \geq \frac{1}{1-\gamma}$ — a scenario that arises frequently in real-world applications — the dependence of our sample complexity (4.2) on the mixing time is at most loose by a factor of $\frac{1}{1-\gamma}$. Unfortunately, our current theory falls short of characterizing the tight dependency of the mixing-time effect on $\frac{1}{1-\gamma}$; a new suite of analysis tools might be needed in order to fully settle this issue.

Before concluding this section, let us briefly highlight the key technical novelty of our analysis. The primary challenge in establishing Theorems 1 and 2 lies in the development of suitable concentration results concerning the sample trajectory of interest. The classical concentration results for Markov chains (e.g. in Paulin (2015)) do not readily address concentration of random variables in the form of

$$\sum_{t=1}^{T} \big(f_t(s_t, a_t) - \mathbb{E}_{(s,a)\sim\mu_b} \left[f_t\left(s, a\right) \right] \big), \tag{4.4}$$

where $f_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ might depend on the historical sample trajectory $\{(s_i, a_i)\}_{1 \leq i \leq t-1}$ (so not just (s_t, a_t)). To handle such complicated statistical dependency, we employ a decoupling argument (cf. Lemma 5) to reduce the concentration of (4.4) into the a number of $t_{\mathsf{mix}} \log(T/\delta)$ martingale concentrations. Then we apply Freedman's inequality (Li et al., 2021a, Theorem 3) to cope with each martingale concentration, thereby achieving sharp non-asymptotic concentration results for (4.4). The interested reader is referred to Step 2.4 in Appendix B.3 for details. It is also worth emphasizing that the proof techniques developed in the current paper are drastically different from the ones used in Li et al. (2023, 2021b). In particular, the techniques introduced in Li et al. (2023, 2021b) impose firm requirements that the data samples cover uniformly all state-action pairs, which are rarely satisfied under the partial-coverage scenario studied herein. What is more, minimax lower bounds related to the mixing time of a Markov chain in the RL context were previously rarely available. The analysis techniques developed here for establishing Theorem 3 might shed light on how to establish lower bounds for other settings that involve in-depth understanding of the mixing time.

5 Related works

Offline RL and pessimism. The principal of pessimism (or conservatism) in the face of uncertainty has recently been employed and studied extensively in offline RL (also called batch RL), e.g., Chang et al. (2021); Jin et al. (2021); Kidambi et al. (2020); Kumar et al. (2020); Li et al. (2022); Liu et al. (2020); Munos (2003, 2007); Rashidinejad et al. (2021); Uehara and Sun (2021); Xie et al. (2021); Yin et al. (2021a,b); Yin and Wang (2021); Yu et al. (2021a,b, 2020); Zhang et al. (2022). Among these prior works, Rashidinejad et al. (2021) studied offline RL for infinite-horizon MDPs when the offline data are i.i.d. samples drawn from some distribution μ satisfying the single policy concentrability condition. They showed that a model-based value iteration algorithm with LCB penalization achieves a sample complexity of $O(\frac{SC^\star}{(1-\gamma)^5\varepsilon^2})$, which is comparable to our bound for Algorithm 1 (see (3.5)) and is worse than our bound for Algorithm 2 (see (4.3)) by a factor of $\frac{1}{(1-\gamma)^2}$ (ignoring the $o(\varepsilon^{-2})$ term and logarithm factors). Note that the setting considered in Rashidinejad et al. (2021) is a special case of our setting by taking $t_{\mathsf{mix}} = 1$. In addition, Jin et al. (2021) proposed a pessimistic variant of the value iteration algorithm, which achieves appealing performance under the episodic linear MDP setting. Furthermore, the recent works Shi et al. (2022); Xie et al. (2021) proposed several pessimistic variants of RL algorithms for finite-horizon episodic MDPs. Focusing on offline RL with episodic data generated using some reference policy satisfying the single policy concentrability condition, these algorithms achieve a sample complexity of $\widetilde{O}(H^3SC^\star/\varepsilon^2)$. Note, however, that none of these algorithms accommodate the asynchronous case with a single Markovian trajectory.

Q-learning. There are at least two basic forms of Q-learning: the synchronous version and the asynchronous counterpart. Synchronous Q-learning typically assumes access to a simulator that generates independent samples for all state-action pairs, and attempts to update all entries of the Q-function estimates simultaneously (Beck and Srikant, 2012; Bowen et al., 2021; Chen et al., 2020; Even-Dar et al., 2003; Li et al., 2023; Wainwright, 2019a; Wang et al., 2021). The current paper studies the asynchronous form of Q-learning, which naturally arises when the data is a Markovian trajectory induced by a behavior policy (Chen et al., 2022, 2021; Even-Dar et al., 2003; Jaakkola et al., 1994; Li et al., 2023, 2021b; Qu and Wierman, 2020; Shah and Xie, 2018; Tsitsiklis, 1994; Xiong et al., 2020). However, most prior works focused on the case when the observed trajectory is able to cover all state-action pairs with sufficient frequency (Beck and Srikant, 2012; Chen et al., 2021; Even-Dar et al., 2003; Li et al., 2023, 2021b; Qu and Wierman, 2020). For instance, the recent work Qu and Wierman (2020) demonstrated that the sample complexity of asynchronous Q-learning is at most $\widetilde{O}\big(\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}\big)$, which was subsequently sharpened by Li et al. (2023) to $\widetilde{O}\big(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2}+\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{mix}}(1-\gamma)}\big)$. It is also worth noting that some variants of model-free algorithms (e.g., the variant coupled with upper confidence bounds) have proven effective for online exploratory RL (Bai et al., 2019; Dong et al., 2019; Jin et al., 2018; Li et al., 2021a; Ménard et al., 2021; Pazis et al., 2016; Strehl et al., 2006; Yang et al., 2021; Zhang et al., 2021); while online RL is beyond the scope of the current paper, the analysis framework therein based on the optimism principle shed light on our setting as well. In comparison to the model-based approach (Agarwal et al., 2020a,b; Azar et al., 2017; Li et al., 2020), model-free algorithms like Q-learning often incur lower memory and computational complexities.

Variance reduction. The idea of variance reduction first appeared in the stochastic optimization literature (Johnson and Zhang, 2013) and has been recently employed in RL to speed up various algorithms (Du et al., 2017; Khamaru et al., 2021a,b; Li et al., 2021a,b; Shi et al., 2022; Sidford et al., 2018a,b; Wai et al., 2019; Wainwright, 2019b; Xu et al., 2019; Yang and Wang, 2019; Zhang et al., 2020, 2021). Among these works, Wainwright (2019b); Yang and Wang (2019) showed that in the synchronous case, variance-reduced Q-learning is minimax optimal, both in tabular MDPs and the ones with function approximation. Li et al. (2021b) showed that the sample complexity of variance-reduced asynchronous Q-learning algorithm scales as $\widetilde{O}(\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2})$ for small enough accuracy level ε, thereby matching the lower bound in the synchronous counterpart.

6 Discussion

In this paper, we have revisited the paradigm of asynchronous Q-learning, which was designed to accommodate Markovian sample trajectories. Noteworthily, all prior theory for asynchronous Q-learning becomes vacuous when the observed sample trajectory falls short of providing uniform coverage of all state-action pairs, even when the observed data is produced by an expert that intentionally leaves out suboptimal actions. To address this issue, we have designed two algorithms — asynchronous Q-learning algorithms with LCB penalization and its variance-reduced variant — based on the principle of pessimism in the face of uncertainty. The sample complexities of these two algorithms scale as $\widetilde{O}\big(\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}\big)$ and $\widetilde{O}\big(\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}\big)$, respectively, provided that the target accuracy level ε is sufficiently small; in particular, the latter one matches the lower bound established for the case with i.i.d. data and is hence unimprovable. Compared to prior literature, we have established the first theory that supports the use of pessimism principle despite the Markovian structure of data. Moving forward, there are numerous directions that are worthy of further exploration. For example, the dependency of our sample complexity on the mixing time scales as $\widetilde{O}\big(\frac{St_{\mathsf{mix}}C^\star}{(1-\gamma)^2\varepsilon}+\frac{t_{\mathsf{mix}}C^\star}{(1-\gamma)^2\varepsilon}\big)$, while the minimax lower bound we have developed scales as $\Omega\big(\frac{St_{\mathsf{mix}}}{(1-\gamma)\varepsilon}\big)$; it remains unclear what the optimal dependency on t_{mix} is, as well as how to achieve it. Additionally, the current work focuses solely on tabular MDPs; it would be of interest to extend the current analysis to accommodate reduced-dimensional function approximation. Going beyond offline RL, our analysis framework might shed light on how to improve the sample complexity analysis for discounted infinite-horizon MDPs in *online exploratory* RL (note that the state-of-the-art sample complexity bounds in this case (Zhang et al., 2021) remain highly suboptimal except for very small ε (i.e., $\varepsilon \leq \frac{(1-\gamma)^{14}}{S^2A^2}$)).

Acknowledgements

Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009 and CCF-1907661. J. Fan is supported in part by NSF DMS-2210833, DMS-2052926, DMS-2053832 and ONR grant N00014-22-1-2340. Y. Yan is supported in part by Charlotte Elizabeth Procter Honorific

Fellowship from Princeton University. Part of this work was done while Y. Yan, G. Li and Y. Chen were visiting the Simons Institute for the Theory of Computing.

A Notation

We now introduce several notation that will be used multiple times throughout this paper. For any positive integer n, we define $[n] \coloneqq \{1, \cdots, n\}$. For any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, define

$$P_{s,a} = P(\cdot \,|\, s, a) \in \mathbb{R}^{1 \times S}$$

to be the (s,a)-th row of a probability transition matrix $P \in \mathbb{R}^{SA \times S}$. For any $t \geq 0$, we define $P_t \in \mathbb{R}^{SA \times S}$ to be an empirical probability transition matrix such that

$$P_t\big(s' \,|\, s, a\big) = \begin{cases} 1, & \text{if } (s, a, s') = (s_{t-1}, a_{t-1}, s_t) \\ 0, & \text{otherwise} \end{cases} \tag{A.1}$$

for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. For any deterministic policy π, we introduce two probability transition kernels $P_\pi : \mathcal{S} \to \Delta(\mathcal{S})$ and $P^\pi : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S} \times \mathcal{A})$, defined in a way that

$$P_\pi(s' \,|\, s) = P\big(s' \,|\, s, \pi(s)\big) \tag{A.2a}$$

$$P^\pi\left(s', a' \,|\, s, a\right) = \begin{cases} P\left(s' \,|\, s, a\right), & \text{if } a' = \pi\left(s'\right) \\ 0, & \text{otherwise} \end{cases} \tag{A.2b}$$

for any $(s, a), (s', a') \in \mathcal{S} \times \mathcal{A}$. In addition, we define $\rho^{\pi^\star}$ to be a distribution on $\mathcal{S} \times \mathcal{A}$ such that

$$\rho^{\pi^\star}\left(s, a\right) = \begin{cases} \rho\left(s\right), & \text{if } a = \pi^\star\left(s\right), \\ 0, & \text{otherwise.} \end{cases} \tag{A.3}$$

For any two vectors $a = [a_i]_{i=1}^n \in \mathbb{R}^n$ and $b = [b_i]_{i=1}^n \in \mathbb{R}^n$, we define the Hadamard product $a \circ b = [a_i b_i]_{i=1}^n$, as well as the concise notation $a^2 = a \circ a$. We also use $a \leq b$ (resp. $a \geq b$) to denote $a_i \leq b_i$ (resp. $a_i \geq b_i$) for all $i \in [n]$. Moreover, for two vectors $a = [a_1, \cdots, a_n]$ and $b = [b_1, \cdots, b_n]^\top$, we abuse the notation by letting

$$\langle a, b \rangle = \sum_{i=1}^n a_i b_i$$

even when a is a row vector and b is a column vector. For any $s \in \mathcal{S}$, $a \in \mathcal{A}$ and any vector $V \in \mathbb{R}^S$, we define and denote

$$\mathsf{Var}_{s,a}(V) \coloneqq \mathsf{Var}_{s' \sim P(\cdot \,|\, s,a)}\big(V(s')\big) = P_{s,a}\left(V^2\right) - (P_{s,a}V)^2.$$

We let $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ to denote $|f(n)| \leq C g(n)$ for some constant $C > 0$ when n is sufficiently large; we use $f(n) \gtrsim g(n)$ to indicate that $f(n) \geq C|g(n)|$ for some constant $C > 0$ when n is sufficiently large; and we let $f(n) \asymp g(n)$ represent the condition that $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold simultaneously. Throughout this paper, we define $0/0 = 0$. For any sequence $\{a_i\}_{i=n_1}^{n_2}$ and two integers m_1 and m_2, we define

$$\sum_{i=m_1}^{m_2} a_i = \begin{cases} \sum_{i=\max\{n_1, m_1\}}^{\min\{n_2, m_2\}} a_i, & \text{if } \max\{n_1, m_1\} \leq \min\{n_2, m_2\}, \\ 0, & \text{else.} \end{cases}$$

B Analysis for Q-learning with LCB penalization (Theorem 1)

In this section, we present the proof of Theorem 1, which consists of several steps to be detailed below.

B.1 Preliminary facts and additional notation

Before proceeding, we first introduce the following quantities regarding the learning rates:

$$\eta_0^t \coloneqq \prod_{j=1}^{t}(1-\eta_j) \qquad \text{and} \qquad \eta_i^t \coloneqq \begin{cases} \eta_i \prod_{j=i+1}^{t}(1-\eta_j), & \text{if } t > i, \\ \eta_i, & \text{if } t = i, \\ 0, & \text{if } t < i, \end{cases} \tag{B.1}$$

where we recall our choice $\eta_j = (H+1)/(H+j)$. We make note of the following results that have been established in prior works (e.g., Jin et al. (2018, Lemma 4.1) and Li et al. (2021a, Lemma 1)).

Lemma 1. *The learning rates satisfy the following properties.*

1. *For any integer $t \geq 1$, $\sum_{i=1}^{t} \eta_i^t = 1$ and $\eta_0^t = 0$.*
2. *For any integer $t \geq 1$ and any $1/2 \leq a \leq 1$,*

$$\frac{1}{t^a} \leq \sum_{i=1}^{t} \frac{1}{i^a}\eta_i^t \leq \frac{2}{t^a}.$$

3. *For any integer $t \geq 1$,*

$$\max_{i\in[t]} \eta_i^t \leq \frac{2H}{t} \qquad \textit{and} \qquad \sum_{i=1}^{t}\left(\eta_i^t\right)^2 \leq \frac{2H}{t}.$$

4. *For any integer $i \geq 1$,*

$$\sum_{t=i}^{\infty} \eta_i^t = 1 + \frac{1}{H}.$$

For any iteration $t \leq T$, we remind the reader that n_t represents the number of times (s,a) has been visited prior to iteration t (see Algorithm 1). For notational simplicity, let $n = n_t(s,a)$ when it is clear from the context, and suppose that (s,a) has been visited during the iterations $k_1 < \cdots < k_n < t$. We also find it convenient to define the (deterministic) policy estimate $\pi_t : \mathcal{S} \to \mathcal{A}$ recursively as follows

$$\pi_t(s) \coloneqq \begin{cases} \arg\max_{a\in\mathcal{A}} Q_t(s_{t-1}, a), & \text{if } s = s_{t-1} \text{ and } V_t(s) > V_{t-1}(s), \\ \pi_{t-1}(s), & \text{otherwise.} \end{cases} \tag{B.2}$$

If there are multiple $a \in \mathcal{A}$ that maximize $Q_t(s_{t-1}, a)$, we can pick any of these actions.

The following lemma provides a useful upper bound on $Q^\star - Q_t$, and in the meantime, justifies that the value function estimate V_t is always a pessimistic view of V^{π_t} (and hence $V^\star$). The proof of this lemma is postponed to Appendix C.1.

Lemma 2. *With probability exceeding $1-\delta$, for all $s \in \mathcal{S}$ and $t \in [T]$, it holds that*

$$Q^\star\big(s, \pi^\star(s)\big) - Q_t\big(s, \pi^\star(s)\big) \leq \gamma \sum_{i=1}^{n} \eta_i^n P_{s,\pi^\star(s)}\big(V^\star - V_{k_i}\big) + \beta_n\big(s, \pi^\star(s)\big),$$

where $n = n_t(s, \pi^\star(s))$ and we define

$$\beta_n\big(s, \pi^\star(s)\big) \equiv \beta_n \coloneqq 3C_{\mathsf{b}}\sqrt{\frac{H\iota}{n(1-\gamma)^2}};$$

in addition, we also have

$$V_t(s) \leq V^{\pi_t}(s) \leq V^\star(s), \qquad \forall s \in \mathcal{S}.$$

Next, let us define two disjoint sets of state-action pairs, divided based on the associated occupancy probability induced by the behavior policy:

$$\mathcal{I} \coloneqq \left\{ \left(s, \pi^\star(s)\right) \mid s \in \mathcal{S}, \mu_{\mathsf{b}}\left(s, \pi^\star(s)\right) \geq \frac{\delta}{ST} \right\}, \tag{B.3a}$$

$$\mathcal{I}^c \coloneqq \left\{ \left(s, \pi^\star(s)\right) \mid s \in \mathcal{S}, \mu_{\mathsf{b}}\left(s, \pi^\star(s)\right) < \frac{\delta}{ST} \right\}. \tag{B.3b}$$

It turns out that the state-action pairs in $\mathcal{I}^c$ are rarely visited, as formalized by the following lemma. The proof is deferred to Appendix C.2.

Lemma 3. *With probability exceeding* $1-\delta$, *we have*

$$\mathcal{I}^c \cap \left\{(s_t, a_t)\right\}_{t=t_{\mathsf{mix}}(\delta)}^T = \varnothing.$$

B.2 Step 1: error decomposition

Before proceeding, let us introduce several quantities that will play an important role in our analysis:

$$\begin{aligned}
\alpha_j &\coloneqq \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \sum_{t=1}^T \left\langle \rho(P_{\pi^\star})^j, V^\star - V_t \right\rangle, \\
\theta_j &\coloneqq \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \sum_{t=1}^T \sum_{s\in\mathcal{S}} \left[\rho(P_{\pi^\star})^j\right]\left(s, \pi^\star(s)\right) \min\left\{\beta_{n_t(s,\pi^\star(s))}\left(s, \pi^\star(s)\right), \frac{1}{1-\gamma}\right\}, \\
\xi_j &\coloneqq \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \left\langle \rho(P_{\pi^\star})^j, V^\star - V_t \right\rangle + \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^{j+1} \left\langle \rho(P_{\pi^\star})^{j+1}, V^\star - V_0 \right\rangle, \\
\psi_j &\coloneqq \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \sum_{t=t_{\mathsf{mix}}(\delta)}^T \Bigg[\sum_{s\in\mathcal{S}, a\in\mathcal{A}} \left[\rho^{\pi^\star}(P^{\pi^\star})^j\right](s,a) \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} P_{s,a}\left(V^\star - V_{k_i(s,a)}\right) \\
&\qquad - \left(1+\frac{1}{H}\right) \frac{\left[\rho^{\pi^\star}(P^{\pi^\star})^j\right](s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)} \sum_{i=1}^{n_t(s_t,a_t)} \eta_i^{n_t(s_t,a_t)} P_{s_t,a_t}\left(V^\star - V_{k_i(s_t,a_t)}\right)\Bigg], \\
\phi_j &\coloneqq \gamma^{j+1}\left(1+\frac{1}{H}\right)^{3j+2} \sum_{t=0}^T \mathbb{1}_{(s_t,a_t)\in\mathcal{I}} \Bigg[\frac{\left[\rho^{\pi^\star}(P^{\pi^\star})^j\right](s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)} P_{s_t,a_t}\left(V^\star - V_t\right) \\
&\qquad - \left(1+\frac{1}{H}\right) \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \left[\rho^{\pi^\star}(P^{\pi^\star})^j\right](s,a)\, P_{s,a}\left(V^\star - V_t\right)\Bigg],
\end{aligned}$$

where we recall the definition of $\mathcal{I}_1$ in (B.3).

Let us begin with the following basic inequality:

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) = \left\langle \rho, V^\star - V^{\widehat{\pi}} \right\rangle \overset{\text{(i)}}{\leq} \langle \rho, V^\star - V_T\rangle \overset{\text{(ii)}}{\leq} \frac{1}{T}\sum_{t=1}^T \langle \rho, V^\star - V_t\rangle \overset{\text{(iii)}}{=} \frac{1}{T}\alpha_0. \tag{B.4}$$

Here, (i) holds true according to Lemma 2; (ii) follows from the monotonicity of V_t in t (by construction); and (iii) follows simply from the definition of α_0. We then turn attention to bounding α_0, towards which we observe that

$$\begin{aligned}
\alpha_0 &= \sum_{t=1}^{t_{\mathsf{mix}}(\delta)-1} \langle \rho, V^\star - V_t\rangle + \sum_{t=t_{\mathsf{mix}}(\delta)}^T \sum_{s\in\mathcal{S}} \rho(s) \min\left\{Q^\star\left(s, \pi^\star(s)\right) - V_t(s), \frac{1}{1-\gamma}\right\} \\
&\leq \sum_{t=1}^{t_{\mathsf{mix}}(\delta)-1} \langle \rho, V^\star - V_t\rangle + \sum_{t=t_{\mathsf{mix}}(\delta)}^T \sum_{s\in\mathcal{S}} \rho(s) \min\left\{Q^\star\left(s, \pi^\star(s)\right) - Q_t\left(s, \pi^\star(s)\right), \frac{1}{1-\gamma}\right\}
\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t\rangle + \underbrace{\gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S}} \rho\left(s\right) \sum_{i=1}^{n_t(s,\pi^\star(s))} \eta_i^{n_t(s,\pi^\star(s))} P_{s,\pi^\star(s)}\left(V^\star - V_{k_i}\right)}_{=:\,\zeta} \\
&\quad + \underbrace{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \rho\left(s\right) \min\left\{\beta_{n_t(s,\pi^\star(s))}\big(s,\pi^\star(s)\big), \frac{1}{1-\gamma}\right\}}_{=\,\theta_0}.
\end{aligned}
$$

Here, the first identity holds since $V^\star(s) = Q^\star\big(s,\pi^\star(s)\big)$ and $0 \le V^\star(s) - V_t(s) \le 1/(1-\gamma)$ for all $s \in \mathcal{S}$, the second line relies on the fact that $V_t(s) \ge \max_a Q_t(s,a) \ge Q_t(s,\pi^\star(s))$, while the last line invokes Lemma 2. With probability exceeding $1-\delta$, the first term ζ can be upper bounded by

$$
\begin{aligned}
\zeta &\le \gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S}} \rho\left(s\right) \sum_{i=1}^{n_t(s,\pi^\star(s))} \eta_i^{n_t(s,\pi^\star(s))} P_{s,\pi^\star(s)}\left(V^\star - V_{k_i(s,\pi^\star(s))}\right) \\
&= \gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu_{\mathsf{b}}\left(s,a\right) \frac{\rho^{\pi^\star}\left(s,a\right)}{\mu_{\mathsf{b}}\left(s,a\right)} \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} P_{s,\pi^\star(s)}\left(V^\star - V_{k_i}\right) \\
&\overset{\text{(i)}}{\le} \gamma\left(1+\frac{1}{H}\right) \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \mathbb{1}\{(s_t,a_t)\in\mathcal{I}\} \frac{\rho^{\pi^\star}\left(s_t,a_t\right)}{\mu_{\mathsf{b}}\left(s_t,a_t\right)} \sum_{i=1}^{n_t(s_t,a_t)} \eta_i^{n_t(s_t,a_t)} P_{s_t,a_t}\left(V^\star - V_{k_i(s_t,a_t)}\right) + \psi_0 \\
&\overset{\text{(ii)}}{=} \gamma\left(1+\frac{1}{H}\right) \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \mathbb{1}\{(s_t,a_t)\in\mathcal{I}\} \frac{\rho^{\pi^\star}\left(s_t,a_t\right)}{\mu_{\mathsf{b}}\left(s_t,a_t\right)} \left(\sum_{j=n_t(s_t,a_t)}^{n_T(s_t,a_t)} \eta^j_{n_t(s_t,a_t)}\right) P_{s_t,a_t}\left(V^\star - V_t\right) + \psi_0 \\
&\overset{\text{(iii)}}{\le} \gamma\left(1+\frac{1}{H}\right)^2 \sum_{t=0}^{T} \mathbb{1}\{(s_t,a_t)\in\mathcal{I}\} \frac{\rho^{\pi^\star}\left(s_t,a_t\right)}{\mu_{\mathsf{b}}\left(s_t,a_t\right)} P_{s_t,a_t}\left(V^\star - V_t\right) + \psi_0 \\
&= \gamma\left(1+\frac{1}{H}\right)^3 \sum_{t=0}^{T} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \rho^{\pi^\star}\left(s,a\right) P_{s,a}\left(V^\star - V_t\right) + \psi_0 + \phi_0 \\
&= \gamma\left(1+\frac{1}{H}\right)^3 \sum_{t=0}^{T} \langle \rho P_{\pi^\star}, V^\star - V_t\rangle + \psi_0 + \phi_0 \\
&\le \alpha_1 + \psi_0 + \phi_0 + \gamma\left(1+\frac{1}{H}\right)^3 \langle \rho P_{\pi^\star}, V^\star - V_0\rangle,
\end{aligned}
$$

where we remind the reader of our notation $\rho^{\pi^\star}$ in (A.3). Here, (i) is valid (i.e., $\rho(s_t,a_t)/\mu_{\mathsf{b}}(s,a)$ is well defined for $t \ge t_{\mathsf{mix}}(\delta)$) due to Lemma 3; (ii) holds by grouping the terms in the previous line; and (iii) utilizes Lemma 1 and the property that $V^\star \ge V_t$ (cf. Lemma 2). Therefore, we arrive at

$$
\begin{aligned}
\alpha_0 &\le \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t\rangle + \zeta + \theta_0 \\
&\le \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t\rangle + \alpha_1 + \psi_0 + \phi_0 + \gamma\left(1+\frac{1}{H}\right)^3 \langle \rho P_{\pi^\star}, V^\star - V_0\rangle + \theta_0 \\
&= \alpha_1 + \xi_0 + \theta_0 + \psi_0 + \phi_0,
\end{aligned}
$$

where we have used the definition of ξ_0. Repeat the same argument to reach

$$
\alpha_j \le \alpha_{j+1} + \xi_j + \theta_j + \psi_j + \phi_j
$$

for all $j \geq 1$. This in turn allows us to conclude that

$$\alpha_0 \leq \underbrace{\limsup_{j\to\infty} \alpha_j}_{=:\,\alpha} + \underbrace{\sum_{j=0}^{\infty} \xi_j}_{=:\,\xi} + \underbrace{\sum_{j=0}^{\infty} \theta_j}_{=:\,\theta} + \underbrace{\sum_{j=0}^{\infty} \psi_j}_{=:\,\psi} + \underbrace{\sum_{j=0}^{\infty} \phi_j}_{=:\,\phi} . \tag{B.5}$$

We will then bound the terms α, ξ, θ, ψ and ϕ separately in the subsequent steps. Before continuing, we make note of a useful result.

Lemma 4. *Recall that* $H = \left\lceil \frac{4}{1-\gamma} \log \frac{ST}{\delta} \right\rceil$ *for some* $0 < \delta < 1$. *For any vector with non-negative entries* $V \in \mathbb{R}^d$ *, we have*

$$\sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \left\langle \rho (P_{\pi^\star})^j, V \right\rangle \lesssim \frac{1}{1-\gamma} \left\langle d_\rho^\star, V \right\rangle + \frac{\delta}{ST^4 (1-\gamma)} \|V\|_\infty . \tag{B.6}$$

Proof. See Appendix C.3. ☐

B.3 Step 2: bounding each term in (B.5)

Step 2.1: bounding α. It is first observed that

$$\alpha = \limsup_{j\to\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \sum_{t=1}^{T} \left\langle \rho (P_{\pi^\star})^j, V^\star - V_t \right\rangle \overset{\text{(i)}}{\leq} \frac{T}{1-\gamma} \limsup_{k\to\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^k \overset{\text{(ii)}}{=} 0.$$

Here, (i) is valid since $\rho (P_{\pi^\star})^j$ is a probability distribution over $\mathcal{S}$ and $0 \leq V^\star - V_t \leq 1/(1-\gamma)$ holds for all $1 \leq t \leq T$; (ii) holds since

$$\gamma \left(1 + \frac{1}{H} \right)^3 \leq \gamma \left(1 + \frac{1-\gamma}{4} \right)^2 < 1 \tag{B.7}$$

for all $\gamma < 1$.

Step 2.2: bounding ξ. By utilizing (B.6) and (B.7), we can demonstrate that

$$\begin{aligned}
\xi &= \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \left\{ \sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \left\langle \rho P_{\pi^\star}^j, V^\star - V_t \right\rangle \right\} + \sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^{j+1} \left\langle \rho (P_{\pi^\star})^{j+1}, V^\star - V_0 \right\rangle \\
&\lesssim \frac{1}{1-\gamma} \sum_{t=0}^{t_{\mathsf{mix}}(\delta)} \left\langle d_\rho^\star, V^\star - V_t \right\rangle + \frac{1}{ST^4 (1-\gamma)} \frac{t_{\mathsf{mix}}(\delta) + 1}{1-\gamma} \\
&\lesssim \frac{t_{\mathsf{mix}}(\delta)}{(1-\gamma)^2} + \frac{t_{\mathsf{mix}}(\delta)}{T^4 (1-\gamma)^2} \\
&\lesssim \frac{t_{\mathsf{mix}}}{(1-\gamma)^2} \log \frac{1}{\delta} + \frac{t_{\mathsf{mix}}}{T^4 (1-\gamma)^2} \log \frac{1}{\delta}.
\end{aligned}$$

Here, the second line holds due to (B.6) and the basic fact $0 \leq V^\star(s) - V_t(s) \leq \frac{1}{1-\gamma}$, the penultimate line makes use of the fact $\|V^\star - V_t\|_\infty \leq \frac{1}{1-\gamma}$ once again, whereas the last line holds since $t_{\mathsf{mix}}(\delta) \lesssim t_{\mathsf{mix}} \log \frac{1}{\delta}$.

Step 2.3: bounding θ. When it comes to θ, we can deduce that

$$\theta = \sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H} \right)^3 \right]^j \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \left[\rho (P_{\pi^\star})^j \right] (s) \min \left\{ \beta_{n_t(s, \pi^\star(s))}, \frac{1}{1-\gamma} \right\}$$

$$
\begin{aligned}
&= \sum_{t=1}^{T}\sum_{j=0}^{\infty}\left[\gamma\left(1+\frac{1}{H}\right)^{3}\right]^{j}\sum_{s\in\mathcal{S}}\left[\rho(P_{\pi^\star})^{j}\right](s)\min\left\{\beta_{n_t(s,\pi^\star(s))},\frac{1}{1-\gamma}\right\}\\
&\overset{\text{(i)}}{\lesssim} \frac{1}{1-\gamma}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}d_{\rho}^{\star}(s)\min\left\{\beta_{n_t(s,\pi^\star(s))},\frac{1}{1-\gamma}\right\}+\frac{1}{ST^4(1-\gamma)}\frac{T}{1-\gamma}\\
&\lesssim \sum_{s\in\mathcal{S}}\sum_{t=1}^{t_{\mathsf{burn\text{-}in}}(s)}\frac{d_{\rho}^{\star}(s)}{(1-\gamma)^2}+\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)+1}^{T}d_{\rho}^{\star}(s)\sqrt{\frac{H\iota}{n_t\big(s,\pi^\star(s)\big)(1-\gamma)^4}}+\frac{1}{T^3(1-\gamma)^2}\\
&\overset{\text{(ii)}}{\asymp} \sum_{s\in\mathcal{S}}\frac{d_{\rho}^{\star}(s)}{\mu_{\mathsf{b}}(s,\pi^\star(s))}\frac{t_{\mathsf{mix}}\iota}{(1-\gamma)^2}+\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)+1}^{T}d_{\rho}^{\star}(s)\sqrt{\frac{H\iota}{t\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)(1-\gamma)^4}}+\frac{1}{T^3(1-\gamma)^2}\\
&\overset{\text{(iii)}}{\lesssim} \frac{C^\star St_{\mathsf{mix}}\iota}{(1-\gamma)^2}+\sum_{s\in\mathcal{S}}d_{\rho}^{\star}\big(s,\pi^\star(s)\big)\sqrt{\frac{HT\iota}{\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)(1-\gamma)^4}}+\frac{1}{T^3(1-\gamma)^2}\\
&\overset{\text{(iv)}}{\lesssim} \frac{C^\star St_{\mathsf{mix}}\iota}{(1-\gamma)^2}+\sqrt{\frac{C^\star HT\iota}{(1-\gamma)^4}}\sum_{s\in\mathcal{S}}\sqrt{d_{\rho}^{\star}\big(s,\pi^\star(s)\big)}\\
&\overset{\text{(v)}}{\lesssim} \frac{C^\star St_{\mathsf{mix}}\iota}{(1-\gamma)^2}+\sqrt{\frac{C^\star ST\iota^2}{(1-\gamma)^5}},
\end{aligned}
$$

where we define, for each $s\in\mathcal{S}$,

$$
t_{\mathsf{burn\text{-}in}}(s):=C_{\mathsf{burn\text{-}in}}\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)}\log\left(\frac{ST}{\delta}\right)
$$

for some sufficiently large constant $C_{\mathsf{burn\text{-}in}}>0$. Here, (i) relies on (B.6); (ii) utilizes Li et al. (2021b, Lemma 8); (iii) follows from the fact that

$$
\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\le 1+\int_{1}^{T}\frac{1}{\sqrt{x}}\mathrm{d}x=1+2\left(\sqrt{T}-1\right)\le 2\sqrt{T}; \tag{B.8}
$$

(iv) uses Assumption 2; and (v) invokes the Cauchy-Schwarz inequality and the fact that $\sum_s d_{\rho}^{\star}\big(s,\pi^\star(s)\big)=1$.

Step 2.4: bounding ψ. Recall that

$$
\begin{aligned}
\psi_j := \gamma\left[\gamma\left(1+\frac{1}{H}\right)^{3}\right]^{j}\sum_{t=t_{\mathsf{mix}}(\delta)}^{T}\Bigg[&\sum_{s\in\mathcal{S},a\in\mathcal{A}}\left[\rho^{\pi^\star}(P^{\pi^\star})^{j}\right](s,a)\sum_{i=1}^{n_t(s,a)}\eta_i^{n_t(s,a)}P_{s,a}\left(V^\star-V_{k_i(s,a)}\right)\\
&-\left(1+\frac{1}{H}\right)\frac{\left[\rho^{\pi^\star}(P^{\pi^\star})^{j}\right](s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)}\sum_{i=1}^{n_t(s_t,a_t)}\eta_i^{n_t(s_t,a_t)}P_{s_t,a_t}\left(V^\star-V_{k_i(s_t,a_t)}\right)\Bigg].
\end{aligned}
$$

In order to bound ψ, we make the observation that

$$
\begin{aligned}
\psi &= \sum_{j=0}^{\infty}\gamma\left[\gamma\left(1+\frac{1}{H}\right)^{3}\right]^{j}\sum_{t=t_{\mathsf{mix}}(\delta)}^{T}\Bigg[\sum_{s\in\mathcal{S},a\in\mathcal{A}}\left[\rho^{\pi^\star}(P^{\pi^\star})^{j}\right](s,a)\sum_{i=1}^{n_t(s,a)}\eta_i^{n_t(s,a)}P_{s,a}\left(V^\star-V_{k_i(s,a)}\right)\\
&\qquad -\left(1+\frac{1}{H}\right)\frac{\left[\rho^{\pi^\star}(P^{\pi^\star})^{j}\right](s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)}\sum_{i=1}^{n_t(s_t,a_t)}\eta_i^{n_t(s_t,a_t)}P_{s_t,a_t}\left(V^\star-V_{k_i(s_t,a_t)}\right)\Bigg]\\
&= \sum_{t=t_{\mathsf{mix}}(\delta)}^{T}\Bigg[\sum_{s\in\mathcal{S},a\in\mathcal{A}}\widetilde{d}(s,a)\sum_{i=1}^{n_t(s,a)}\eta_i^{n_t(s,a)}P_{s,a}\left(V^\star-V_{k_i(s,a)}\right)
\end{aligned}
$$

$$-\left(1+\frac{1}{H}\right)\frac{\widetilde{d}\left(s_t,a_t\right)}{\mu_{\mathsf{b}}\left(s_t,a_t\right)}\sum_{i=1}^{n_t(s_t,a_t)}\eta_i^{n_t(s_t,a_t)}P_{s_t,a_t}\left(V^\star-V_{k_i(s_t,a_t)}\right)\Bigg].$$

Here, $\widetilde{d}(\cdot,\cdot)$ is defined such that

$$\widetilde{d}\left(s,a\right):=\sum_{j=0}^{\infty}\gamma\left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j\left[\rho^{\pi^\star}(P^{\pi^\star})^j\right](s,a)$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. For any $t_{\mathsf{mix}}(\delta)\leq t\leq T$ and any $(s,a)\in\mathcal{I}$, let us define

$$f_t\left(s,a\right)=\frac{\widetilde{d}\left(s,a\right)}{\mu_{\mathsf{b}}\left(s,a\right)}\sum_{i=1}^{n_t(s,a)}\eta_i^{n_t(s,a)}P_{s,a}\left(V^\star-V_{k_i(s,a)}\right),$$

allowing us to rewrite

$$\psi=\sum_{t=t_{\mathsf{mix}}(\delta)}^{T}\left\{\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_t\left(s,a\right)\right]-\left(1+\frac{1}{H}\right)f_t\left(s_t,a_t\right)\right\}. \tag{B.9}$$

Let us take a moment to look at some properties of f_t. It is straightforward to check that

(i) when $a\neq\pi^\star(s)$, one has $f_t(s,a)=0$;

(ii) $f_t(s,a)$ is monotonically decreasing in t.

The latter property follows from the non-decreasing property of V_t in t (by construction), and that $\eta_i^{n_t(s,a)}$ is decreasing in t, as well as $\sum_{i=1}^{n_t(s,a)}\eta_i^{n_t(s,a)}=1$ (cf. Lemma 1). On the other hand, when $a=\pi^\star(s)$ and $(s,a)\in\mathcal{I}$, we can invoke (B.6) to arrive at

$$\widetilde{d}\big(s,\pi^\star(s)\big)\lesssim\frac{1}{1-\gamma}d_\rho^\star(s)+\frac{\delta}{ST^4(1-\gamma)}, \tag{B.10}$$

and consequently,

$$\begin{aligned}
f_t\left(s,a\right)&\lesssim\left\{\frac{1}{1-\gamma}\frac{d_\rho^\star\left(s,a\right)}{\mu_{\mathsf{b}}\left(s,a\right)}+\frac{\delta}{ST^4\left(1-\gamma\right)}\frac{1}{\mu_{\mathsf{b}}\left(s,a\right)}\right\}\sum_{i=1}^{n_t(s,a)}\eta_i^{n_t(s,a)}P_{s,a}\left(V^\star-V_{k_i(s,a)}\right)\\
&\lesssim\frac{1}{1-\gamma}\frac{d_\rho^\star\left(s,a\right)}{\mu_{\mathsf{b}}\left(s,a\right)}+\frac{\delta}{ST^4\left(1-\gamma\right)}\frac{1}{\mu_{\mathsf{b}}\left(s,a\right)}\\
&\leq\frac{C^\star}{\left(1-\gamma\right)^2}+\frac{\delta}{ST^4\left(1-\gamma\right)}\frac{ST}{\delta}\\
&\leq\frac{c_{10}C^\star}{\left(1-\gamma\right)^2}:=C_f.
\end{aligned} \tag{B.11}$$

for some constant $c_{10}\geq1$. Here, the second line follows from Assumption 2, the properties that $\sum_{i=1}^{n_t(s,a)}\eta_i^{n_t(s,a)}=1$, $0\leq V^\star(s)-V_t(s)\leq1/(1-\gamma)$ for all $0\leq t\leq T$; the third line is valid since $\mu_{\mathsf{b}}(s,\pi^\star(s))\geq\delta/(ST)$ when $(s,\pi^\star(s))\in\mathcal{I}$.

We now proceed to bound (B.9). It is worth noting that both f_t and (s_t,a_t) depend on $s_0,a_0,s_1,\ldots,s_{t-1},a_{t-1}$. To handle such statistical dependency, we define

$$K:=\left\lfloor\frac{T}{\tau}\right\rfloor\qquad\text{where}\quad\tau:=t_{\mathsf{mix}}(\delta/T^2)\lesssim t_{\mathsf{mix}}\log\frac{T}{\delta}.$$

Armed with this notation, one can decompose

$$\psi=\sum_{t=1}^{\tau}\sum_{k=1}^{K-1}\left\{\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_{k\tau+t}\left(s,a\right)\right]-\left(1+\frac{1}{H}\right)f_{k\tau+t}\left(s_{k\tau+t},a_{k\tau+t}\right)\right\}$$

$$
\begin{aligned}
&+\left(\sum_{t=t_{\mathsf{mix}}(\delta)}^{\tau}+\sum_{t=K\tau+1}^{T}\right)\left\{\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_t\left(s,a\right)\right]-\left(1+\frac{1}{H}\right)f_t\left(s_t,a_t\right)\right\}\\
&=\underbrace{\sum_{i=1}^{\tau}\sum_{k=1}^{K-1}\left\{\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_{(k-1)\tau+i}\left(s,a\right)\right]-\left(1+\frac{1}{H}\right)f_{(k-1)\tau+i}\left(s_{k\tau+i},a_{k\tau+i}\right)\right\}}_{=:\kappa_1}\\
&\quad+\underbrace{\sum_{i=1}^{\tau}\sum_{k=1}^{K-1}\left\{\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_{k\tau+i}\left(s,a\right)\right]-\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_{(k-1)\tau+i}\left(s,a\right)\right]\right\}}_{=:\kappa_2}\\
&\quad+\underbrace{\left(1+\frac{1}{H}\right)\sum_{i=1}^{\tau}\sum_{k=1}^{K-1}\left[f_{(k-1)\tau+i}\left(s_{k\tau+i},a_{k\tau+i}\right)-f_{k\tau+i}\left(s_{k\tau+i},a_{k\tau+i}\right)\right]}_{=:\kappa_3}\\
&\quad+\underbrace{\left(\sum_{t=t_{\mathsf{mix}}(\delta)}^{\tau}+\sum_{t=K\tau+1}^{T}\right)\left\{\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_t\left(s,a\right)\right]-\left(1+\frac{1}{H}\right)f_t\left(s_t,a_t\right)\right\}}_{=:\kappa_4}.
\end{aligned}
$$

In what follows, we bound κ_1, κ_2, κ_3 and κ_4 respectively.

- The term κ_4 can be easily bounded using (B.11) as follows

$$
\kappa_4\le\left(\sum_{t=t_{\mathsf{mix}}(\delta)}^{\tau}+\sum_{t=K\tau+1}^{T}\right)\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_t\left(s,a\right)\right]\le 2\tau C_f\asymp\frac{C^\star t_{\mathsf{mix}}}{\left(1-\gamma\right)^2}\log\left(\frac{T}{\delta}\right).
$$

- With regards to κ_3, we make the observation that

$$
\begin{aligned}
\kappa_3&=\left(1+\frac{1}{H}\right)\sum_{t=\tau+1}^{K\tau}\left[f_{t-\tau}\left(s_t,a_t\right)-f_t\left(s_t,a_t\right)\right]\\
&\overset{\text{(i)}}{\le}\left(1+\frac{1}{H}\right)\sum_{t=\tau+1}^{K\tau}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\left[f_{t-\tau}\left(s,a\right)-f_t\left(s,a\right)\right]\\
&\overset{\text{(ii)}}{=}\left(1+\frac{1}{H}\right)\sum_{t=\tau+1}^{K\tau}\sum_{s\in\mathcal{S}}\left[f_{t-\tau}\big(s,\pi^\star(s)\big)-f_t\big(s,\pi^\star(s)\big)\right]\\
&=\left(1+\frac{1}{H}\right)\left\{\sum_{t=1}^{\tau}\sum_{s\in\mathcal{S}}f_t\big(s,\pi^\star(s)\big)-\sum_{t=(K-1)\tau+1}^{K\tau}\sum_{s\in\mathcal{S}}f_t\big(s,\pi^\star(s)\big)\right\}\\
&\le 2\left(1+\frac{1}{H}\right)\tau SC_f\asymp\frac{C^\star St_{\mathsf{mix}}}{\left(1-\gamma\right)^2}\log\left(\frac{T}{\delta}\right).
\end{aligned}
$$

  Here, (i) holds since $f_t(s,a)$ is monotonically decreasing in t; and (ii) holds since, by definition, $f(s,a)=0$ if $a\neq\pi^\star(s)$.

- Similarly, κ_2 can be bounded by

$$
\begin{aligned}
\kappa_2&=\sum_{t=\tau+1}^{K\tau}\left\{\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_t\left(s,a\right)\right]-\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_{t-\tau}(s,a)\right]\right\}\\
&=\sum_{t=(K-1)\tau+1}^{K\tau}\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_t\left(s,a\right)\right]-\sum_{t=1}^{\tau}\mathbb{E}_{(s,a)\sim\mu_{\mathsf{b}}}\left[f_t\left(s,a\right)\right]
\end{aligned}
$$

$$\lesssim \tau C_f \asymp \frac{C^\star t_{\mathsf{mix}}}{(1-\gamma)^2} \log\left(\frac{T}{\delta}\right).$$

- Finally, we turn attention to bounding κ_1. For each $1 \le i \le \tau$, we will bound

$$\xi_i \coloneqq \sum_{k=1}^{K-1} \left\{ \mathop{\mathbb{E}}_{(s,a)\sim\mu_{\mathsf{b}}} \left[f_{(k-1)\tau+i}(s,a) \right] - \left(1+\frac{1}{H}\right) f_{(k-1)\tau+i}(s_{k\tau+i}, a_{k\tau+i}) \right\}$$

respectively. We need the following lemma to decouple the complicated statistical dependency.

Lemma 5. *One can construct an auxiliary set of random variables* $\left\{\left(s_k^i, a_k^i\right) : 1 \le k \le K-1\right\}$ *satisfying*

$$\left\{ \left(s_k^i, a_k^i\right) : 1 \le k \le K-1 \right\} \overset{\text{i.i.d.}}{\sim} \mu_{\mathsf{b}}, \tag{B.12a}$$

$$\mathbb{P}\left\{ \left(s_k^i, a_k^i\right) = (s_{k\tau+i}, a_{k\tau+i}) \quad \textit{for all } 1 \le k \le K-1 \right\} \ge 1 - \frac{\delta}{T}, \tag{B.12b}$$

and

$$\left(s_k^i, a_k^i\right) \ \textit{is independent of} \left\{ (s_t, a_t) : 0 \le t \le (k-1)\tau + i \right\}. \tag{B.12c}$$

Proof. See Appendix C.4. ☐

With the above set of auxiliary random variables $\left\{\left(s_k^i, a_k^i\right) : 1 \le k \le K-1\right\}$ in place, one can obtain

$$\begin{aligned}
\xi_i &= \sum_{k=1}^{K-1} \left\{ \mathop{\mathbb{E}}_{(s,a)\sim\mu_{\mathsf{b}}} \left[f_{(k-1)\tau+i}(s,a) \right] - \left(1+\frac{1}{H}\right) f_{(k-1)\tau+i}\left(s_k^i, a_k^i\right) \right\} \\
&= -\left(1+\frac{1}{H}\right) \sum_{k=1}^{K-1} \left\{ f_{(k-1)\tau+i}\left(s_k^i, a_k^i\right) - \mathop{\mathbb{E}}_{(s,a)\sim\mu_{\mathsf{b}}} \left[f_{(k-1)\tau+i}(s,a) \right] \right\} - \frac{1}{H} \sum_{k=1}^{K-1} \mathop{\mathbb{E}}_{(s,a)\sim\mu_{\mathsf{b}}} \left[f_{(k-1)\tau+i}(s,a) \right]
\end{aligned}$$

with probability exceeding $1-\delta/T$. Recognizing the property (B.12c), we are ready to use the Freedman inequality (cf. Li et al. (2021a, Theorem 3)) to bound ξ_i. Introduce the random variable

$$X_k = f_{(k-1)\tau+i}(s_k^i, a_k^i) - \mathop{\mathbb{E}}_{(s,a)\sim\mu_{\mathsf{b}}} \left[f_{(k-1)\tau+i}(s,a) \right], \tag{B.13}$$

and define a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_{K-1}$ with

$$\mathcal{F}_{k-1} = \sigma\left\{ \left\{ \left(s_k^i, a_k^i\right) \right\}_{k=1}^{k-1}, \left\{ (s_t, a_t) \right\}_{t=0}^{(k-1)\tau+i} \right\} \quad \text{for} \quad 1 \le k \le K-1.$$

It is straightforward to verify that

$$|X_k| \le R \coloneqq C_f, \quad \mathbb{E}\left[X_k \mid \mathcal{F}_{k-1} \right] = 0 \quad \text{for all } 1 \le k \le K-1,$$

and

$$\begin{aligned}
W &\coloneqq \sum_{k=1}^{K-1} \mathbb{E}\left[X_k^2 \mid \mathcal{F}_{k-1} \right] \le \sum_{k=1}^{K-1} \mathbb{E}\left[f_{(k-1)\tau+i}^2(s_k^i, a_k^i) \mid \mathcal{F}_{k-1} \right] \\
&\le C_f \sum_{k=1}^{K-1} \mathbb{E}\left[f_{(k-1)\tau+i}(s_k^i, a_k^i) \mid \mathcal{F}_{k-1} \right] = C_f \sum_{k=1}^{K-1} \mathop{\mathbb{E}}_{(s,a)\sim\mu_{\mathsf{b}}} \left[f_{(k-1)\tau+i}(s,a) \right] && \text{(B.14)} \\
&\le C_f^2 K. && \text{(B.15)}
\end{aligned}$$

Invoke the Freedman inequality in Li et al. (2021a, Theorem 3) to show that for any integer $m \ge 1$,

$$\left| \sum_{k=1}^{K-1} X_k \right| \le \sqrt{8 \max\left\{ W, \frac{C_f^2 K}{2^m} \right\} \log \frac{2Tm}{\delta}} + \frac{4}{3} R \log \frac{2Tm}{\delta}$$

$$
\begin{aligned}
&\le \sqrt{8W \log \frac{2Tm}{\delta}} + \sqrt{8\frac{C_f^2 K}{2^m} \log \frac{2Tm}{\delta}} + \frac{4}{3} C_f \log \frac{2Tm}{\delta} \\
&\le \frac{1}{2HC_f} W + 4HC_f \log \frac{2Tm}{\delta} + C_f \sqrt{8 \frac{K}{2^m} \log \frac{2Tm}{\delta}} + \frac{4}{3} C_f \log \frac{2Tm}{\delta} \\
&= \frac{1}{2H} \sum_{k=1}^{K-1} \mathbb{E}_{(s,a)\sim\mu_\mathsf{b}} \left[f_{(k-1)\tau+i}(s,a) \right] + O\left(HC_f \log \frac{T}{\delta} \right)
\end{aligned}
$$

holds with probability exceeding $1-\delta/T$. Here, the penultimate line relies on the AM-GM inequality, whereas the last line holds by using (B.14) and taking $m \asymp \log K \lesssim \log T \lesssim T$. Consequently, we see that with probability exceeding $1-\delta/T$,

$$
\begin{aligned}
\xi_i &= -\left(1+\frac{1}{H}\right) \sum_{k=1}^{K-1} X_k - \frac{1}{H} \sum_{k=1}^{K-1} \mathop{\mathbb{E}}_{(s,a)\sim\mu_\mathsf{b}} \left[f_{(k-1)\tau+i}(s,a) \right] \\
&\le 2 \left| \sum_{k=1}^{K-1} X_k \right| - \frac{1}{H} \sum_{k=1}^{K-1} \mathop{\mathbb{E}}_{(s,a)\sim\mu_\mathsf{b}} \left[f_{(k-1)\tau+i}(s,a) \right] \\
&\lesssim HC_f \log \frac{T}{\delta} \lesssim \frac{C^\star \iota}{(1-\gamma)^3} \log \frac{T}{\delta}.
\end{aligned}
$$

As a result, with probability exceeding $1-\delta$ we can guarantee that

$$
\kappa_1 \le \sum_{i=1}^{\tau} \xi_i \lesssim \frac{C^\star \tau \iota}{(1-\gamma)^3} \log\left(\frac{T}{\delta}\right) \asymp \frac{C^\star t_{\mathsf{mix}} \iota}{(1-\gamma)^3} \log^2\left(\frac{T}{\delta}\right).
$$

The above bounds taken collectively allow us to conclude that

$$
\psi \le \kappa_1 + \kappa_2 + \kappa_3 + \kappa_4 \lesssim \frac{C^\star t_{\mathsf{mix}} \iota}{(1-\gamma)^3} \log^2\left(\frac{T}{\delta}\right) + \frac{C^\star S t_{\mathsf{mix}}}{(1-\gamma)^2} \log\left(\frac{T}{\delta}\right).
$$

Step 2.5: bounding ϕ. By replacing $f_t(s,a)$ in Step 4 with

$$
g_t(s,a) = P_{s,a}(V^\star - V_t),
$$

we can employ an analogous argument to show that ϕ admits the same bound as ψ, namely,

$$
\phi \lesssim \frac{C^\star t_{\mathsf{mix}} \iota}{(1-\gamma)^3} \log^2\left(\frac{T}{\delta}\right) + \frac{C^\star S t_{\mathsf{mix}}}{(1-\gamma)^2} \log\left(\frac{T}{\delta}\right).
$$

We omit this part for the sake of brevity.

B.4 Step 3: putting all pieces together

To finish up, taking the bounds on α, θ, ψ and ϕ collectively gives

$$
\begin{aligned}
\alpha_0 &\le \alpha + \xi + \theta + \psi + \phi \\
&\lesssim \frac{C^\star S t_{\mathsf{mix}} \iota}{(1-\gamma)^2} + \sqrt{\frac{C^\star S T \iota^2}{(1-\gamma)^5}} + \frac{C^\star t_{\mathsf{mix}} \iota}{(1-\gamma)^3} \log^2\left(\frac{T}{\delta}\right) + \frac{C^\star S t_{\mathsf{mix}}}{(1-\gamma)^2} \log\left(\frac{T}{\delta}\right) \\
&\asymp \sqrt{\frac{C^\star S T \iota^2}{(1-\gamma)^5}} + \frac{C^\star S t_{\mathsf{mix}} \iota}{(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}} \iota}{(1-\gamma)^3} \log^2\left(\frac{T}{\delta}\right).
\end{aligned}
$$

Consequently, we can invoke (B.4) to conclude that

$$
V^\star(\rho) - V^{\widehat{\pi}}(\rho) \le \frac{\alpha_0}{T} \lesssim \sqrt{\frac{C^\star S \iota^2}{T(1-\gamma)^5}} + \frac{C^\star S t_{\mathsf{mix}} \iota}{T(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}} \iota^2}{T(1-\gamma)^3}.
$$

C Auxiliary lemmas for Theorem 1

C.1 Proof of Lemma 2

Consider any given pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. For notational simplicity, we write $n = n_t(s,a)$, the total number of times that (s,a) has been visited prior to time t. We also set $k_0 = -1$, and let

$$k_i \coloneqq \min\Big\{\big\{0 \le k < T : k > k_{i-1}, (s_k, a_k) = (s,a)\big\}, T\Big\} \tag{C.1}$$

for each $1 \le i \le T$. Clearly, each k_i is a stopping time. In view of the update rule in Algorithm 1, we have

$$Q_t\left(s,a\right) = \sum_{i=1}^{n} \eta_i^n \Big\{ r\left(s,a\right) + \gamma V_{k_i}\left(s_{k_i+1}\right) - b_i\left(s,a\right) \Big\},$$

which together with the Bellman optimality equation $Q^\star = r + \gamma P V^\star$ gives

$$\begin{aligned}
\left(Q^\star - Q_t\right)\left(s,a\right) &= r\left(s,a\right) + \gamma P_{s,a} V^\star - \sum_{i=1}^{n} \eta_i^n \Big\{ r\left(s,a\right) + \gamma V_{k_i}\left(s_{k_i+1}\right) - b_i\left(s,a\right) \Big\} \\
&= \gamma P_{s,a} V^\star - \sum_{i=1}^{n} \eta_i^n \Big\{ \gamma V_{k_i}\left(s_{k_i+1}\right) - b_i\left(s,a\right) \Big\} \\
&= \sum_{i=1}^{n} \eta_i^n \gamma P_{s,a}\left(V^\star - V_{k_i}\right) + \sum_{i=1}^{n} \eta_i^n \gamma \Big(\big(P - P_{k_i}\big) V_{k_i} \Big)\left(s,a\right) + \sum_{i=1}^{n} \eta_i^n b_i\left(s,a\right),
\end{aligned} \tag{C.2}$$

where the last two lines are valid since $\sum_{i=1}^n \eta_i^n = 1$ (cf. Lemma 1).

From now on we only focus on the case where $a = \pi^\star(s)$. Define $\mathcal{F}_i$ to be the σ-field generated by $\{(s_i, a_i)\}_{i=0}^{k_i}$. It is straightforward to check that for any $1 \le \tau \le T$,

$$\Big\{ \mathbb{1}_{k_i < T} \Big(\big(P - P_{k_i}\big) V_{k_i} \Big) \big(s, \pi^\star(s)\big) \Big\}_{i=1}^{\tau}$$

is a martingale difference sequence with respect to $\{\mathcal{F}_i\}_{i \ge 0}$. Then, we can invoke the Azuma-Hoeffding inequality together with the basic bound $\|V_{k_i}\|_\infty \le \frac{1}{1-\gamma}$ to show that for any fixed $s \in \mathcal{S}$ and $\tau \in [T]$,

$$\begin{aligned}
\left| \sum_{i=1}^{\tau} \mathbb{1}_{k_i < T}\, \eta_i^\tau \Big(\big(P - P_{k_i}\big) V_{k_i} \Big) \big(s, \pi^\star(s)\big) \right| &\lesssim \frac{1}{1-\gamma} \sqrt{\sum_{i=1}^{\tau} \left(\eta_i^\tau\right)^2 \log \frac{ST}{\delta}} \\
&\lesssim \sqrt{\frac{H}{\tau\left(1-\gamma\right)^2} \log \frac{ST}{\delta}}
\end{aligned}$$

holds with probability exceeding $1 - \delta/(ST)$. Here, the last line utilizes Lemma 1. Taking the union bound over $\tau \le T$ allows us to replace τ with $n = n_t(s,a)$ in the above inequality, namely, for any fixed $s \in \mathcal{S}$ and $a \in \mathcal{A}$, with probability exceeding $1 - \delta/S$ we have

$$\left| \sum_{i=1}^{n} \eta_i^n \gamma \Big(\big(P - P_{k_i}\big) V_{k_i} \Big) \big(s, \pi^\star(s)\big) \right| \lesssim \sqrt{\frac{H\iota}{n\left(1-\gamma\right)^2}} \tag{C.3}$$

holds for all $n = n_t(s, \pi^\star(s))$ with $1 \le t \le T$. In view of Lemma 1, for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$ we know that

$$C_{\mathsf{b}} \sqrt{\frac{H\iota}{n_t(s,a)\left(1-\gamma\right)^2}} \le \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} b_i\left(s,a\right) \le 2C_{\mathsf{b}} \sqrt{\frac{H\iota}{n_t(s,a)\left(1-\gamma\right)^2}}. \tag{C.4}$$

Therefore, when C_{b} is sufficiently large, it follows that

$$\left(Q^\star - Q_t\right)\big(s, \pi^\star(s)\big) \le \gamma \sum_{i=1}^{n} \eta_i^n P_{s,\pi^\star(s)}\left(V^\star - V_{k_i}\right) + 3C_{\mathsf{b}} \sqrt{\frac{H\iota}{n\left(1-\gamma\right)^2}}.$$

Taking the union bound over $s \in \mathcal{S}$ and defining

$$\beta_n\big(s, \pi^\star(s)\big) \coloneqq 3C_{\mathsf{b}}\sqrt{\frac{H\iota}{n\left(1-\gamma\right)^2}},$$

we can conclude that with probability exceeding $1-\delta$,

$$\left(Q^\star - Q_t\right)\big(s, \pi^\star(s)\big) \le \gamma \sum_{i=1}^{n} \eta_i^n P_{s,\pi^\star(s)}\left(V^\star - V_{k_i}\right) + \beta_n\big(s, \pi^\star(s)\big)$$

for all $s \in \mathcal{S}$ and $t \in [T]$.

Additionally, observe that $V^\star \ge V^{\pi_t}$ holds trivially due to the optimality of $V^\star$. We are therefore left with showing $V^{\pi_t} \ge V_t$. Suppose for the moment that with probability exceeding $1-\delta$, for all $s \in \mathcal{S}$, $t \in [T]$ and $j \in [t]$, it holds that

$$\left(Q^{\pi_t} - Q_j\right)\big(s, \pi_t(s)\big) \ge \gamma P_{s,\pi_t(s)}\left(V^{\pi_t} - V_j\right) \mathbb{1}\left\{n_t\big(s, \pi_t(s)\big) \ge 1\right\}; \tag{C.5}$$

the proof of this claim (C.5) is deferred to Appendix C.1.1. As a consequence, for every $s \in \mathcal{S}$ and $t \in [T]$, there exists $j(t) \in [t]$ such that

$$\begin{aligned}\left(V^{\pi_t} - V_t\right)(s) &\overset{\text{(i)}}{=} Q^{\pi_t}\big(s, \pi_t(s)\big) - Q_{j(t)}\big(s, \pi_t(s)\big) \overset{\text{(ii)}}{=} Q^{\pi_t}\big(s, \pi_t(s)\big) - Q_{j(t)}\big(s, \pi_{j(t)}(s)\big) \\ &\overset{\text{(iii)}}{\ge} \min\left\{\gamma P_{s,\pi_t(s)}\left(V^{\pi_t} - V_{j(t)}\right), 0\right\} \overset{\text{(iv)}}{\ge} \min\left\{\gamma P_{s,\pi_{j(t)}(s)}\left(V^{\pi_t} - V_t\right), 0\right\}.\end{aligned}$$

Here, (i) and (ii) hold since the update rule of Algorithm 1 asserts that there must exist some $j(t) \le t$ such that $V_t(s) = V_{j(t)}(s) = Q_{j(t)}(s, \pi_{j(t)}(s))$ and $\pi_t(s) = \pi_{j(t)}(s)$; (iii) utilizes (C.5); and (iv) follows from the monotonicity of V_t in t (by construction). By setting

$$s_{\min} \coloneqq \arg\min_{s \in \mathcal{S}} \left(V^{\pi_t} - V_t\right)(s),$$

we can deduce that

$$\begin{aligned}\left(V^{\pi_t} - V_t\right)(s_{\min}) &\ge \min\left\{\gamma P_{s_{\min},\pi_{j(t)}(s_{\min})}\left(V^{\pi_t} - V_t\right), 0\right\} \\ &\ge \min\left\{\gamma \min_{s \in \mathcal{S}}\left(V^{\pi_t} - V_t\right)(s), 0\right\} \\ &= \min\left\{\gamma\left(V^{\pi_t} - V_t\right)(s_{\min}), 0\right\},\end{aligned}$$

which together with the assumption $0 < \gamma < 1$ immediately gives

$$\left(V^{\pi_t} - V_t\right)(s_{\min}) \ge 0.$$

Given that $\left(V^{\pi_t} - V_t\right)(s) \ge \left(V^{\pi_t} - V_t\right)(s_{\min})$ for every $s \in \mathcal{S}$, we conclude the proof.

C.1.1 Proof of inequality (C.5)

First of all, if $n_t\big(s, \pi_t(s)\big) = 0$, then for all $j \in [t]$, $Q_j\big(s, \pi_t(s)\big) = 0$ since it is never updated; therefore, (C.5) holds true. From now on, we shall only focus on the case when $n_t\big(s, \pi_t(s)\big) \ge 1$.

Consider any $s \in \mathcal{S}$, $t \in [T]$ and $j \in [t]$. For the moment, let us define $\{k_i\}_{i=1}^T$ w.r.t. the state-action pair $\big(s, \pi_t(s)\big)$ in the same way as (C.1). We can then repeat the argument in (C.2) to decompose

$$\begin{aligned}&\big(Q^{\pi_t} - Q_j\big)\big(s, \pi_t(s)\big) \\ &= \big(r + \gamma P V^{\pi_t}\big)\big(s, \pi_t(s)\big) - \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))}\Big\{r\big(s, \pi_t(s)\big) + \gamma V_{k_i}\left(s_{k_i+1}\right) - b_i\big(s, \pi_t(s)\big)\Big\}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} \gamma \Big\{ P_{s,\pi_t(s)} V^{\pi_t} - V_{k_i}(s_{k_i+1}) \Big\} + \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} b_i\big(s,\pi_t(s)\big) \\
&= \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} \gamma \Big\{ P_{s,\pi_t(s)} (V^{\pi_t} - V_{k_i}) + \Big((P - P_{k_i}) V_{k_i} \Big) \big(s,\pi_t(s)\big) \Big\} + \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} b_i\big(s,\pi_t(s)\big) \\
&\geq \left(\sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} \right) \gamma \min_{1\leq i\leq n} P_{s,\pi_t(s)} (V^{\pi_t} - V_{k_i}) + \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} \gamma \Big((P - P_{k_i}) V_{k_i} \Big) \big(s,\pi_t(s)\big) \\
&\quad + \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} b_i\big(s,\pi_t(s)\big) \\
&\geq \gamma P_{s,\pi_t(s)} (V^{\pi_t} - V_t) + \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} \gamma \Big((P - P_{k_i}) V_{k_i} \Big) \big(s,\pi_t(s)\big) + C_{\mathsf{b}} \sqrt{\frac{H\iota}{n_j\big(s,\pi_t(s)\big)(1-\gamma)^2}}.
\end{aligned}
$$

Here, the last inequality follows from (C.4), as well as the facts that $\sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} = 1$ (cf. Lemma 1) and that V_t is non-decreasing in t. It thus boils down to showing that for every $s \in \mathcal{S}$, $t \in [T]$ and $j \in [t]$,

$$
\sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} \gamma \Big((P - P_{k_i}) V_{k_i} \Big) \big(s,\pi_t(s)\big) \lesssim \sqrt{\frac{H\iota}{n_j\big(s,\pi_t(s)\big)(1-\gamma)^2}}. \tag{C.6}
$$

If this were true and if C_{b} is sufficiently large, then we could combine the above two inequalities to conclude the proof of (C.5).

We then prove the inequality (C.6). Notice that for all $(s,\pi_t(s))$ such that $n_t(s,\pi_t(s)) \geq 1$, it must appear at least once in the sample trajectory. Therefore it suffices to show that for all $0 \leq l < T$ and $t \in [T]$, it holds that

$$
\sum_{i=1}^{n_t(s_l,a_l)} \eta_i^{n_t(s_l,a_l)} \gamma \Big((P - P_{k_i}) V_{k_i} \Big) (s_l,a_l) \lesssim \sqrt{\frac{H\iota}{n_t(s_l,a_l)(1-\gamma)^2}},
$$

where we abuse the notation by defining $\{k_i\}_{i=1}^T$ for the state-action pair (s_l,a_l) in the same way as (C.1). Furthermore, it suffices to only check those (s_l,a_l) in the sample trajectory that were visited for the first time, i.e., $n_l(s_l,a_l) = 0$ and $n_{l+1}(s_l,a_l) = 1$. It is straightforward to check that, for any $1 \leq \tau \leq T$,

$$
\Big\{ \mathbb{1}_{k_i<T} \Big((P - P_{k_i}) V_{k_i} \Big) (s_l,a_l) \Big\}_{i=1}^{\tau}
$$

is a martingale difference sequence with respect to $\{\mathcal{F}_i\}_{i\geq 0}$, where $\mathcal{F}_i$ is the σ-field generated by $\{(s_i,a_i)\}_{i=0}^{k_i}$. Then we can invoke the Azuma-Hoeffding inequality to show that: for any such (s_l,a_l) and any $\tau \in [T]$, with probability exceeding $1-\delta/T^2$,

$$
\left| \sum_{i=1}^{\tau} \mathbb{1}_{k_i<T}\, \eta_i^{\tau} \Big((P - P_{k_i}) V_{k_i} \Big) (s_l,a_l) \right| \lesssim \frac{1}{1-\gamma} \sqrt{\sum_{i=1}^{\tau} (\eta_i^{\tau})^2 \log\frac{T}{\delta}} \lesssim \sqrt{\frac{H\iota}{\tau(1-\gamma)^2}}.
$$

Taking the union bound over $\tau \in [T]$ allows us to replace τ with $n_t(s_l,a_l)$ in the above inequality, namely, this shows that for any such (s_l,a_l), with probability exceeding $1-\delta/T$ we have

$$
\left| \sum_{i=1}^{n_t(s_l,a_l)} \eta_i^{n_t(s_l,a_l)} \Big((P - P_{k_i}) V_{k_i} \Big) (s_l,a_l) \right| \lesssim \sqrt{\frac{H\iota}{n_t(s_l,a_l)(1-\gamma)^2}}
$$

for all $t \in [T]$. Taking the union bound over all such (s_l,a_l) (which are concerned with at most T pairs), we see that with probability exceeding $1-\delta$,

$$
\left| \sum_{i=1}^{n_t(s_l,a_l)} \eta_i^{n_t(s_l,a_l)} \Big((P - P_{k_i}) V_{k_i} \Big) (s_l,a_l) \right| \lesssim \sqrt{\frac{H\iota}{n_t(s_l,a_l)(1-\gamma)^2}}
$$

is valid for any $0 \leq j < T$ and any $t \in [T]$. This establishes the inequality (C.6), thus concluding the proof.

C.2 Proof of Lemma 3

For each $\big(s,\pi^\star(s)\big)\in\mathcal{I}^c$, we first have

$$\mathbb{P}\Big\{\,(s_t,a_t)=\big(s,\pi^\star(s)\big)\mid(s_0,a_0)\sim\mu_{\mathsf{b}}\Big\}=\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)<\frac{\delta}{ST},$$

given that μ_{b} is taken to be the stationary distribution of the sample trajectory. By virtue of the union bound, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{I}^c\cap\{(s_t,a_t)\}_{t=t_{\mathsf{mix}}(\delta)}^{T}=\varnothing\mid(s_0,a_0)\sim\mu_{\mathsf{b}}\right)\\
&\geq 1-\sum_{t=t_{\mathsf{mix}}}^{T}\sum_{s:(s,\pi^\star(s))\in\mathcal{I}^c}\mathbb{P}\Big\{\,(s_t,a_t)=(s,\pi^\star(s))\mid(s_0,a_0)\sim\mu_{\mathsf{b}}\Big\}\\
&>1-\delta.
\end{aligned}$$

In addition, for an arbitrary pair $(s,a)\in\mathcal{S}\times\mathcal{A}$, the definition of the mixing time gives

$$\left|\mathbb{P}\left(\big\{\,(s_t,a_t)\,\big\}_{t=t_{\mathsf{mix}}(\delta)}^{T}\subseteq\mathcal{I}\mid(s_0,a_0)\sim\mu_{\mathsf{b}}\right)-\mathbb{P}\left(\{(s_t,a_t)\}_{t=t_{\mathsf{mix}}}^{T}\subseteq\mathcal{I}\mid(s_0,a_0)=(s,a)\right)\right|\leq\delta.$$

Combine the above results to yield

$$\mathbb{P}\left(\big\{\,(s_t,a_t)\,\big\}_{t=t_{\mathsf{mix}}(\delta)}^{T}\subseteq\mathcal{I}\mid(s_0,a_0)=(s,a)\right)\geq 1-2\delta$$

for an arbitrary pair $(s,a)\in\mathcal{S}\times\mathcal{A}$.

C.3 Proof of Lemma 4

For any given integer $K>0$, one can decompose

$$\begin{aligned}
\sum_{j=0}^{\infty}\left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j\big\langle\rho(P_{\pi^\star})^j,V\big\rangle&=\sum_{j=0}^{K-1}\left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j\big\langle\rho(P_{\pi^\star})^j,V\big\rangle+\sum_{j=K}^{\infty}\left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j\big\langle\rho(P_{\pi^\star})^j,V\big\rangle\\
&\leq\left(1+\frac{1}{H}\right)^{3K}\sum_{j=0}^{K-1}\gamma^j\big\langle\rho(P_{\pi^\star})^j,V\big\rangle+\sum_{j=K}^{\infty}\left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j\|V\|_\infty\\
&\leq\underbrace{\left(1+\frac{1}{H}\right)^{3K}\frac{1}{1-\gamma}\big\langle d_\rho^\star,V\big\rangle}_{=:\,\alpha_1}+\underbrace{\gamma^K\left(1+\frac{1}{H}\right)^{3K}\frac{1}{1-\gamma\left(1+\frac{1}{H}\right)^3}\|V\|_\infty}_{=:\,\alpha_2}.
\end{aligned}$$

Here, the last inequality holds since $d_\rho^\star=(1-\gamma)\sum_{j=0}^{\infty}\gamma^j\rho\big(P_{\pi^\star}\big)^j$.

By taking

$$K=H=\left\lceil\frac{4}{1-\gamma}\log\frac{ST}{\delta}\right\rceil,$$

we can derive

$$\left(1+\frac{1}{H}\right)^{3K}=\left(1+\frac{1}{H}\right)^{3H}\overset{\text{(i)}}{\leq}e^3=O(1)$$

and

$$\gamma^K=e^{K\log[1-(1-\gamma)]}\overset{\text{(ii)}}{\leq}e^{-K(1-\gamma)}=\frac{\delta}{ST^4}.$$

Here, (i) holds since $(1+1/x)^x\leq e$ for all $x>0$; (ii) is valid since $\log(1-x)\leq -x$ for all $x\in(0,1)$. It is also worth noting that

$$\frac{1}{1-\gamma\left(1+\frac{1}{H}\right)^3}\leq\frac{1}{1-\gamma\left(1+\frac{1-\gamma}{4}\right)^3}\overset{\text{(iii)}}{\leq}\frac{1}{1-\gamma\left[1+\frac{61}{64}(1-\gamma)\right]}=\frac{1}{(1-\gamma)\left(1-\frac{61}{64}\gamma\right)}\lesssim\frac{1}{1-\gamma},\tag{C.7}$$

where (iii) holds since $(1+x)^3 \le 1 + 61x/16$ for all $0 < x \le 1/4$. We then immediately arrive at

$$\alpha_1 \lesssim \frac{1}{1-\gamma} \left\langle d_\rho^\star, V \right\rangle$$

and

$$\alpha_2 \lesssim \frac{\delta}{ST^4 (1-\gamma)} \|V\|_\infty .$$

Taking the upper bounds on α_1 and α_2 collectively leads to the advertised inequality

$$\sum_{j=0}^{\infty} \left[\gamma \left(1 + \frac{1}{H}\right)^3 \right]^j \left\langle \rho(P_{\pi^\star})^j, V \right\rangle \lesssim \frac{1}{1-\gamma} \left\langle d_\rho^\star, V \right\rangle + \frac{\delta}{ST^4 (1-\gamma)} \|V\|_\infty .$$

C.4 Proof of Lemma 5

For notational simplicity, we denote

$$X_t \coloneqq (s_t, a_t), \qquad 1 \le t \le T;$$

clearly, $\{X_t\}_{t\ge 0}$ forms a Markov chain on $\mathcal{X} \triangleq \mathcal{S} \times \mathcal{A}$, with stationary distribution μ_{b}. In what follows, we demonstrate how to construct the sequence

$$Y_{K-1}^i = (s_{K-1}^i, a_{K-2}^i), \quad Y_{K-2}^i = (s_{K-2}^i, a_{K-2}^i), \quad \cdots, \quad Y_1^i = (s_1^i, a_1^i)$$

so as to satisfy the desired properties.

Let us start by constructing Y_{K-1}^i. Recall from the definition of the mixing time that: for any fixed state-action pairs $x_0, x_1, \cdots, x_{(K-2)\tau+i} \in \mathcal{X}$, one has

$$\mathsf{TV}\left(\mathcal{L}\left(X_{(K-1)\tau+i} \mid X_0 = x_0, \dots, X_{(K-2)\tau+i} = x_{(K-2)\tau+i}\right), \mu_{\mathsf{b}}\right) \le \frac{\delta}{T^2}.$$

where $\mathcal{L}(\cdot)$ denotes the law of the random variable. In view of the definition of the total-variation distance, we know that there exists a random variable $Y_{K-1}^{x_0,\dots,x_{(K-2)\tau+i}}$ such that conditional on the event $X_0 = x_0, \dots, X_{(K-2)\tau+i} = x_{(K-2)\tau+i}$,

(i) the law of $Y_{K-1}^{x_0,\dots,x_{(K-2)\tau+i}}$ obeys

$$\mathcal{L}\left(Y_{K-1}^{x_0,\dots,x_{(K-2)\tau+i}} \mid X_0 = x_0, \dots, X_{(K-2)\tau+i} = x_{(K-2)\tau+i}\right) = \mu_{\mathsf{b}}$$

(ii) $Y_{K-1}^{x_0,\dots,x_{(K-2)\tau+i}}$ is almost identical to $X_{(K-1)\tau+i}$ in the sense that

$$\mathbb{P}\left\{X_{(K-1)\tau+i} \ne Y_{K-1}^{x_0,\,\dots,\,x_{(K-2)\tau+i}} \mid X_0 = x_0, \dots, X_{(K-2)\tau+i} = x_{(K-2)\tau+i}\right\} \le \frac{\delta}{T^2}.$$

As a consequence, we can construct Y_{K-1}^i as follows

$$Y_{K-1}^i \coloneqq \sum_{x_0,\dots,x_{(K-2)\tau+i} \in \mathcal{X}} Y_{K-1}^{x_0,\dots,x_{(K-2)\tau+i}} \mathbb{1}\{X_0 = x_0, \dots, X_{(K-2)\tau+i} = x_{(K-2)t_{\mathsf{mix}}+i}\};$$

as can be easily verified, for any $x_0, x_1, \cdots, x_{(K-2)t_{\mathsf{mix}}+i} \in \mathcal{X}$ one has

$$\begin{aligned}
&\mathcal{L}\left(Y_{K-1}^i \mid X_0 = x_0, \dots, X_{(K-2)\tau+i} = x_{(K-2)\tau+i}\right) \\
&\qquad = \mathcal{L}\left(Y_{K-1}^{x_0,\dots,x_{(K-2)\tau+i}} \mid X_0 = x_0, \dots, X_{(K-2)\tau+i} = x_{(K-2)\tau+i}\right) = \mu_{\mathsf{b}}.
\end{aligned}$$

All this in turn implies that

$$Y_{K-1}^i \sim \mu_{\mathsf{b}} \qquad \text{and} \qquad Y_{K-1}^i \perp\!\!\!\perp \left\{X_0, X_1 \dots, X_{(K-2)\tau+i}\right\}.$$

In addition, it is also seen that

$$\begin{aligned}
\mathbb{P}\left(Y_{K-1}^{i} \neq X_{(K-1)\tau+i}\right) &= \sum_{x_0,\dots,x_{(K-2)\tau+i}\in\mathcal{X}} \mathbb{P}\left(X_0 = x_0,\dots,X_{(K-2)\tau+i} = x_{(K-2)\tau+i}\right) \\
&\qquad \cdot \mathbb{P}\Big\{X_{(K-1)\tau+i} \neq Y_{K-1}^{x_0,\dots,x_{(K-2)\tau+i}} \mid X_0 = x_0,\dots,X_{(K-2)\tau+i} = x_{(K-2)\tau+i}\Big\} \\
&\leq \frac{\delta}{T^2} \sum_{x_0,\dots,x_{(K-2)\tau+i}\in\mathcal{X}} \mathbb{P}\left(X_0 = x_0,\dots,X_{(K-2)\tau+i} = x_{(K-2)\tau+i}\right) \\
&= \frac{\delta}{T^2}.
\end{aligned}$$

Next, we turn to the construction of Y_{K-2}^{i}. Consider any fixed $x_0, x_1, \cdots, x_{(K-3)\tau+i}, y_{K-1}^{i} \in \mathcal{X}$. Given that $Y_{K-1}^{i} \perp\!\!\!\perp \{X_0, X_1 \dots, X_{(K-2)\tau+i}\}$, the conditional law of $X_{(K-2)\tau+i}$ obeys

$$\begin{aligned}
&\mathcal{L}\left(X_{(K-2)\tau+i} \mid X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y_{K-1}^{i} = y_{k-1}^{i}\right) \\
&\quad = \mathcal{L}\left(X_{(K-2)\tau+i} \mid X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}\right).
\end{aligned}$$

This in turn allows one to obtain

$$\begin{aligned}
&\mathsf{TV}\Big(\mathcal{L}\left(X_{(K-2)\tau+i} \mid X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y_{K-1}^{i} = y_{K-1}^{i}\right), \mu_{\mathsf{b}}\Big) \\
&\quad = \mathsf{TV}\left(\mathcal{L}\left(X_{(K-2)\tau+i} \mid X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}\right), \mu_{\mathsf{b}}\right) \\
&\quad \leq \frac{\delta}{T^2}.
\end{aligned}$$

According to the definition of the total-variation distance, there exists a random variable $Y_{K-2}^{x_0,x_1,\cdots,x_{(K-3)\tau+i},y_{K-1}^{i}}$ such that: conditional on the event $X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y_{K-1}^{i} = y_{k-1}^{i}$,

(i) the law of $Y_{K-2}^{x_0,x_1,\cdots,x_{(K-3)\tau+i},y_{K-1}^{i}}$ obeys

$$\mathcal{L}\left(Y_{K-2}^{x_0,x_1,\cdots,x_{(K-3)\tau+i},y_{K-1}^{i}} \mid X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y_{K-1}^{i} = y_{K-1}^{i}\right) = \mu_{\mathsf{b}};$$

(ii) $Y_{K-2}^{x_0,x_1,\cdots,x_{(K-3)\tau+i},y_{K-1}^{i}}$ is almost identical to $X_{(K-2)\tau+i}$ in the following sense

$$\mathbb{P}\left(X_{(K-2)\tau+i} \neq Y_{K-2}^{x_0,x_1,\cdots,x_{(K-3)\tau+i},y_{K-1}^{i}} \mid X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y_{K-1}^{i} = y_{K-1}^{i}\right) \leq \frac{\delta}{T^2}.$$

With the above set of random variables in mind, we can readily construct Y_{K-2}^{i} as follows:

$$Y_{K-2}^{i} \coloneqq \sum_{x_0,x_1,\cdots,x_{(K-3)\tau+i},y_{K-1}^{i}\in\mathcal{X}} Y_{K-2}^{x_0,x_1,\cdots,x_{(K-3)\tau+i},y_{K-1}^{i}} \mathbb{1}\left\{X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y_{K-1}^{i} = y_{k-1}^{i}\right\}.$$

As can be straightforwardly verified, for any $x_0, x_1, \cdots, x_{(K-3)\tau+i}, y_{K-1}^{i} \in \mathcal{X}$ we have

$$\begin{aligned}
&\mathcal{L}\left(Y_{K-2}^{i} \mid X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y_{K-1}^{i} = y_{k-1}^{i}\right) \\
&\quad = \mathcal{L}\left(Y_{K-2}^{x_0,x_1,\cdots,x_{(K-3)\tau+i},y_{K-1}^{i}} \mid X_0 = x_0,\dots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y_{K-1}^{i} = y_{k-1}^{i}\right) = \mu_{\mathsf{b}},
\end{aligned}$$

thus implying that $Y_{K-2}^{i} \sim \mu_{\mathsf{b}}$ and

$$Y_{K-2}^{i} \perp\!\!\!\perp \left\{X_0, X_1 \dots, X_{(K-3)\tau+i}, Y_{K-1}^{i}\right\}.$$

This reveals that $Y^i_{K-1}, Y^i_{K-2} \overset{\text{i.i.d.}}{\sim} \mu_{\mathsf{b}}$. In addition, we can also show that

$$
\begin{aligned}
&\mathbb{P}\left(Y^i_{K-2} \neq X_{(K-2)\tau+i}\right) \\
&\quad = \sum_{x_0,x_1,\cdots,x_{(K-3)\tau+i},y^i_{K-1}\in\mathcal{X}} \mathbb{P}\left(X_0 = x_0,\ldots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y^i_{K-1} = y^i_{K-1}\right) \\
&\qquad\quad \cdot \mathbb{P}\left(X_{(K-2)\tau+i} \neq Y^{x_0,x_1,\cdots,x_{(K-3)\tau+i},y^i_{K-1}}_{K-2} \mid X_0 = x_0,\ldots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y^i_{K-1} = y^i_{K-1}\right) \\
&\quad \leq \frac{\delta}{T^2} \sum_{x_0,x_1,\cdots,x_{(K-3)\tau+i},y^i_{K-1}\in\mathcal{X}} \mathbb{P}\left(X_0 = x_0,\ldots,X_{(K-3)\tau+i} = x_{(K-3)\tau+i}, Y^i_{K-1} = y^i_{K-1}\right) \\
&\quad = \frac{\delta}{T^2}.
\end{aligned}
$$

Moving forward, we can employ similar arguments to construct $Y^i_{K-3},\ldots,Y^i_1$ sequentially such that:

(i) $Y^i_1, Y^i_2, \ldots, Y^i_{K-1} \overset{\text{i.i.d.}}{\sim} \mu_{\mathsf{b}}$;

(ii) for all $1 \leq k \leq K-1$,

$$
Y^i_k \perp\!\!\!\perp \left\{X_0, X_1, \ldots, X_{(k-1)\tau+i}\right\} \qquad \text{and} \qquad \mathbb{P}\left(Y^i_k \neq X_{k\tau+i}\right) \leq \frac{\delta}{T^2}.
$$

As a result, we arrive at

$$
\mathbb{P}\left(Y^i_1 = X_{\tau+i}, \cdots, Y^i_{K-1} = X_{(K-1)\tau+i}\right) \geq 1 - \sum_{k=1}^{K-1} \mathbb{P}\left(Y^i_k = X_{k\tau+i}\right) \geq 1 - \frac{\delta}{T}.
$$

This concludes the proof.

D Analysis for variance-reduced Q-learning with LCB penalization (Theorem 2)

This section presents the proof of Theorem 2, which is concerned with the performance of variance-reduced Q-learning with LCB penalization. Recall that $\overline{V}_{k+1} = V_{T_k}$, that is, the value estimate in the last iterate of the k-th epoch is also used as the reference for the $(k+1)$-th epoch. For each $1 \leq k \leq K$, we define

$$
\Lambda_k \coloneqq \sum_{s\in\mathcal{S}} \rho\left(s\right)\left(V^\star - \overline{V}_k\right)(s) \tag{D.1}
$$

Clearly, the proof of Theorem 2 boils down to bounding Λ_K. As we shall see momentarily, obtaining a tight bound on Λ_k relies on bounding another closely related quantity Δ_{K-1}, define for each $1 \leq k \leq K$ as follows:

$$
\Delta_k \coloneqq \sum_{s\in\mathcal{S}} \widetilde{\rho}\left(s\right)\left(V^\star - \overline{V}_k\right)(s)\,. \tag{D.2}
$$

Here, we set

$$
\widetilde{\rho} \coloneqq \frac{d^\star_\rho - (1-\gamma)\,\rho}{\gamma}. \tag{D.3}
$$

The sequence $\{\Delta_k\}_{k=1}^K$ will be bounded by induction in the sequel. We shall present our proof by describing three key steps following some preliminary facts.

D.1 Preliminary facts about the k-th epoch

Let us first look at what happens in the k-th epoch. For notational simplicity, we will denote $\overline{V} := \overline{V}_{k-1}$. Similar to the proof of Theorem 1, for any iterate $t \leq T_k$, let $n = n_t(s,a)$ and assume that (s,a) has been visited during the iterations $k_1 < \cdots < k_n < t$. We also need to define the policy $\pi_t : \mathcal{S} \to \mathcal{A}$ as follows

$$\pi_t(s) := \begin{cases} \arg\max_{a\in\mathcal{A}} Q_t(s_{t-1}, a), & \text{if } s = s_{t-1} \text{ and } V_t(s) > V_{t-1}(s), \\ \pi_{t-1}(s), & \text{otherwise.} \end{cases}$$

If there are multiple $a \in \mathcal{A}$ that maximize $Q_t(s_{t-1}, a)$ simultaneously, then we can go with any of them. We make note of the following lemma.

Lemma 6. *With probability exceeding $1-\delta$, for any $s \in \mathcal{S}$ and $t \in [T]$ we have*

$$(Q^\star - Q_t)\big(s, \pi^\star(s)\big) \leq \gamma \sum_{i=1}^{n} \eta_i^n P_{s,a}(V^\star - V_{k_i}) + \beta_n\big(s, \pi^\star(s)\big),$$

where $n = n_t\big(s, \pi^\star(s)\big)$ and

$$\begin{aligned}\beta_n(s,a) := {} & 3C_{\mathsf{b}}\sqrt{\frac{H\iota}{n}\Big\{\sigma_n^{\mathsf{adv}}(s,a) - [\mu_n^{\mathsf{adv}}(s,a)]^2\Big\}} + 3C_{\mathsf{b}}\frac{H^{3/4}\iota^{3/4}}{n^{3/4}(1-\gamma)} + 3C_{\mathsf{b}}\frac{H\iota}{n(1-\gamma)} \\ & + 3C_{\mathsf{b}}\sqrt{\frac{\iota}{n^{\mathsf{ref}}(s,a)}\Big\{\sigma^{\mathsf{ref}}(s,a) - [\mu^{\mathsf{ref}}(s,a)]^2\Big\}} + 3C_{\mathsf{b}}\frac{\iota^{3/4}}{(1-\gamma)[n^{\mathsf{ref}}(s,a)]^{3/4}} \\ & + 3C_{\mathsf{b}}\frac{\iota}{(1-\gamma)n^{\mathsf{ref}}(s,a)}. \end{aligned} \tag{D.4}$$

In addition, it holds that

$$V_t(s) \leq V^{\pi_t}(s) \leq V^\star(s) \qquad \text{for all } s \in \mathcal{S} \text{ and } 1 \leq t \leq T_k.$$

Proof. See Appendix E.1. □

Moreover, both $\sigma_n^{\mathsf{adv}}(s,a)$ and $\sigma^{\mathsf{ref}}(s,a) - [\mu^{\mathsf{ref}}(s,a)]^2$ play an important role in determining the variance of the update, and we are in need of the following bounds on these two quantities.

Lemma 7. *With probability exceeding $1-\delta$, for all $s \in \mathcal{S}$ and $t \in [T_k]$ we have*

$$\sigma^{\mathsf{adv}}_{n_t(s,\pi^\star(s))}\big(s, \pi^\star(s)\big) \leq P_{s,\pi^\star(s)}\big(V^\star - \overline{V}\big)^2 + O\left(\frac{1}{(1-\gamma)^2}\sqrt{\frac{H\iota}{n_t\big(s,\pi^\star(s)\big)}}\right)$$

and

$$\sigma^{\mathsf{ref}}\big(s, \pi^\star(s)\big) - \big[\mu^{\mathsf{ref}}\big(s, \pi^\star(s)\big)\big]^2 = \mathsf{Var}_{s,\pi^\star(s)}(\overline{V}) + O\left(\frac{1}{(1-\gamma)^2}\sqrt{\frac{\iota}{n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)}}\right).$$

In addition, it holds that

$$\sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\rho^\star(s,a)\,\mathsf{Var}_{s,a}(V^\star - \overline{V}) \leq \frac{1}{1-\gamma}\Delta_{k-1}; \tag{D.5a}$$

$$\sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\rho^\star(s,a)\,\mathsf{Var}_{s,a}(\overline{V}) \leq \frac{8}{1-\gamma} + \frac{2}{1-\gamma}\Delta_{k-1}. \tag{D.5b}$$

Proof. See Appendix E.2. □

D.2 Step 1: connecting Λ_k with Δ_{k-1}

In this step, we aim to establish a connection between Λ_k (cf. (D.1)) and Δ_{k-1} (cf. (D.2)). In view of the monotonicity of V_t in t (by construction) and Lemma 6, we can derive

$$\Lambda_k = \left\langle \rho, V^\star - V_{T_k} \right\rangle \le \frac{1}{T_k} \sum_{t=1}^{T_k} \left\langle \rho, V^\star - V_t \right\rangle . \tag{D.6}$$

Before continuing, we find it convenient to introduce a set of quantities (similar to our proof for Theorem 1):

$$\begin{aligned}
\alpha_j &\coloneqq \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \sum_{t=1}^{T_k} \left\langle \rho (P_{\pi^\star})^j, V^\star - V_t \right\rangle, \\
\theta_j &\coloneqq \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \sum_{t=1}^{T_k} \sum_{s\in\mathcal{S}} \left[\rho (P_{\pi^\star})^j\right]\left(s,\pi^\star(s)\right) \min\left\{\beta_{n_t(s,\pi^\star(s))}\left(s,\pi^\star(s)\right), \frac{1}{1-\gamma}\right\}, \\
\xi_j &\coloneqq \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \left\langle \rho (P_{\pi^\star})^j, V^\star - V_t \right\rangle + \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^{j+1} \left\langle \rho (P_{\pi^\star})^{j+1}, V^\star - V_0 \right\rangle, \\
\psi_j &\coloneqq \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \Bigg[\sum_{s\in\mathcal{S},a\in\mathcal{A}} \left[\rho^{\pi^\star} (P^{\pi^\star})^j\right](s,a) \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} P_{s,a}\left(V^\star - V_{k_i(s,a)}\right) \\
&\qquad\qquad - \left(1+\frac{1}{H}\right) \frac{\left[\rho^{\pi^\star} (P^{\pi^\star})^j\right](s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)} \sum_{i=1}^{n_t(s_t,a_t)} \eta_i^{n_t(s_t,a_t)} P_{s_t,a_t}\left(V^\star - V_{k_i(s_t,a_t)}\right) \Bigg], \\
\phi_j &\coloneqq \gamma^{j+1}\left(1+\frac{1}{H}\right)^{3j+2} \sum_{t=0}^{T_k} \mathbb{1}_{(s_t,a_t)\in\mathcal{I}} \Bigg[\frac{\left[\rho^{\pi^\star} (P^{\pi^\star})^j\right](s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)} P_{s_t,a_t}\left(V^\star - V_t\right) \\
&\qquad\qquad - \left(1+\frac{1}{H}\right) \sum_{s\in\mathcal{S},a\in\mathcal{A}} \left[\rho^{\pi^\star} (P^{\pi^\star})^j\right](s,a) P_{s,a}\left(V^\star - V_t\right) \Bigg].
\end{aligned}$$

Repeat the same analysis as in Step 1 of the proof of Theorem 1 (which we omit here for brevity) to yield

$$\alpha_0 \le \underbrace{\limsup_{j\to\infty} \alpha_j}_{=:\,\alpha} + \underbrace{\sum_{j=0}^{\infty} \xi_j}_{=:\,\xi} + \underbrace{\sum_{j=0}^{\infty} \theta_j}_{=:\,\theta} + \underbrace{\sum_{j=0}^{\infty} \psi_j}_{=:\,\psi} + \underbrace{\sum_{j=0}^{\infty} \phi_j}_{=:\,\phi},$$

as well as the properties that $\alpha = 0$,

$$\begin{aligned}
\xi &\lesssim \frac{2t_{\mathsf{mix}}}{1-\gamma}\log\frac{1}{\delta} + \frac{t_{\mathsf{mix}}}{T^4(1-\gamma)^2}\log\frac{1}{\delta}, \\
\psi &\lesssim \frac{C^\star t_{\mathsf{mix}}\iota}{(1-\gamma)^3}\log^2\left(\frac{T}{\delta}\right) + \frac{C^\star S t_{\mathsf{mix}}}{(1-\gamma)^2}\log\left(\frac{T}{\delta}\right), \\
\phi &\lesssim \frac{C^\star t_{\mathsf{mix}}\iota}{(1-\gamma)^3}\log^2\left(\frac{T}{\delta}\right) + \frac{C^\star S t_{\mathsf{mix}}}{(1-\gamma)^2}\log\left(\frac{T}{\delta}\right).
\end{aligned}$$

It then comes down to bounding θ, which is different from what has been done in the proof of Theorem 1. Towards this, we first invoke Lemma 4 to reach

$$\begin{aligned}
\theta &= \sum_{t=1}^{T_k}\sum_{j=0}^{\infty} \left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \sum_{s\in\mathcal{S}} \left[\rho (P_{\pi^\star})^j\right]\left(s,\pi^\star(s)\right) \min\left\{\beta_{n_t(s,\pi^\star(s))}\left(s,\pi^\star(s)\right), \frac{1}{1-\gamma}\right\} \\
&\lesssim \frac{1}{1-\gamma} \sum_{t=1}^{T_k}\sum_{s\in\mathcal{S}} d_\rho^\star(s) \min\left\{\beta_{n_t(s,\pi^\star(s))}\left(s,\pi^\star(s)\right), \frac{1}{1-\gamma}\right\} + \frac{1}{ST^4(1-\gamma)}\frac{T}{1-\gamma}.
\end{aligned} \tag{D.7}$$

To proceed, let us use the definition of $\beta_n(s,a)$ (cf. (D.4)) to decompose

$$
\begin{aligned}
&\sum_{s\in\mathcal{S}}\sum_{t=1}^{T_k} d_\rho^\star(s)\min\left\{\beta_{n_t(s,\pi^\star(s))}\big(s,\pi^\star(s)\big),\frac{1}{1-\gamma}\right\}\\
&\lesssim \underbrace{\sum_{s\in\mathcal{S}}\sum_{t=1}^{t_{\mathsf{burn\text{-}in}}(s)} d_\rho^\star(s)\frac{1}{1-\gamma}}_{=:\,\omega_0}+\underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)}^{T_k} d_\rho^\star(s)\sqrt{\frac{H\iota}{n_t\big(s,\pi^\star(s)\big)}\left\{\sigma_n^{\mathsf{adv}}\big(s,\pi^\star(s)\big)-\big[\mu_n^{\mathsf{adv}}\big(s,\pi^\star(s)\big)\big]^2\right\}}}_{=:\,\omega_1}\\
&\quad+\underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)}^{T_k} d_\rho^\star(s)\sqrt{\frac{\iota}{n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)}\left\{\sigma^{\mathsf{ref}}\big(s,\pi^\star(s)\big)-\big[\mu^{\mathsf{ref}}\big(s,\pi^\star(s)\big)\big]^2\right\}}}_{=:\,\omega_2}\\
&\quad+\underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)}^{T_k} d_\rho^\star(s)\frac{H^{3/4}\iota^{3/4}}{n_t^{3/4}\big(s,\pi^\star(s)\big)(1-\gamma)}}_{=:\,\omega_3}+\underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)}^{T_k} d_\rho^\star(s)\frac{H\iota}{n_t\big(s,\pi^\star(s)\big)(1-\gamma)}}_{=:\,\omega_4}\\
&\quad+\underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)}^{T_k} d_\rho^\star(s)\frac{\iota^{3/4}}{(1-\gamma)\big[n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)\big]^{3/4}}}_{=:\,\omega_5}+\underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)}^{T_k} d_\rho^\star(s)\frac{\iota}{(1-\gamma)\,n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)}}_{=:\,\omega_6},
\end{aligned}
\tag{D.8}
$$

where we define, for each $s\in\mathcal{S}$, that

$$
t_{\mathsf{burn\text{-}in}}(s) := C_{\mathsf{burn\text{-}in}}\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)}\log\left(\frac{ST}{\delta}\right)
$$

for some sufficiently large constant $C_{\mathsf{burn\text{-}in}}>0$.

Before continuing, we first collect a few immediate and useful results of Li et al. (2021b, Lemma 8): with probability exceeding $1-\delta$, we have

$$
n_t\big(s,\pi^\star(s)\big)\gtrsim t\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big),\qquad \forall s\in\mathcal{S}\ \text{ and }\ t_{\mathsf{burn\text{-}in}}(s)\le t\le T_k \tag{D.9}
$$

and when $T_k^{\mathsf{ref}}\asymp T_k\ge t_{\mathsf{burn\text{-}in}}(s)$, one has

$$
n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)\gtrsim T_k^{\mathsf{ref}}\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big),\qquad \forall s\in\mathcal{S}, \tag{D.10}
$$

provided that $C_{\mathsf{burn\text{-}in}}$ is large enough. We then bound the terms $\omega_0,\ldots,\omega_6$ separately.

- The first bound ω_0 can be easily bounded by

$$
\begin{aligned}
\omega_0 &\le \sum_{s\in\mathcal{S}} t_{\mathsf{burn\text{-}in}}(s)d_\rho^\star(s)\frac{1}{1-\gamma}\lesssim\sum_{s\in\mathcal{S}}\frac{t_{\mathsf{mix}}\iota}{\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)}d_\rho^\star\big(s,\pi^\star(s)\big)\frac{1}{1-\gamma}\\
&\le\sum_{s\in\mathcal{S}}\frac{C^\star t_{\mathsf{mix}}\iota}{1-\gamma}\asymp\frac{C^\star S t_{\mathsf{mix}}\iota}{1-\gamma},
\end{aligned}
$$

  where the last line follows from Assumption 2.

- To control ω_1, we observe that

$$
\begin{aligned}
\omega_1 &\lesssim \sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)}^{T_k} d_\rho^\star(s)\sqrt{\frac{H\iota}{n_t\big(s,\pi^\star(s)\big)}\sigma_n^{\mathsf{adv}}\big(s,\pi^\star(s)\big)}\\
&\lesssim \underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)}^{T_k} d_\rho^\star(s)\sqrt{\frac{H\iota}{n_t\big(s,\pi^\star(s)\big)}P_{s,\pi^\star(s)}\big(V^\star-\overline{V}\big)^2}}_{=:\,\omega_{1,1}}+\underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\mathsf{burn\text{-}in}}(s)}^{T_k} d_\rho^\star(s)\frac{H^{3/4}\iota^{3/4}}{(1-\gamma)\,n_t^{3/4}\big(s,\pi^\star(s)\big)}}_{=:\,\omega_{1,2}},
\end{aligned}
$$

where the last inequality follows from Lemma 7. The first term $\omega_{1,1}$ admits the following bound

$$\begin{aligned}
\omega_{1,1} &\overset{\text{(i)}}{\asymp} \sum_{s\in\mathcal{S}}\sum_{t=t_{\text{burn-in}}(s)}^{T_k} d_\rho^\star\big(s,\pi^\star(s)\big)\sqrt{\frac{H\iota}{t\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)}P_{s,\pi^\star(s)}\left(V^\star-\overline{V}\right)^2} \\
&\overset{\text{(ii)}}{\lesssim} \sum_{s\in\mathcal{S}}\sum_{t=1}^{T_k}\sqrt{\frac{C^\star H\iota}{t}d_\rho^\star\big(s,\pi^\star(s)\big)P_{s,\pi^\star(s)}\left(V^\star-\overline{V}\right)^2} \\
&\overset{\text{(iii)}}{\lesssim} \sqrt{C^\star H\iota T_k}\sum_{s\in\mathcal{S}}\sqrt{d_\rho^\star\big(s,\pi^\star(s)\big)P_{s,\pi^\star(s)}\left(V^\star-\overline{V}\right)^2} \\
&\overset{\text{(iv)}}{\lesssim} \sqrt{C^\star SH\iota T_k}\sqrt{\sum_{s\in\mathcal{S}}d_\rho^\star\big(s,\pi^\star(s)\big)P_{s,\pi^\star(s)}\left(V^\star-\overline{V}\right)^2} \\
&\asymp \sqrt{C^\star SH\iota T_k}\sqrt{\sum_{s\in\mathcal{S},a\in\mathcal{A}}d_\rho^\star\left(s,a\right)P_{s,a}\left(V^\star-\overline{V}\right)^2} \\
&\overset{\text{(v)}}{\lesssim} \sqrt{\frac{C^\star S\iota^2 T_k}{(1-\gamma)^2}}\sqrt{\Delta_{k-1}}.
\end{aligned}$$

Here, (i) follows from (D.9); (ii) utilizes Assumption 2; (iii) arises from (B.8); (iv) applies the Cauchy-Schwarz inequality; and (v) comes from Lemma 7 and the definition of H (i.e., $H\asymp\frac{\iota}{1-\gamma}$). The other term $\omega_{1,2}$ is identical to ω_3, which we shall bound momentarily.

- When it comes to ω_2, we apply Lemma 7 to reach

$$\omega_2 \lesssim \underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\text{burn-in}}(s)}^{T_k} d_\rho^\star\left(s\right)\sqrt{\frac{\iota}{n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)}\mathsf{Var}_{s,\pi^\star(s)}(\overline{V})}}_{=:\,\omega_{2,1}} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{t=t_{\text{burn-in}}(s)}^{T_k} d_\rho^\star\left(s\right)\frac{\iota^{3/4}}{(1-\gamma)\left[n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)\right]^{3/4}}}_{=:\,\omega_{2,2}}.$$

Regarding $\omega_{2,1}$, we make the observation that

$$\begin{aligned}
\omega_{2,1} &\overset{\text{(i)}}{\asymp} \sum_{s\in\mathcal{S}}\sum_{t=t_{\text{burn-in}}(s)}^{T_k} d_\rho^\star\big(s,\pi^\star(s)\big)\sqrt{\frac{\iota}{T_k\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)}\mathsf{Var}_{s,\pi^\star(s)}(\overline{V})} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{C^\star\iota T_k}\sum_{s\in\mathcal{S}}\sqrt{d_\rho^\star\big(s,\pi^\star(s)\big)\mathsf{Var}_{s,\pi^\star(s)}(\overline{V})} \\
&\asymp \sqrt{C^\star\iota T_k}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\sqrt{d_\rho^\star\left(s,a\right)\mathsf{Var}_{s,a}(\overline{V})} \\
&\overset{\text{(iii)}}{\lesssim} \sqrt{C^\star S\iota T_k}\sqrt{\sum_{s\in\mathcal{S},a\in\mathcal{A}}d_\rho^\star\left(s,a\right)\mathsf{Var}_{s,a}(\overline{V})} \\
&\overset{\text{(iv)}}{\lesssim} \sqrt{C^\star S\iota T_k}\sqrt{\frac{1}{1-\gamma}+\frac{\Delta_{k-1}}{1-\gamma}} \\
&\asymp \sqrt{\frac{C^\star S\iota T_k}{1-\gamma}}+\sqrt{\frac{C^\star S\iota T_k}{1-\gamma}}\sqrt{\Delta_{k-1}}.
\end{aligned}$$

Here, (i) relies on (D.10); (ii) invokes Assumption 2; (iii) utilizes the Cauchy-Schwarz inequality; and (iv) follows from Lemma 7. The other term $\omega_{2,2}$ is the same as ω_5, which will be bounded momentarily.

- Regarding ω_3, we have

$$\omega_3 \overset{\text{(i)}}{\asymp} \sum_{s\in\mathcal{S}}\sum_{t=t_{\text{burn-in}}(s)}^{T_k} d_\rho^\star\big(s,\pi^\star(s)\big)\frac{H^{3/4}\iota^{3/4}}{(1-\gamma)\,t^{3/4}\mu_{\mathsf{b}}^{3/4}\big(s,\pi^\star(s)\big)}$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{\lesssim} \frac{(C^\star)^{3/4} H^{3/4} \iota^{3/4}}{1-\gamma} \sum_{s\in\mathcal{S}} \sum_{t=1}^{T_k} \frac{\left[d^\star\big(s,\pi^\star(s)\big)\right]^{1/4}}{t^{3/4}} \\
&\overset{\text{(iii)}}{\lesssim} T_k^{1/4} \frac{(C^\star)^{3/4} H^{3/4} \iota^{3/4}}{1-\gamma} \sum_{s\in\mathcal{S}} \left[d^\star\big(s,\pi^\star(s)\big)\right]^{1/4} \\
&\overset{\text{(iv)}}{\lesssim} T_k^{1/4} \frac{(C^\star)^{3/4} H^{3/4} \iota^{3/4}}{1-\gamma} \left(\sum_{s\in\mathcal{S}} 1\right)^{3/4} \left(\sum_{s\in\mathcal{S}} d^\star\big(s,\pi^\star(s)\big)\right)^{1/4} \\
&\asymp T_k^{1/4} \frac{(C^\star)^{3/4} S^{3/4} \iota^{3/2}}{(1-\gamma)^{7/4}}.
\end{aligned}
$$

Here, (i) follows from (D.9); (ii) utilizes Assumption 2; (iii) follows from the fact that for any $T \geq 1$,

$$
\sum_{t=1}^{T} \frac{1}{t^{3/4}} \leq 1 + \int_1^T x^{-3/4} \mathrm{d}x = 1 + 4\left(T^{1/4} - 1\right) \leq 4T^{1/4}; \tag{D.11}
$$

(iv) follows from Hölder's inequality; and the last line holds since $\sum_s d_\rho^\star\big(s,\pi^\star(s)\big) = 1$.

- Regarding ω_4, we have

$$
\begin{aligned}
\omega_4 &\overset{\text{(i)}}{\asymp} \sum_{s\in\mathcal{S}} \sum_{t=t_{\textsf{burn-in}}(s)}^{T_k} d_\rho^\star\big(s,\pi^\star(s)\big) \frac{H\iota}{(1-\gamma)\, t\mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)} \\
&\overset{\text{(ii)}}{\lesssim} \frac{C^\star H \iota}{1-\gamma} \sum_{s\in\mathcal{S}} \sum_{t=1}^{T_k} \frac{1}{t} \overset{\text{(iii)}}{\lesssim} \frac{C^\star S \iota^2 \log T_k}{(1-\gamma)^2}.
\end{aligned}
$$

Here, (i) utilizes (D.9); (ii) relies on Assumption 2; and (iii) follows from the fact that for any $T \geq 1$,

$$
\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \int_1^T x^{-1} \mathrm{d}x = 1 + (\log T - 1) \leq \log T; \tag{D.12}
$$

- Moving on to ω_5, we develop the following upper bound:

$$
\begin{aligned}
\omega_5 &\overset{\text{(i)}}{\asymp} \sum_{s\in\mathcal{S}} \sum_{t=t_{\textsf{burn-in}}(s)}^{T_k} d_\rho^\star\big(s,\pi^\star(s)\big) \frac{\iota^{3/4}}{(1-\gamma)\,(T_k^{\textsf{ref}})^{3/4} \mu_{\mathsf{b}}^{3/4}\big(s,\pi^\star(s)\big)} \\
&\overset{\text{(ii)}}{\lesssim} \frac{(C^\star)^{3/4} \iota^{3/4}}{1-\gamma} \sum_{s\in\mathcal{S}} \sum_{t=1}^{T_k} \frac{\left[d_\rho^\star\big(s,\pi^\star(s)\big)\right]^{1/4}}{(T_k^{\textsf{ref}})^{3/4}} \\
&\overset{\text{(iii)}}{\asymp} T_k^{1/4} \frac{(C^\star)^{3/4} \iota^{3/4}}{1-\gamma} \sum_{s\in\mathcal{S}} \left[d_\rho^\star\big(s,\pi^\star(s)\big)\right]^{1/4} \\
&\overset{\text{(iv)}}{\lesssim} T_k^{1/4} \frac{(C^\star)^{3/4} S^{3/4} \iota^{3/4}}{1-\gamma}.
\end{aligned}
$$

Here, (i) follows from (D.10); (ii) results from Assumption 2; (iii) holds since $T_k^{\textsf{ref}} \asymp T_k$; and (iv) invokes Hölder's inequality and $\sum_s d_\rho^\star\big(s,\pi^\star(s)\big) = 1$ once again.

- We are left with bounding the last term ω_6, towards which we observe

$$
\omega_6 \overset{\text{(i)}}{\asymp} \sum_{s\in\mathcal{S}} \sum_{t=t_{\textsf{burn-in}}(s)}^{T_k} d_\rho^\star\big(s,\pi^\star(s)\big) \frac{\iota}{(1-\gamma)\, T_k^{\textsf{ref}} \mu_{\mathsf{b}}\big(s,\pi^\star(s)\big)}
$$

$$\overset{\text{(ii)}}{\lesssim} \frac{C^\star \iota}{1-\gamma}\sum_{s\in\mathcal{S}}\frac{T_k}{T_k^{\mathsf{ref}}}$$
$$\overset{\text{(iii)}}{\asymp} \frac{C^\star S\iota}{1-\gamma}.$$

Here, (i) follows from (D.10); (ii) utilizes Assumption 2; and (iii) holds since $T_k^{\mathsf{ref}} \asymp T_k$.

Taking the preceding bounds on ω_0, ω_1, ω_2, ω_3, ω_4, ω_5 and ω_6 together with (D.7) and (D.8) yields

$$\begin{aligned}
\theta &\lesssim \frac{1}{1-\gamma}\sum_{t=1}^{T_k}\sum_{s\in\mathcal{S}} d_\rho^\star(s)\min\left\{\beta_{n_t(s,\pi^\star(s))}, \frac{1}{1-\gamma}\right\} + \frac{1}{ST^4(1-\gamma)}\frac{T}{1-\gamma} \\
&\lesssim \frac{1}{1-\gamma}\left(\omega_0+\omega_1+\omega_2+\omega_3+\omega_4+\omega_5+\omega_6\right) + \frac{1}{ST^4(1-\gamma)}\frac{T}{1-\gamma} \\
&\lesssim \frac{C^\star S t_{\mathsf{mix}}\iota}{(1-\gamma)^2} + \sqrt{\frac{C^\star S\iota^2 T_k}{(1-\gamma)^4}}\sqrt{\Delta_{k-1}} + \sqrt{\frac{C^\star S\iota T_k}{(1-\gamma)^3}} + T_k^{1/4}\frac{(C^\star)^{3/4}S^{3/4}\iota^{3/2}}{(1-\gamma)^{11/4}} + \frac{C^\star S\iota^2\log T_k}{(1-\gamma)^2} \\
&\lesssim \frac{C^\star S t_{\mathsf{mix}}\iota}{(1-\gamma)^2} + \sqrt{\frac{C^\star S\iota^2 T_k}{(1-\gamma)^4}}\sqrt{\Delta_{k-1}} + \sqrt{\frac{C^\star S\iota T_k}{(1-\gamma)^3}} + \frac{C^\star S\iota^3}{(1-\gamma)^4},
\end{aligned}$$

where the last line has invoked the AM-GM inequality:

$$2T_k^{1/4}\frac{(C^\star)^{3/4}S^{3/4}\iota^{3/2}}{(1-\gamma)^{11/4}} = 2\frac{T_k^{1/4}(C^\star)^{1/4}S^{1/4}}{(1-\gamma)^{3/4}}\cdot\frac{(C^\star)^{1/2}S^{1/2}\iota^{3/2}}{(1-\gamma)^2} \le \frac{T_k^{1/2}(C^\star)^{1/2}S^{1/2}}{(1-\gamma)^{3/2}} + \frac{C^\star S\iota^3}{(1-\gamma)^4}.$$

Putting all of the above results together, we can conclude that

$$\begin{aligned}
\Lambda_k &\le \frac{1}{T_k}\alpha_0 \le \frac{1}{T_k}\left(\alpha+\xi+\theta+\psi+\phi\right) \\
&\lesssim \sqrt{\frac{C^\star S\iota^2}{T_k(1-\gamma)^4}}\sqrt{\Delta_{k-1}} + \sqrt{\frac{C^\star S\iota}{T_k(1-\gamma)^3}} + \frac{C^\star S\iota^3}{T_k(1-\gamma)^4} + \frac{C^\star S t_{\mathsf{mix}}\iota}{T_k(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}}\iota^2}{T_k(1-\gamma)^3}. \qquad \text{(D.13)}
\end{aligned}$$

D.3 Step 2: bounding Δ_k by induction

Thus far, we have established an intimate connection between Λ_k and Δ_k (see (D.13)). In order to bound Δ_{k-1}, we find it helpful to look at an auxiliary test distribution

$$\widetilde{\rho} = \frac{d_\rho^\star - (1-\gamma)\rho}{\gamma}$$

instead of ρ. The following property about $\widetilde{\rho}$ plays an important role in the subsequent analysis.

Lemma 8. *Suppose that $1/2 \le \gamma < 1$. It holds that*

$$\sum_{j=0}^{\infty}\left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j\left\langle\widetilde{\rho}(P_{\pi^\star})^j, V\right\rangle \lesssim \frac{1}{1-\gamma}\left\langle d_\rho^\star, V\right\rangle\log\frac{ST}{\delta} + \frac{\delta}{ST^4(1-\gamma)}\|V\|_\infty. \qquad \text{(D.14)}$$

Proof. See Appendix E.3. □

Armed with Lemma 8, we can repeat the same analysis in Step 1 to bound each Δ_k. The difference between (B.6) and (D.14) requires us to replace $d_\rho^\star$ in Step 1 with $d_\rho^\star\log(ST/\delta)$, which leads to

$$\Delta_k \lesssim \sqrt{\frac{C^\star S\iota^4}{T_k(1-\gamma)^4}}\sqrt{\Delta_{k-1}} + \sqrt{\frac{C^\star S\iota^3}{T_k(1-\gamma)^3}} + \frac{C^\star S\iota^4}{T_k(1-\gamma)^4} + \frac{C^\star S t_{\mathsf{mix}}\iota^2}{T_k(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}}\iota^3}{T_k(1-\gamma)^3}$$

for each $1 \le k \le K$. The above inequality can be expressed as follows

$$\Delta_k \le \alpha_k \Delta_{k-1}^{1/2} + \beta_k,$$

where

$$\alpha_k = C\sqrt{\frac{C^\star S\iota^4}{T_k\left(1-\gamma\right)^4}} = 2^{-k}A \qquad \text{with} \qquad A = C\sqrt{\frac{C^\star S\iota^4}{\left(1-\gamma\right)^4}}$$

and

$$\beta_k = C\sqrt{\frac{C^\star S\iota^3}{T_k\left(1-\gamma\right)^3}} + C\frac{C^\star S\iota^4}{T_k\left(1-\gamma\right)^4} + C\frac{C^\star S t_{\mathsf{mix}}\iota^2}{T_k\left(1-\gamma\right)^2} + C\frac{C^\star t_{\mathsf{mix}}\iota^3}{T_k\left(1-\gamma\right)^3}$$

for some sufficiently large constant $C > 0$. In addition, it is also observed that

$$\Delta_0 \le \frac{1}{1-\gamma}.$$

By induction, for each $1 \le j \le K$ we have

$$\begin{aligned}\Delta_j \le\ & \underbrace{\beta_j}_{=:\delta_j} + \underbrace{\alpha_j\beta_{j-1}^{1/2}}_{=:\delta_{j-1}} + \underbrace{\alpha_j\alpha_{j-1}^{1/2}\beta_{j-2}^{1/4}}_{=:\delta_{j-2}} + \cdots + \underbrace{\alpha_j\alpha_{j-1}^{1/2}\alpha_{j-2}^{1/4}\cdots\alpha_2^{1/2^{j-2}}\beta_1^{1/2^{j-1}}}_{=:\delta_1} \\ & + \underbrace{\alpha_j\alpha_{j-1}^{1/2}\alpha_{j-2}^{1/4}\cdots\alpha_2^{1/2^{j-2}}\alpha_1^{1/2^{j-1}}\Delta_0^{1/2^j}}_{=:\delta_0}.\end{aligned}$$

In the sequel, we bound each term δ_i, $0 \le i \le j$ separately.

- Let us begin with δ_0, which can be calculated as follows

$$\alpha_j\alpha_{j-1}^{1/2}\alpha_{j-2}^{1/4}\cdots\alpha_2^{1/2^{j-2}}\alpha_1^{1/2^{j-1}} = A^{2-1/2^{j-1}}2^{-j-\frac{j-1}{2}-\frac{j-2}{4}-\cdots-\frac{1}{2^{j-1}}}.$$

  Note that

$$\begin{aligned}j + \frac{j-1}{2} + \frac{j-2}{4} + \cdots + \frac{1}{2^{j-1}} &= \sum_{k=0}^{j-1}\frac{j-k}{2^k} = j\sum_{k=0}^{j-1}\frac{1}{2^k} - \sum_{k=0}^{j-1}\frac{k}{2^k} \\ &= j\left(2-\frac{1}{2^{j-1}}\right) - 2 + \frac{j+1}{2^{j-1}} \\ &= 2j - 2 + \frac{1}{2^{j-1}},\end{aligned}$$

  where the penultimate line holds since

$$\sum_{k=0}^{j-1}\frac{k}{2^k} = \sum_{k=1}^{j-1}\frac{k}{2^{k-1}} - \sum_{k=0}^{j-1}\frac{k}{2^k} = \sum_{k=0}^{j-2}\frac{k+1}{2^k} - \sum_{k=0}^{j-1}\frac{k}{2^k} = \sum_{k=0}^{j-2}\frac{1}{2^k} - \frac{j-1}{2^{j-1}} = 2 - \frac{j+1}{2^{j-1}}.$$

  Therefore we have

$$\begin{aligned}\delta_0 = A^{2-1/2^{j-1}}4^{-j+1-1/2^j}\Delta_0^{1/2^j} &\asymp \frac{1}{4^j}\left[C\sqrt{\frac{C^\star S\iota^4}{\left(1-\gamma\right)^4}}\right]^{2-1/2^{j-1}}\left(\frac{1}{1-\gamma}\right)^{1/2^j} \\ &\lesssim \frac{1}{T_j}\left[\frac{C^\star S\iota^4}{\left(1-\gamma\right)^4}\right]^{1-1/2^j}\left(\frac{1}{1-\gamma}\right)^{1/2^j} \lesssim \frac{C^\star S\iota^4}{T_j\left(1-\gamma\right)^4}.\end{aligned}$$

- Next, we develop a uniform bound on every δ_i, $1 \le i \le j-1$. We first observe that

$$\alpha_j\alpha_{j-1}^{1/2}\alpha_{j-2}^{1/4}\cdots\alpha_{j+1}^{1/2^{j-i-1}}\beta_i^{1/2^{j-i}} = A^{2-1/2^{j-i-1}}2^{-j-\frac{j-1}{2}-\frac{j-2}{4}-\cdots-\frac{i+1}{2^{j-i-1}}}\beta_i^{1/2^{j-i}},$$

and

$$
\begin{aligned}
j + \frac{j-1}{2} + \frac{j-2}{4} + \cdots + \frac{j-j}{2^j} &= \sum_{k=0}^{j-i} \frac{j-i}{2^k} = j \sum_{k=0}^{j-i} \frac{1}{2^k} - \sum_{k=0}^{n-i} \frac{k}{2^k} \\
&= j\left(2 - \frac{1}{2^{j-i}}\right) - 2 + \frac{j+2-i}{2^{j-i}} \\
&= 2j - 2 + \frac{2-i}{2^{j-i}} \\
&\geq 2j - \frac{i}{2^{j-i}} - 2,
\end{aligned}
$$

where the penultimate line holds since

$$
\sum_{k=0}^{j-i} \frac{k}{2^k} = 2 - \frac{j+2-i}{2^{j-i}}.
$$

These properties allow one to derive

$$
\begin{aligned}
\delta_i &\leq A^{2-1/2^{j-i-1}} 4^{-j+i/2^{j-i+1}+1} \beta_i^{1/2^{j-i}} \asymp 4^{-j+i/2^{j-i+1}} \left(\frac{C^\star S \iota^4}{(1-\gamma)^4}\right)^{1-1/2^{j-i}} \beta_i^{1/2^{j-i}} \\
&\asymp \frac{1}{T_j} \left(\frac{C^\star S \iota^4}{(1-\gamma)^4}\right)^{1-1/2^{j-i}} \left(\sqrt{\frac{C^\star S \iota^3}{(1-\gamma)^3}}\right)^{1/2^{j-i}} \\
&\quad + \frac{4^{-j/2^{j-i+1}}}{T_j} \left(\frac{C^\star S \iota^4}{(1-\gamma)^4}\right)^{1-1/2^{j-i}} \left(\frac{C^\star S \iota^4}{(1-\gamma)^4}\right)^{1/2^{j-i}} \\
&\quad + \frac{4^{-j/2^{j-i+1}}}{T_j} \left(\frac{C^\star S \iota^4}{(1-\gamma)^4}\right)^{1-1/2^{j-i}} \left(\frac{C^\star S t_{\mathsf{mix}} \iota^2}{(1-\gamma)^2}\right)^{1/2^{j-i}} \\
&\quad + \frac{4^{-j/2^{j-i+1}}}{T_j} \left(\frac{C^\star S \iota^4}{(1-\gamma)^4}\right)^{1-1/2^{j-i}} \left(\frac{C^\star t_{\mathsf{mix}} \iota^3}{(1-\gamma)^3}\right)^{1/2^{j-i}} \\
&\lesssim \frac{C^\star S \iota^4}{T_j (1-\gamma)^4} + \frac{1}{T_j} \sqrt{\frac{C^\star S \iota^3}{(1-\gamma)^3}} + \frac{C^\star S t_{\mathsf{mix}} \iota^2}{T_j (1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}} \iota^3}{T_j (1-\gamma)^3}.
\end{aligned}
$$

Here, the last line has used the weighted AM-GM inequality that $\alpha x + \beta y \geq (\alpha+\beta) x^{\alpha/(\alpha+\beta)} y^{\beta/(\alpha+\beta)}$ for all $\alpha, \beta, x, y > 0$.

Armed with the above results, we can readily conclude that

$$
\begin{aligned}
\Delta_j &\leq \delta_0 + \delta_j + j \max_{1 \leq i \leq j-1} \delta_i = \delta_0 + \beta_j + j \max_{1 \leq i \leq j-1} \delta_i \\
&\lesssim \frac{C^\star S \iota^4 j}{T_j (1-\gamma)^4} + \frac{j}{T_j} \sqrt{\frac{C^\star S \iota^3}{(1-\gamma)^3}} + \frac{C^\star S t_{\mathsf{mix}} \iota^2 j}{T_j (1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}} \iota^3 j}{T_j (1-\gamma)^3} + \sqrt{\frac{C^\star S \iota^3}{T_k (1-\gamma)^3}} \\
&\lesssim \sqrt{\frac{C^\star S \iota^3}{T_j (1-\gamma)^3}} + \frac{C^\star S \iota^4 j}{T_j (1-\gamma)^4} + \frac{C^\star S t_{\mathsf{mix}} \iota^2 j}{T_j (1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}} \iota^3 j}{T_j (1-\gamma)^3},
\end{aligned}
\tag{D.15}
$$

where the last line relies on the fact that $T_j^{1/2} \asymp 2^j \gtrsim j$.

D.4 Step 3: putting all this together

Recall from (D.13) that

$$\Lambda_K \lesssim \sqrt{\frac{C^\star S\iota^2}{T(1-\gamma)^4}}\sqrt{\Delta_{K-1}} + \sqrt{\frac{C^\star S\iota}{T(1-\gamma)^3}} + \frac{C^\star S\iota^3}{T(1-\gamma)^4} + \frac{C^\star S t_{\mathsf{mix}}\iota}{T(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}}\iota^2}{T(1-\gamma)^3},$$

which invokes the fact that $T_K \asymp T$. From (D.15) and the fact that $T_{K-1} \asymp T$, $K \asymp \log T$, we see that

$$\Delta_{K-1} \lesssim \sqrt{\frac{C^\star S\iota^3}{T(1-\gamma)^3}} + \frac{C^\star S\iota^4 \log T}{T(1-\gamma)^4} + \frac{C^\star S t_{\mathsf{mix}}\iota^2 \log T}{T(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}}\iota^3 \log T}{T(1-\gamma)^3},$$

which in turn allows one to deduce that

$$\begin{aligned}
\sqrt{\frac{C^\star S\iota^2}{T(1-\gamma)^4}}\sqrt{\Delta_{k-1}} &\lesssim \underbrace{\left(\frac{C^\star S\iota^2}{T(1-\gamma)^4}\right)^{1/2}\left(\frac{C^\star S\iota^3}{T(1-\gamma)^3}\right)^{1/4}}_{=:\zeta_1} + \underbrace{\left(\frac{C^\star S\iota^2}{T(1-\gamma)^4}\right)^{1/2}\left(\frac{C^\star S\iota^4 \log T}{T(1-\gamma)^4}\right)^{1/2}}_{=:\zeta_2} \\
&\quad + \underbrace{\left(\frac{C^\star S\iota^2}{T(1-\gamma)^4}\right)^{1/2}\left(\frac{C^\star S t_{\mathsf{mix}}\iota^2 \log T}{T(1-\gamma)^2}\right)^{1/2}}_{=:\zeta_3} + \underbrace{\left(\frac{C^\star S\iota^2}{T(1-\gamma)^4}\right)^{1/2}\left(\frac{C^\star t_{\mathsf{mix}}\iota^3 \log T}{T(1-\gamma)^3}\right)^{1/2}}_{=:\zeta_4} \\
&\lesssim \sqrt{\frac{C^\star S\iota}{T(1-\gamma)^3}} + \frac{C^\star S\iota^3 \log T}{T(1-\gamma)^4} + \frac{C^\star S t_{\mathsf{mix}}\iota}{T(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}}\iota^2}{T(1-\gamma)^3}.
\end{aligned}$$

Here, the last step follows by applying the AM-GM inequality as follows:

$$\begin{aligned}
\zeta_1 &\lesssim \frac{C^\star S\iota^3}{T(1-\gamma)^4} + \sqrt{\frac{C^\star S\iota}{T(1-\gamma)^3}}, \\
\zeta_2 &\lesssim \frac{C^\star S\iota^3\sqrt{\log T}}{T(1-\gamma)^4}, \\
\zeta_3 &\lesssim \frac{C^\star S\iota^3 \log T}{T(1-\gamma)^4} + \frac{C^\star S t_{\mathsf{mix}}\iota}{T(1-\gamma)^2} \\
\zeta_4 &\lesssim \frac{C^\star S\iota^3 \log T}{T(1-\gamma)^4} + \frac{C^\star t_{\mathsf{mix}}\iota^2}{T(1-\gamma)^3}.
\end{aligned}$$

The above bounds taken collectively demonstrate that

$$\Lambda_K \lesssim \sqrt{\frac{C^\star S\iota}{T(1-\gamma)^3}} + \frac{C^\star S\iota^3 \log T}{T(1-\gamma)^4} + \frac{C^\star S t_{\mathsf{mix}}\iota}{T(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}}\iota^2}{T(1-\gamma)^3}.$$

To finish up, combine the preceding bound with (D.6) to reach

$$\begin{aligned}
V^\star(\rho) - V^{\widehat{\pi}}(\rho) = \left\langle \rho, V^\star - V^{\widehat{\pi}} \right\rangle &\overset{\text{(i)}}{\leq} \langle \rho, V^\star - V_{T_K} \rangle \overset{\text{(ii)}}{\leq} \frac{1}{T_K}\sum_{t=1}^{T_K} \langle \rho, V^\star - V_t \rangle = \Lambda_K \\
&\lesssim \sqrt{\frac{C^\star S\iota}{T(1-\gamma)^3}} + \frac{C^\star S\iota^3 \log T}{T(1-\gamma)^4} + \frac{C^\star S t_{\mathsf{mix}}\iota}{T(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}}\iota^2}{T(1-\gamma)^3},
\end{aligned}$$

where (i) holds true according to Lemma 6, and (ii) follows due to the monotonicity of V_t in t. This concludes the proof of Theorem 2.

E Auxiliary lemmas for Theorem 2

E.1 Proof of Lemma 6

Consider any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, and let $n = n_t(s,a)$. For each $1 \le i \le T_k$, define

$$k_i \coloneqq \min\Big\{\big\{0 \le j < T_k \mid j > k_{i-1}, (s_j, a_j) = (s,a)\big\},\, T_k\Big\},$$

and denote $k_0 = 0$. Clearly, each k_i is a stopping time. From the update rule in Algorithm 3, we can write

$$Q_t\left(s,a\right) = \sum_{i=1}^{n} \eta_i^n \left[r\left(s,a\right) + \gamma V_{k_i}\left(s_{k_i+1}\right) - \gamma \overline{V}\left(s_{k_i+1}\right) + \gamma \widetilde{P}_{s,a}\overline{V} - b_i\left(s,a\right)\right].$$

This taken together with the elementary fact $\sum_{i=1}^{n} \eta_i^n = 1$ gives

$$\begin{aligned}
\left(Q^\star - Q_t\right)\left(s,a\right) &= \left(r + \gamma P V^\star\right)\left(s,a\right) - \sum_{i=1}^{n} \eta_i^n \left[r\left(s,a\right) + \gamma V_{k_i}\left(s_{k_i+1}\right) - \gamma \overline{V}\left(s_{k_i+1}\right) + \gamma \widetilde{P}_{s,a}\overline{V} - b_i\left(s,a\right)\right] \\
&= \gamma \sum_{i=1}^{n} \eta_i^n P_{s,a}\left(V^\star - V_{k_i}\right) + \underbrace{\gamma \sum_{i=1}^{n} \eta_i^n \left(\left(P - P_{k_i}\right)\left(V_{k_i} - \overline{V}\right)\right)\left(s,a\right)}_{=:\,\alpha_1} \\
&\quad + \underbrace{\gamma \sum_{i=1}^{n} \eta_i^n \left(\big(P - \widetilde{P}\big)\overline{V}\right)\left(s,a\right)}_{=:\,\alpha_2} + \sum_{i=1}^{n} \eta_i^n b_i(s,a).
\end{aligned} \tag{E.1}$$

From the update rules in Algorithms 3-4 as well as Lemma 1, we see that

$$\sum_{i=1}^{n} \eta_n^i b_i(s,a) \in \left[\widetilde{\beta}_n\left(s,a\right), 2\widetilde{\beta}_n\left(s,a\right)\right], \tag{E.2}$$

where

$$\begin{aligned}
\widetilde{\beta}_n\left(s,a\right) \coloneqq\ & C_{\mathsf{b}}\sqrt{\frac{H\iota}{n}\left\{\sigma_n^{\mathsf{adv}}\left(s,\pi^\star\left(s\right)\right) - \left[\mu_n^{\mathsf{adv}}\left(s,\pi^\star\left(s\right)\right)\right]^2\right\}} + C_{\mathsf{b}}\frac{H^{3/4}\iota^{3/4}}{n^{3/4}\left(1-\gamma\right)} + C_{\mathsf{b}}\frac{H\iota}{n\left(1-\gamma\right)} \\
& + C_{\mathsf{b}}\sqrt{\frac{\iota}{n^{\mathsf{ref}}\left(s,a\right)}\left\{\sigma^{\mathsf{ref}}\left(s,a\right) - \left[\mu^{\mathsf{ref}}\left(s,a\right)\right]^2\right\}} + C_{\mathsf{b}}\frac{\iota^{3/4}}{\left(1-\gamma\right)\left[n^{\mathsf{ref}}\left(s,a\right)\right]^{3/4}} + C_{\mathsf{b}}\frac{\iota}{\left(1-\gamma\right)n^{\mathsf{ref}}\left(s,a\right)}.
\end{aligned}$$

From now on, we shall focus on the case where $a = \pi^\star(s)$. The terms α_1 and α_2 are controlled separately in the following.

- Regarding α_1, we first define a filtration $\{\mathcal{F}_i\}_{i=0}^{T_k-1}$ as

  $$\mathcal{F}_i \coloneqq \sigma\Big\{\left\{\left(s_j^k, a_j^k\right) : 1 \le j \le k_i\right\}, \cup_{j=1}^{k}\mathcal{D}_j^{\mathsf{ref}}, \cup_{j=1}^{k-1}\mathcal{D}_j\Big\}.$$

  Here, (s_i^j, a_i^j) (resp. $(s_i^{j,\mathsf{ref}}, a_i^{j,\mathsf{ref}})$) is defined to be the i-th state-action pair used to update the Q-function estimate (resp. construct the empirical transition kernel) within the j-th epoch; and we set

  $$\mathcal{D}_j^{\mathsf{ref}} \coloneqq \left\{\left(s_i^{j,\mathsf{ref}}, a_i^{j,\mathsf{ref}}\right) : 0 \le i < T_j^{\mathsf{ref}}\right\}, \qquad \mathcal{D}_j \coloneqq \left\{\left(s_i^j, a_i^j\right) : 0 \le i < T_j\right\}. \tag{E.3}$$

  It is straightforward to check that for any $1 \le \tau \le T$,

  $$\left\{\mathbb{1}_{k_i<T}\left(\left(P - P_{k_i}\right)\left(V_{k_i} - \overline{V}\right)\right)\left(s,\pi^\star(s)\right)\right\}_{i=1}^{\tau}$$

is a martingale difference sequence with respect to $\{\mathcal{F}_i\}_{i\geq 0}$. Then we can invoke the Freedman inequality to obtain: for any fixed $s\in\mathcal{S}$ and $\tau\in[T]$, with probability exceeding $1-\delta/(ST)$,

$$\begin{aligned}\left|\sum_{i=1}^{\tau}\mathbb{1}_{k_i\leq T_k}\,\eta_i^{\tau}\left(\left(P-P_{k_i}\right)\left(V_{k_i}-\overline{V}\right)\right)\left(s,\pi^{\star}(s)\right)\right| &\lesssim \sqrt{\sum_{i=1}^{\tau}\left(\eta_i^{\tau}\right)^2\mathsf{Var}_{s,\pi^{\star}(s)}\left(V_{k_i}-\overline{V}\right)\iota}+\frac{1}{1-\gamma}\max_{1\leq i\leq\tau}\eta_i^{\tau}\iota\\ &\lesssim\sqrt{\frac{H\iota}{\tau}\sum_{i=1}^{\tau}\eta_i^{\tau}\mathsf{Var}_{s,\pi^{\star}(s)}\left(V_{k_i}-\overline{V}\right)}+\frac{H\iota}{(1-\gamma)\,\tau}.\end{aligned}$$

Invoke the union bound to show that with probability at least $1-\delta$, the above inequality holds simultaneously for all $\tau\in[T]$ and $s\in\mathcal{S}$. Replacing τ with $n=n_t(s,\pi^{\star}(s))$ yields that, with probability exceeding $1-\delta$,

$$|\alpha_1|\lesssim\sqrt{\frac{H\tau}{n}\sum_{i=1}^{n}\eta_i^{n}\mathsf{Var}_{s,\pi^{\star}(s)}\left(V_{k_i}-\overline{V}\right)}+\frac{H\tau}{(1-\gamma)\,n}\tag{E.4}$$

holds for all $s\in\mathcal{S}$ and $t\in[T_k]$, where $n=n_t(s,\pi^{\star}(s))$. In addition, the update rules in Algorithm 2 tell us that

$$\mu_n^{\mathsf{adv}}(s,a)=\sum_{i=1}^{n}\eta_i^{n}\left[V_{k_i}\left(s_{k_i+1}\right)-\overline{V}\left(s_{k_i+1}\right)\right]=\sum_{i=1}^{n}\eta_i^{n}\Big(P_{k_i}\left(V_{k_i}-\overline{V}\right)\Big)(s,a)\,;\tag{E.5}$$

$$\sigma_n^{\mathsf{adv}}(s,a)=\sum_{i=1}^{n}\eta_i^{n}\left[V_{k_i}\left(s_{k_i+1}\right)-\overline{V}\left(s_{k_i+1}\right)\right]^2=\sum_{i=1}^{n}\eta_i^{n}\left(P_{k_i}\left(V_{k_i}-\overline{V}\right)^2\right)(s,a)\,.\tag{E.6}$$

Recognizing that

$$\sum_{i=1}^{n}\eta_i^{n}\mathsf{Var}_{s,a}\left(V_{k_i}-\overline{V}\right)=\sum_{i=1}^{n}\eta_i^{n}P_{s,a}\left(V_{k_i}-\overline{V}\right)^2-\sum_{i=1}^{n}\eta_i^{n}\left[P_{s,a}\left(V_{k_i}-\overline{V}\right)\right]^2,$$

we can connect $\mathsf{Var}_{s,a}\left(V_{k_i}-\overline{V}\right)$ with μ_n^{adv} and σ_n^{adv} as follows

$$\begin{aligned}&\sum_{i=1}^{n}\eta_i^{n}\mathsf{Var}_{s,\pi^{\star}(s)}\left(V_{k_i}-\overline{V}\right)-\left\{\sigma_n^{\mathsf{adv}}\left(s,\pi^{\star}(s)\right)-\left[\mu_n^{\mathsf{adv}}\left(s,\pi^{\star}(s)\right)\right]^2\right\}\\ &=\sum_{i=1}^{n}\eta_i^{n}P_{s,\pi^{\star}(s)}\left(V_{k_i}-\overline{V}\right)^2-\sum_{i=1}^{n}\eta_i^{n}\left[\left(P\left(V_{k_i}-\overline{V}\right)\right)\left(s,\pi^{\star}(s)\right)\right]^2\\ &\quad-\sum_{i=1}^{n}\eta_i^{n}\left(P_{k_i}\left(V_{k_i}-\overline{V}\right)^2\right)\left(s,\pi^{\star}(s)\right)+\left[\sum_{i=1}^{n}\eta_i^{n}\left(P_{k_i}\left(V_{k_i}-\overline{V}\right)\right)\left(s,\pi^{\star}(s)\right)\right]^2\\ &=\underbrace{\sum_{i=1}^{n}\eta_i^{n}\left(\left(P-P_{k_i}\right)\left(V_{k_i}-\overline{V}\right)^2\right)\left(s,\pi^{\star}(s)\right)}_{=:\,\alpha_{1,1}}+\underbrace{\left[\sum_{i=1}^{n}\eta_i^{n}\left(P_{k_i}\left(V_{k_i}-\overline{V}\right)\right)\left(s,\pi^{\star}(s)\right)\right]^2-\sum_{i=1}^{n}\eta_i^{n}\left[P_{s,\pi^{\star}(s)}\left(V_{k_i}-\overline{V}\right)\right]^2}_{=:\,\alpha_{1,2}},\end{aligned}$$

leaving us with two terms to control.

– The first term $\alpha_{1,1}$ can be bounded by the Azuma-Hoeffding inequality. We can employ similar arguments as used when proving (C.3) and invoke the Azuma-Hoeffding inequality to show that: with probability exceeding $1-\delta/S$, for all $s\in\mathcal{S}$ and $t\in[T_k]$, it holds that

$$|\alpha_{1,1}|\lesssim\sqrt{\frac{H\iota}{n\left(1-\gamma\right)^4}}.\tag{E.7}$$

– Moving on to the second term $\alpha_{1,2}$, we invoke the identity $\sum_{i=1}^{n}\eta_i^n = 1$ to deduce that

$$\begin{aligned}
\alpha_{1,2} &= \left[\sum_{i=1}^{n}\eta_i^n\left(P_{k_i}\left(V_{k_i}-\overline{V}\right)\right)\left(s,\pi^\star(s)\right)\right]^2 - \left(\sum_{i=1}^{n}\eta_i^n\right)\sum_{i=1}^{n}\eta_i^n\left[P_{s,\pi^\star(s)}\left(V_{k_i}-\overline{V}\right)\right]^2 \\
&\overset{\text{(i)}}{\leq} \left[\sum_{i=1}^{n}\eta_i^n\left(P_{k_i}\left(V_{k_i}-\overline{V}\right)\right)\left(s,\pi^\star(s)\right)\right]^2 - \left[\sum_{i=1}^{n}\eta_i^n P_{s,\pi^\star(s)}\left(V_{k_i}-\overline{V}\right)\right]^2 \\
&= \left[\sum_{i=1}^{n}\eta_i^n\left(\left(P_{k_i}-P\right)\left(V_{k_i}-\overline{V}\right)\right)\left(s,\pi^\star(s)\right)\right]\left[\sum_{i=1}^{n}\eta_i^n\left(\left(P_{k_i}+P\right)\left(V_{k_i}-\overline{V}\right)\right)\left(s,\pi^\star(s)\right)\right] \\
&\overset{\text{(ii)}}{\leq} \frac{1}{1-\gamma}\left|\sum_{i=1}^{n}\eta_i^n\left(\left(P_{k_i}-P\right)\left(V_{k_i}-\overline{V}\right)\right)\left(s,\pi^\star(s)\right)\right| \\
&\overset{\text{(iii)}}{\lesssim} \sqrt{\frac{H\iota}{n(1-\gamma)^4}}.
\end{aligned}$$

Here, (i) arises from the Cauchy-Schwarz inequality; (ii) follows from the fact that $0 \leq V_{k_i} - \overline{V} \leq 1/(1-\gamma)$ and the identity $\sum_{i=1}^{n}\eta_i^n = 1$; and (iii) follows by repeating the argument used to establish (C.3) and invoking the Azuma-Hoeffding inequality (which we omite here for the sake of brevity).

With the preceding bounds in place, we conclude that with probability exceeding $1-O(\delta)$,

$$\begin{aligned}
\sum_{i=1}^{n}\eta_n^i\mathsf{Var}_{s,\pi^\star(s)}\left(V_{k_i}-\overline{V}\right) &\leq \sigma_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right) - \left[\mu_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right)\right]^2 + \alpha_{1,1} + \alpha_{1,2} \\
&\leq \sigma_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right) - \left[\mu_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right)\right]^2 + O\left(\sqrt{\frac{H\iota}{n(1-\gamma)^4}}\right)
\end{aligned}$$

holds for all $s \in \mathcal{S}$ and $t \in [T_k]$. Putting the above results together and using the fact $\sigma_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right) \geq \left[\mu_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right)\right]^2$ (due to Jensen's inequality) reveal that with probability exceeding $1-O(\delta)$,

$$\begin{aligned}
|\alpha_1| &\lesssim \sqrt{\frac{H\iota}{n}\sum_{i=1}^{n}\eta_i^n\mathsf{Var}_{s,\pi^\star(s)}\left(V_{k_i}-\overline{V}\right)} + \frac{H\iota}{(1-\gamma)n} \\
&\lesssim \sqrt{\frac{H\iota}{n}\left\{\sigma_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right) - \left[\mu_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right)\right]^2 + O\left(\sqrt{\frac{H\iota}{n(1-\gamma)^4}}\right)\right\}} + \frac{H\iota}{n(1-\gamma)} \\
&\lesssim \sqrt{\frac{H\iota}{n}\left\{\sigma_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right) - \left[\mu_n^{\mathsf{adv}}\left(s,\pi^\star(s)\right)\right]^2\right\}} + \frac{H^{3/4}\iota^{3/4}}{n^{3/4}(1-\gamma)} + \frac{H\iota}{n(1-\gamma)}
\end{aligned}$$

holds for all $s \in \mathcal{S}$ and $t \in [T_k]$.

- Regarding α_2, we first recall that $n^{\mathsf{ref}}(s,a)$ denotes the number of visit to (s,a) among the samples used to compute $\widetilde{P}$. Let $k_0 = -1$, and for each $1 \leq i \leq T_k^{\mathsf{ref}}$, define

$$k_i \coloneqq \min\left\{\left\{0 \leq k < T_k^{\mathsf{ref}} \mid k > k_{i-1}, (s_k,a_k) = (s,a)\right\}, T_k^{\mathsf{ref}}\right\}.$$

Akin to how we establish (E.4), we can use the Freedman inequality to show that: for any fixed $s \in \mathcal{S}$, with probability exceeding $1-\delta/S$,

$$\begin{aligned}
|\alpha_2| &= \left|\gamma\left(\left(P-\widetilde{P}\right)\overline{V}\right)\left(s,\pi^\star(s)\right)\right| \\
&= \gamma\left|\frac{1}{n^{\mathsf{ref}}\left(s,\pi^\star(s)\right)}\sum_{i=0}^{T_k^{\mathsf{ref}}}\left(P-P_i^{\mathsf{ref}}\right)_{s,a}\overline{V}\,\mathbb{1}\left\{\left(s_i^{\mathsf{ref}},a_i^{\mathsf{ref}}\right) = \left(s,\pi^\star(s)\right)\right\}\right|
\end{aligned}$$

$$
\begin{aligned}
&= \gamma \left| \frac{1}{n^{\mathsf{ref}}\big(s, \pi^{\star}(s)\big)} \sum_{i=0}^{n^{\mathsf{ref}}(s,\pi^{\star}(s))} \big(P - P_{k_i}^{\mathsf{ref}}\big)_{s,\pi^{\star}(s)} \overline{V} \right| \\
&\lesssim \sqrt{\frac{\mathsf{Var}_{s,\pi^{\star}(s)}(\overline{V})}{n^{\mathsf{ref}}\big(s, \pi^{\star}(s)\big)} \iota} + \frac{\iota}{(1-\gamma)\, n^{\mathsf{ref}}\big(s, \pi^{\star}\,(s)\,\big)}.
\end{aligned}
$$

Here, the first line results from the identity $\sum_{i=1}^{n} \eta_i^n = 1$. It follows from the update rule in Algorithm 2 that

$$
\mu^{\mathsf{ref}}(s,a) = \frac{1}{n^{\mathsf{ref}}(s,a)} \sum_{i=1}^{n^{\mathsf{ref}}(s,a)} \overline{V}\left(s_{k_i+1}\right) = \frac{1}{n^{\mathsf{ref}}(s,a)} \sum_{i=1}^{n^{\mathsf{ref}}(s,a)} \big(P_{k_i}\overline{V}\big)(s,a), \tag{E.8}
$$

$$
\sigma^{\mathsf{ref}}(s,a) = \frac{1}{n^{\mathsf{ref}}(s,a)} \sum_{i=1}^{n^{\mathsf{ref}}(s,a)} \overline{V}^2\left(s_{k_i+1}\right) = \frac{1}{n^{\mathsf{ref}}(s,a)} \sum_{i=1}^{n^{\mathsf{ref}}(s,a)} \big(P_{k_i}\overline{V}^2\big)(s,a), \tag{E.9}
$$

allowing us to deduce that

$$
\begin{aligned}
&\left| \mathsf{Var}_{s,\pi^{\star}(s)}(\overline{V}) - \sigma^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big) + \big[\mu^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big)\big]^2 \right| \\
&= \left| P_{s,\pi^{\star}(s)}(\overline{V}^2) - (P_{s,\pi^{\star}(s)}\overline{V})^2 - \sigma^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big) + \big[\mu^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big)\big]^2 \right| \\
&\le \underbrace{\left| P_{s,\pi^{\star}(s)}(\overline{V}^2) - \frac{1}{n^{\mathsf{ref}}\big(s,\pi^{\star}\,(s)\,\big)} \sum_{i=1}^{n^{\mathsf{ref}}(s,\pi^{\star}(s))} \Big(P_{k_i}\overline{V}^2\Big)\big(s,\pi^{\star}(s)\big) \right|}_{=:\alpha_{2,1}} \\
&\quad + \underbrace{\left| \left[\frac{1}{n^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big)} \sum_{i=1}^{n^{\mathsf{ref}}(s,\pi^{\star}(s))} \big(P_{k_i}\overline{V}\big)\big(s,\pi^{\star}(s)\big) \right]^2 - \big(P_{s,\pi^{\star}(s)}\overline{V}\big)^2 \right|}_{=:\alpha_{2,2}}.
\end{aligned}
$$

Using the similar argument in proving (C.3) and the Azuma-Hoeffding inequality, we can show that with probability exceeding $1-\delta/S$,

$$
\alpha_{2,1} \lesssim \frac{1}{(1-\gamma)^2} \sqrt{\frac{\iota}{n^{\mathsf{ref}}(s,a)}}.
$$

The second term $\alpha_{2,2}$ can be bounded by

$$
\begin{aligned}
\alpha_{2,2} &= \left| \left[\frac{1}{n^{\mathsf{ref}}(s,\pi^{\star}\,(s))} \sum_{i=1}^{n^{\mathsf{ref}}(s,\pi^{\star}(s))} \big((P_{k_i} - P)\overline{V}\big)\big(s,\pi^{\star}(s)\big) \right] \left[\frac{1}{n^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big)} \sum_{i=1}^{n^{\mathsf{ref}}(s,\pi^{\star}(s))} \big((P_{k_i} + P)\overline{V}\big)\big(s,\pi^{\star}(s)\big) \right] \right| \\
&\le \frac{2}{1-\gamma} \left| \frac{1}{n^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big)} \sum_{i=1}^{n^{\mathsf{ref}}(s,\pi^{\star}(s))} \big((P_{k_i} - P)\overline{V}\big)\big(s,\pi^{\star}(s)\big) \right| \\
&\lesssim \frac{1}{(1-\gamma)^2} \sqrt{\frac{\iota}{n^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big)}}.
\end{aligned}
$$

Here, the penultimate line follows from the fact that $0 \le \overline{V}(s) \le 1/(1-\gamma)$ for all $s \in \mathcal{S}$, whereas the last line can be proved by using the similar argument used to establish (C.3) and invoking the Azuma-Hoeffding inequality. These bounds taken collectively allow us to derive

$$
\begin{aligned}
\mathsf{Var}_{s,\pi^{\star}(s)}(\overline{V}) &= \sigma^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big) - \big[\mu^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big)\big]^2 + O\left(\alpha_{2,1} + \alpha_{2,2}\right) \\
&= \sigma^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big) - \big[\mu^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big)\big]^2 + O\left(\frac{1}{(1-\gamma)^2} \sqrt{\frac{\iota}{n^{\mathsf{ref}}\big(s,\pi^{\star}(s)\big)}} \right).
\end{aligned} \tag{E.10}
$$

Consequently, it is immediately seen that: with probability exceeding $1 - O(\delta)$,

$$
\begin{aligned}
|\alpha_2| &\lesssim \sqrt{\frac{\iota}{n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)}\left[\sigma^{\mathsf{ref}}\big(s,\pi^\star(s)\big) - \big[\mu^{\mathsf{ref}}\big(s,\pi^\star(s)\big)\big]^2 + O\left(\frac{1}{(1-\gamma)^2}\sqrt{\frac{1}{n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)}}\right)\right]} + \frac{\iota}{(1-\gamma)\,n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)} \\
&\lesssim \sqrt{\frac{\iota}{n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)}\left\{\sigma^{\mathsf{ref}}\big(s,\pi^\star(s)\big) - \big[\mu^{\mathsf{ref}}\big(s,\pi^\star(s)\big)\big]^2\right\}} + \frac{\iota^{3/4}}{(1-\gamma)\left[n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)\right]^{3/4}} + \frac{\iota}{(1-\gamma)\,n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)}
\end{aligned}
$$

holds simultaneously for all $s \in \mathcal{S}$.

With the above bounds on α_1 and α_2 in place, we can take these together with (E.2) to obtain

$$
0 \le \sum_{i=1}^{n} \eta_i^n b_i\big(s,\pi^\star(s)\big) + \alpha_1 + \alpha_2 \le 3\widetilde{\beta}_n\big(s,\pi^\star(s)\big) = \beta_n\big(s,\pi^\star(s)\big),
$$

with the proviso that $C_{\mathsf{b}} > 0$ is sufficiently large. Substitution into (E.1) then gives: with probability exceeding $1 - O(\delta)$,

$$
\left(Q^\star - Q_t\right)\big(s,\pi^\star(s)\big) \le \gamma \sum_{i=1}^{n} \eta_i^n P_{s,\pi^\star(s)}\left(V^\star - V_{k_i}\right) + \beta_n\big(s,\pi^\star(s)\big)
$$

holds for all $s \in \mathcal{S}$ and $t \in [T_k]$.

The second part of the lemma — namely, $V_t(s) \le V^{\pi_t}(s) \le V^\star(s)$ for all $s \in \mathcal{S}$ and $t \in [T_k]$ — can be proved in a way similar to the proof of the second part of Lemma 2. We omit it here for brevity.

E.2 Proof of Lemma 7

In view of (E.6), we can deduce that

$$
\begin{aligned}
\sigma_n^{\mathsf{adv}}\big(s,\pi^\star(s)\big) &= \sum_{i=1}^{n} \eta_i^n \left(P_{k_i}\left(V_{k_i} - \overline{V}\right)^2\right)\big(s,\pi^\star(s)\big) \\
&= \sum_{i=1}^{n} \eta_i^n P_{s,\pi^\star(s)}\left(V_{k_i} - \overline{V}\right)^2 + \sum_{i=1}^{n} \eta_i^n \left(\left(P_{k_i} - P\right)\left(V_{k_i} - \overline{V}\right)^2\right)\big(s,\pi^\star(s)\big) \\
&\le P_{s,\pi^\star(s)}\left(V^\star - \overline{V}\right)^2 + \sum_{i=1}^{n} \eta_i^n \left(\left(P_{k_i} - P\right)\left(V_{k_i} - \overline{V}\right)^2\right)\big(s,\pi^\star(s)\big) \\
&\le P_{s,\pi^\star(s)}\left(V^\star - \overline{V}\right)^2 + O\left(\sqrt{\frac{H\iota}{n\,(1-\gamma)^4}}\right).
\end{aligned}
$$

Here, the penultimate line follows from the fact that V_t is non-decreasing in t and $\overline{V} \le V_t \le V^\star$, while the last inequality invokes the upper bound on $\alpha_{1,1}$ (cf. (E.7)) derived in Lemma 6. In addition, we observe that

$$
\sigma^{\mathsf{ref}}\big(s,\pi^\star(s)\big) - \big[\mu^{\mathsf{ref}}\big(s,\pi^\star(s)\big)\big]^2 = \mathsf{Var}_{s,\pi^\star(s)}(\overline{V}) + O\left(\frac{1}{(1-\gamma)^2}\sqrt{\frac{\iota}{n^{\mathsf{ref}}\big(s,\pi^\star(s)\big)}}\right),
$$

which follows directly from (E.10).

Next, we turn to bounding the sum $\sum_{s,a} d_\rho^\star(s,a)\mathsf{Var}_{s,a}(\overline{V})$, which can be decomposed into

$$
\sum_{s,a} d_\rho^\star(s,a)\,\mathsf{Var}_{s,a}(\overline{V}) = \underbrace{\sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\rho^\star(s,a)\,\mathsf{Var}_{s,a}(V^\star)}_{=:\alpha_1} + \underbrace{\sum_{s,a} d_\rho^\star(s,a)\left[\mathsf{Var}_{s,a}(\overline{V}) - \mathsf{Var}_{s,a}(V^\star)\right]}_{=:\alpha_2}. \tag{E.11}
$$

This leaves us with two terms α_1 and α_2 to control.

- With regards to α_1, we first define a vector $v = [v_s]_{s\in\mathcal{S}} \in \mathbb{R}^S$ obeying

$$v_s := \mathsf{Var}_{s,\pi^\star(s)}\left(V^\star\right) \qquad \text{for all } s \in \mathcal{S},$$

which clearly satisfies

$$\begin{aligned} v &= P_{\pi^\star}\left[V^\star \circ V^\star\right] - \left(P_{\pi^\star}V^\star\right) \circ \left(P_{\pi^\star}V^\star\right) \\ &= P_{\pi^\star}\left(V^\star \circ V^\star\right) - \frac{1}{\gamma^2}\left(r - V^\star\right) \circ \left(r - V^\star\right) \\ &= P_{\pi^\star}\left(V^\star \circ V^\star\right) - \frac{1}{\gamma^2} r \circ r - \frac{1}{\gamma^2} V^\star \circ V^\star + 2V^\star \circ r \\ &\leq \frac{1}{\gamma^2}\left(\gamma^2 P_{\pi^\star} - I\right)\left(V^\star \circ V^\star\right) + 2V^\star \circ r. \end{aligned} \tag{E.12}$$

Here, the second identity follows from the Bellman optimality equation $V^\star = r + \gamma P_{\pi^\star}V^\star$. Recognizing that $d_\rho^\star(s,a) = d_\rho^\star(s)\,\mathbb{1}\{a = \pi^\star(s)\}$ and $d_\rho^\star = (1-\gamma)\rho(I - \gamma P_{\pi^\star})^{-1}$, we obtain

$$\begin{aligned} \alpha_1 &= \sum_{s\in\mathcal{S}} d_\rho^\star(s)\,\mathsf{Var}_{s,\pi^\star(s)}(V^\star) = \left\langle d_\rho^\star, v\right\rangle = (1-\gamma)\,\rho\left(I - \gamma P_{\pi^\star}\right)^{-1} v \\ &\leq (1-\gamma)\,\|\rho\|_1 \left\|\left(I - \gamma P_{\pi^\star}\right)^{-1} v\right\|_\infty = (1-\gamma)\left\|\left(I - \gamma P_{\pi^\star}\right)^{-1} v\right\|_\infty \\ &\overset{\text{(i)}}{\leq} \frac{1-\gamma}{\gamma^2}\left\|\left(I - \gamma P_{\pi^\star}\right)^{-1}\left(\gamma^2 P_{\pi^\star} - I\right)\left(V^\star \circ V^\star\right)\right\|_\infty + 2(1-\gamma)\left\|\left(I - \gamma P_{\pi^\star}\right)^{-1}\left(V^\star \circ r\right)\right\|_\infty \\ &= \frac{1-\gamma}{\gamma^2}\left\|\left(I - \gamma P_{\pi^\star}\right)^{-1}\left[(1-\gamma)I + \gamma\left(I - \gamma P_{\pi^\star}\right)\right]\left(V^\star \circ V^\star\right)\right\|_\infty + 2(1-\gamma)\|V^\star\|_\infty\left\|\left(I - \gamma P_{\pi^\star}\right)^{-1} r\right\|_\infty \\ &\overset{\text{(ii)}}{\leq} \frac{(1-\gamma)^2}{\gamma^2}\left\|\left(I - \gamma P_{\pi^\star}\right)^{-1}\left(V^\star \circ V^\star\right)\right\|_\infty + \frac{1-\gamma}{\gamma}\|V^\star \circ V^\star\|_\infty + 2(1-\gamma)\|V^\star\|_\infty^2 \\ &\leq \frac{(1-\gamma)^2}{\gamma^2}\frac{1}{1-\gamma}\|V^\star\|_\infty^2 + \frac{1-\gamma}{\gamma}\|V^\star\|_\infty^2 + 2(1-\gamma)\|V^\star\|_\infty^2 \\ &\overset{\text{(iii)}}{\leq} \frac{8}{1-\gamma}. \end{aligned}$$

Here, (i) follows from (E.12); (ii) relies on the triangle inequality as well as the Bellman optimality equation $V^\star = (I - \gamma P_{\pi^\star})^{-1} r$; and (iii) arises from the property $0 \leq V^\star(s) \leq 1/(1-\gamma)$ for all $s \in \mathcal{S}$ as well as the assumption that $\gamma \geq 1/2$.

- Regarding α_2, we make the observation that

$$\begin{aligned} \alpha_2 &= \sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\rho^\star(s,a)\left\{P_{s,a}\overline{V}^2 - \left[P_{s,a}\overline{V}\right]^2 - P_{s,a}\big(V^{\star 2}\big) + \left[P_{s,a}V^\star\right]^2\right\} \\ &= \sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\rho^\star(s,a)\left\{P_{s,a}\overline{V}^2 - P_{s,a}\big(V^{\star 2}\big)\right\} + \sum_{s,a} d_\rho^\star(s,a)\left\{\left[P_{s,a}V^\star\right]^2 - \left[P_{s,a}\overline{V}\right]^2\right\} \\ &\leq \sum_{s,a} d_\rho^\star(s,a)\,P_{s,a}\left(V^\star - \overline{V}\right)P_{s,a}\left(V^\star + \overline{V}\right) \\ &\leq \frac{2}{1-\gamma}\sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\rho^\star(s,a)\,P_{s,a}\left(V^\star - \overline{V}\right), \end{aligned} \tag{E.13}$$

where the third line holds since $\overline{V}^2 \leq V^{\star 2}$, and the last line is valid since $\|P_{s,a}\|_1 = 1$ and $\|\overline{V}\|_\infty \leq \|V^\star\|_\infty \leq \frac{1}{1-\gamma}$. Recognizing that

$$d_\rho^\star = (1-\gamma)\rho + \gamma d_\rho^\star P_{\pi^\star}, \tag{E.14}$$

we can use the fact $d_\rho^\star(s,a) = d_\rho^\star(s)\,\mathbb{1}\{a = \pi^\star(s)\}$ to derive

$$\sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\rho^\star(s,a)\,P_{s,a}\big(V^\star - \overline{V}\big) = \sum_{s\in\mathcal{S}} d_\rho^\star(s)P_{s,\pi^\star(s)}\big(V^\star - \overline{V}\big) = \left\langle d_\rho^\star P_{\pi^\star}, V^\star - \overline{V}\right\rangle$$

$$= \left\langle \frac{d_\rho^\star - (1-\gamma)\,\rho}{\gamma}, V^\star - \overline{V} \right\rangle$$
$$= \left\langle \widetilde{\rho}, V^\star - \overline{V} \right\rangle = \Delta_{k-1}.$$

Substitution into (E.13) leads to

$$\alpha_2 \le \frac{2}{1-\gamma}\Delta_{k-1}.$$

- Take the preceding bounds on α_1 and α_2 together with (E.11) to yield

$$\sum_{s,a} d_\rho^\star(s,a)\,\mathsf{Var}_{s,a}(\overline{V}) \le \alpha_1 + \alpha_2 \le \frac{8}{1-\gamma} + \frac{2}{1-\gamma}\Delta_{k-1}.$$

Finally, we turn attention to $\sum_{s,a} d_\rho^\star(s,a)\,\mathsf{Var}_{s,a}(V^\star - \overline{V})$. This sum can be bounded as follows

$$\begin{aligned}
\sum_{s,a} d_\rho^\star(s,a)\,\mathsf{Var}_{s,a}(V^\star - \overline{V}) &\le \sum_{s\in\mathcal{S},a\in\mathcal{A}} d^\star(s,a)\,P_{s,a}\left(V^\star - \overline{V}\right)^2 = \sum_{s\in\mathcal{S}} d^\star(s)\,P_{s,\pi^\star(s)}\left(V^\star - \overline{V}\right)^2 \\
&= \left\langle \frac{d_\rho^\star - (1-\gamma)\,\rho}{\gamma}, \left(V^\star - \overline{V}\right)^2 \right\rangle \\
&= \sum_{s\in\mathcal{S}} \frac{d^\star(s) - (1-\gamma)\,\rho(s)}{\gamma}\left(V^\star - \overline{V}\right)^2(s) \\
&= \left\langle \widetilde{\rho}, \left(V^\star - \overline{V}\right)^2 \right\rangle \le \left\|V^\star - \overline{V}\right\|_\infty \left\langle \widetilde{\rho}, V^\star - \overline{V} \right\rangle \\
&\le \frac{1}{1-\gamma}\Delta_{k-1},
\end{aligned}$$

where the first identity holds since $d^\star(s,a) = d^\star(s)\,\mathbb{1}\{a = \pi^\star(s)\}$, and the second line invokes (E.14).

E.3 Proof of Lemma 8

Recall that

$$\widetilde{\rho} \coloneqq \frac{d_\rho^\star - (1-\gamma)\,\rho}{\gamma},$$

which is clearly also a probability vector. To prove the lemma, we find it helpful to introduce the following occupancy distribution induced by $\widetilde{\rho}$:

$$d_{\widetilde{\rho}}^\star(s) \coloneqq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\left(s_t = s \mid \pi^\star, s_0 \sim \widetilde{\rho}\right).$$

Repeating the argument used to establish (B.6), we can easily see that: for any vector $V \in \mathbb{R}^d$ with non-negative entries, it holds that

$$\sum_{j=0}^{\infty}\left[\gamma\left(1+\frac{1}{H}\right)^3\right]^j \left\langle \widetilde{\rho}P_{\pi^\star}^j, V\right\rangle \lesssim \frac{1}{1-\gamma}\left\langle d_{\widetilde{\rho}}^\star, V\right\rangle + \frac{\delta}{ST^4(1-\gamma)}\|V\|_\infty. \tag{E.15}$$

Consequently, it boils down to analyzing the distribution $d_{\widetilde{\rho}}^\star$.

For any integer $K \ge 0$, employ the identity $d_\rho^\star = \rho\sum_{i=0}^{\infty}\gamma^i(P_{\pi^\star})^i$ to deduce that

$$\begin{aligned}
d_{\widetilde{\rho}}^\star(s) &= (1-\gamma)\left[\sum_{i=0}^{\infty}\gamma^i\frac{d_\rho^\star - (1-\gamma)\,\rho}{\gamma}(P_{\pi^\star})^i\right](s) \\
&\overset{\text{(i)}}{=} \frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\gamma^{i+j}\rho(P_{\pi^\star})^{i+j}\right](s) - \frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=0}^{\infty}\gamma^i\rho(P_{\pi^\star})^i\right](s)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(1-\gamma)^2}{\gamma}\left[\rho\left(\sum_{l=0}^{\infty}\gamma^l(P_{\pi^\star})^l\right)^2\right](s) - \frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=0}^{\infty}\gamma^i\rho(P_{\pi^\star})^i\right](s)\\
&= \frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=0}^{\infty}(i+1)\,\gamma^i\rho(P_{\pi^\star})^i\right](s) - \frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=0}^{\infty}\gamma^i\rho(P_{\pi^\star})^i\right](s)\\
&= \frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=0}^{\infty}i\gamma^i\rho(P_{\pi^\star})^i\right](s)\\
&= \frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=0}^{K-1}i\gamma^i\rho(P_{\pi^\star})^i\right](s) + \frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=K}^{\infty}i\gamma^i\rho(P_{\pi^\star})^i\right](s)\\
&\leq K\frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=0}^{K-1}\gamma^i\rho(P_{\pi^\star})^i\right](s) + \frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=K}^{\infty}i\gamma^i\rho(P_{\pi^\star})^i\right](s)\\
&\overset{\text{(ii)}}{\leq} 2K\,(1-\gamma)\,d_\rho^\star(s) + \underbrace{\frac{(1-\gamma)^2}{\gamma}\left[\sum_{i=K}^{\infty}i\gamma^i\rho(P_{\pi^\star})^i\right](s)}_{=:\,e(s)}.
\end{aligned}
$$

Here, (i) and (ii) make use of the identity $d_\rho^\star = (1-\gamma)\rho\sum_{j=0}^{\infty}\gamma^j(P_{\pi^\star})^j$ and the assumption that $\gamma \geq 1/2$. By choosing

$$K \coloneqq \left\lceil \frac{C_K}{1-\gamma}\log\frac{ST}{\delta}\right\rceil$$

for some constant $C_K > 0$, we can guarantee that

$$d_{\widetilde{\rho}}^\star(s) \leq 4C_K d_\rho^\star(s)\log\frac{ST}{\delta} + e\,(s)\,. \tag{E.16}$$

This inequality further motivates us to bound $e(s)$. Towards this, note that $e(s)$ satisfies

$$
\begin{aligned}
\sum_{s\in\mathcal{S}} e\,(s) &= \frac{(1-\gamma)^2}{\gamma}\sum_{i=K}^{\infty}i\gamma^i\rho(P_{\pi^\star})^i 1\\
&= \frac{(1-\gamma)^2}{\gamma}\sum_{i=K}^{\infty}i\gamma^i = \frac{1-\gamma}{\gamma}\left(\sum_{i=K}^{\infty}i\gamma^i - \gamma\sum_{i=K}^{\infty}i\gamma^i\right)\\
&= \frac{1-\gamma}{\gamma}\left(\sum_{i=K}^{\infty}i\gamma^i - \sum_{i=K+1}^{\infty}(i-1)\,\gamma^i\right)\\
&= \frac{1-\gamma}{\gamma}\left(K\gamma^K + \sum_{i=K+1}^{\infty}\gamma^i\right)\\
&= \frac{1-\gamma}{\gamma}\left(K\gamma^K + \frac{\gamma^{K+1}}{1-\gamma}\right)\\
&\leq 2C_K\gamma^{K-1}\log\frac{ST}{\delta} + \gamma^K \lesssim \frac{\delta}{ST^4}
\end{aligned}
\tag{E.17}
$$

with 1 the all-one vector, where the second line holds since $\rho(P_{\pi^\star})^i$ remains a probability vector (and hence $\rho(P_{\pi^\star})^i 1 = 1$). Here, the last line follows from our assumption that $\gamma \geq 1/2$ and the fact that

$$\gamma^K = e^{K\log[1-(1-\gamma)]} \leq e^{-K(1-\gamma)} \leq e^{-C_K\log(ST/\delta)} = \left(\frac{\delta}{ST}\right)^{-C_K} \leq \frac{\delta^2}{S^2T^5},$$

provided that $C_K \geq 5$.

We are now ready to establish the claim of this lemma. Substituting the bounds (E.16) and (E.17) into (E.15) leads to

$$\begin{aligned}\left\langle d_{\widetilde{\rho}}^{\star}, V\right\rangle &\lesssim \left\langle d_{\rho}^{\star}, V\right\rangle \log \frac{ST}{\delta} + \sum_{s\in\mathcal{S}} e\left(s\right) \left\|V\right\|_{\infty} \\ &\lesssim \left\langle d_{\rho}^{\star}, V\right\rangle \log \frac{ST}{\delta} + \frac{\delta}{ST^{4}} \left\|V\right\|_{\infty}.\end{aligned}$$

As a result, one can readily conclude that

$$\sum_{j=0}^{\infty}\left[\gamma\left(1+\frac{1}{H}\right)^{3}\right]^{j}\left\langle \widetilde{\rho}P_{\pi^{\star}}^{j}, V\right\rangle \lesssim \frac{1}{1-\gamma}\left\langle d_{\rho}^{\star}, V\right\rangle \log\frac{ST}{\delta} + \frac{\delta}{ST^{4}\left(1-\gamma\right)}\left\|V\right\|_{\infty}.$$

F Analysis for the minimax lower bound in Theorem 3

Without loss of generality, we assume throughout the proof that S is an odd number.

Construction of hard MDP instances. For any vector $\theta = [\theta_s]_{1\leq s\leq S} \in \mathbb{R}^S$ with $\|\theta\|_{\infty} \leq 1/2$, define an MDP $\mathcal{M}_\theta = \{\mathcal{S}, \mathcal{A}, \gamma, P_\theta, r\}$ parameterized by θ with the state space $\mathcal{S} = \left\{0, \pm 1, \ldots, \pm\frac{S-1}{2}\right\}$ and the action space $\mathcal{A} = \{\pm 1\}$. Several key elements are constructed as follows.

- We first define the initial state distribution as

$$\rho(s) = \left(1 - \frac{1}{\tau_{\mathsf{mix}}}\right)\mathbb{1}(s=0) + \frac{1}{(S-1)\tau_{\mathsf{mix}}}\mathbb{1}(s\neq 0), \qquad \forall\, s\in\mathcal{S}, \tag{F.1}$$

  where τ_{mix} is defined to be

$$\tau_{\mathsf{mix}} = \frac{16c_2}{(1-\gamma)\,\varepsilon}. \tag{F.2}$$

  It is self-evident that as long as $c_1 \leq 32c_2$, we have $\tau_{\mathsf{mix}} \geq 2$ and hence ρ is well-defined.

- The transition probability kernel is constructed such that

$$P_\theta\left(s' \,|\, s, a\right) = \left(1 - \frac{1}{t_{\mathsf{mix}}}\right)\left[\left(\frac{1}{2} + \theta_s a\right)\mathbb{1}\left\{s' = s\right\} + \left(\frac{1}{2} - \theta_s a\right)\mathbb{1}\left\{s' = -s\right\}\right] + \frac{1}{t_{\mathsf{mix}}}\rho\left(s'\right) \tag{F.3}$$

  for any $(s, a, s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$.

- The reward function is taken to be

$$r\left(s,a\right) = \begin{cases} 1, & \text{if } s > 0; \\ 0, & \text{if } s \leq 0. \end{cases} \tag{F.4}$$

- The behavior policy is chosen to be

$$\pi_{\mathsf{b}}(a\,|\,s) = \frac{1}{2}, \qquad \forall\,(s,a)\in\mathcal{S}\times\mathcal{A}. \tag{F.5}$$

- Furthermore, we shall define

$$\Theta \coloneqq \left\{\theta\in\mathbb{R}^S : \theta_s \in \left\{\pm\frac{1}{2}\right\} \text{ for } s>0, \text{ and } \theta_s = 0 \text{ for } s\leq 0\right\}, \tag{F.6}$$

  and we consider a class of MDP given by $\{\mathcal{M}_\theta : \theta\in\Theta\}$. In what follows, we shall also let $\mathcal{U}$ represent the uniform distribution over Θ.

Useful properties about the constructed MDP. Next, for any $\theta \in \Theta$, we gather a couple of basic properties about $\mathcal{M}_\theta$ and the Markov chain induced by the behavior policy π_{b}.

- The stationary distribution μ_{b} of the Markov chain induced by π_{b} is given by

$$\mu_{\mathsf{b}}(s,a) = \rho(s)\,\pi_{\mathsf{b}}(a \,|\, s) = \frac{1}{2}\rho(s)\,. \tag{F.7}$$

  for all (s,a). The mixing time of this Markov chain is at most $2t_{\mathsf{mix}}$, since it is clear that

$$d_{\mathsf{TV}}\left(P_\theta^t(\cdot \,|\, s_0, a_0), \mu_{\mathsf{b}}\right) \le \left(1 - \frac{1}{t_{\mathsf{mix}}}\right)^t \le \frac{1}{4}, \qquad \forall (s_0, a_0) \in \mathcal{S} \times \mathcal{A}$$

  as long as $t \ge 2t_{\mathsf{mix}}$.

- For any policy π, it can be verified that

$$d_\rho^\pi(s) \le \frac{3}{2}\rho(s) \tag{F.8}$$

  for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. The proof of (F.8) is deferred to Appendix F.1. This immediately gives

$$C^\star = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_\rho^\star(s,a)}{\mu_{\mathsf{b}}(s,a)} \overset{\text{(i)}}{\le} 2 \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_\rho^\star(s,a)}{\rho(s)} \overset{\text{(ii)}}{\le} 3,$$

  where (i) follows from (F.7), and (ii) utilizes (F.8).

- The optimal policy $\pi_\theta^\star$ is given by

$$\pi_\theta^\star(s) = \mathsf{sign}(\theta_s), \qquad \forall\, s > 0, \tag{F.9}$$

  and $\pi_\theta^\star(s)$ can be either 1 or -1 for $s \le 0$. In addition, for any $s > 0$ and any policy π that is independent of θ_s, we can show that

$$\mathop{\mathbb{E}}_{\mathcal{M}_\theta:\theta\sim\mathcal{U}}\left[V_\theta^\star(s) - V^\pi(s)\right] \ge \frac{1}{35(1-\gamma)}. \tag{F.10}$$

  The proof of (F.9) and (F.10) can be found in Appendix F.2.

Proof of the claimed lower bound. Let $X_1, \ldots, X_T \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(1/t_{\mathsf{mix}})$ and $Y_1, \ldots, Y_T \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(1/\tau_{\mathsf{mix}})$ be two independent sequences of Bernoulli random variables. Let S_T be the number of different state pairs $\{\pm s\}$ visited by the sample trajectory. From the construction of the Markov chain, we know that S_T is stochastically dominated by $1 + \sum_{t=1}^T X_t Y_t$. In view of this, we have

$$\mathbb{E}[S_T] \le \mathbb{E}\left[1 + \sum_{t=1}^T X_t Y_t\right] = 1 + \frac{T}{t_{\mathsf{mix}}\tau_{\mathsf{mix}}} \le 1 + c\frac{S}{(1-\gamma)\tau_{\mathsf{mix}}\varepsilon} \le \frac{S}{8},$$

where the last relation holds since $\tau_{\mathsf{mix}} = \frac{16c_2}{(1-\gamma)\varepsilon}$ and $S \ge 16$. This implies that with probability at least 0.5, $S_T \le S/4$. Then for any estimator π, with probability at least 0.5 we have

$$\begin{aligned}\max_{\mathcal{M}_\theta:\theta\in\Theta} V^\star(\rho) - V^\pi(\rho) &\ge \mathop{\mathbb{E}}_{\mathcal{M}_\theta:\theta\sim\mathcal{U}}\left[V^\star(\rho) - V^\pi(\rho)\right] = \sum_{s\in\mathcal{S}} \rho(s) \mathop{\mathbb{E}}_{\mathcal{M}_\theta:\theta\sim\mathcal{U}}\left[V^\star(s) - V^\pi(s)\right] \\ &\overset{\text{(i)}}{\ge} \left(\frac{S-1}{2} - S_T\right)\frac{1}{(S-1)\tau_{\mathsf{mix}}}\frac{\gamma}{35(1-\gamma)} \overset{\text{(ii)}}{\ge} \frac{\varepsilon}{4800c_2}.\end{aligned}$$

Here, (i) holds since there are at least $(S-1)/2 - S_T$ states s satisfying $s > 0$ and the state pair $\{\pm s\}$ is not visited by the sample trajectory. For any such state s, any estimator based on the observed sample trajectory is independent of θ_s. Then (i) follows from (F.10). In addition, (ii) holds as long as $S_T \le S/4$, $S \ge 16$ and $\gamma \ge 1/2$. By replacing ε with $4800c_2\varepsilon$, we arrive at the desired lower bound.

F.1 Proof of Equation (F.8)

Let $P_\theta^\pi : \mathcal{S} \to \mathcal{S}$ be the probability transition kernel defined as

$$P_\theta^\pi(s' \,|\, s) = \sum_{a \in \mathcal{A}} P_\theta(s' \,|\, s, a)\pi(a \,|\, s), \qquad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}.$$

For any policy π and any $\nu \in \Delta(\mathcal{S})$ satisfying $\nu(s) + \nu(-s) = \rho(s) + \rho(-s) = 2\rho(s)$ for all $s \in \mathcal{S}$, we have

$$\begin{aligned}
(\nu P_\theta^\pi)(s') &= \sum_s \nu(s)\left[\pi(1 \,|\, s) P_\theta(s' \,|\, s, 1) + \pi(-1 \,|\, s) P_\theta(s' \,|\, s, -1)\right] \\
&= \sum_s \nu(s)\Bigg[\pi(1 \,|\, s)\left(1 - \frac{1}{t_{\mathsf{mix}}}\right)\left[\left(\frac{1}{2} + \theta_s\right)\mathbb{1}\{s' = s\} + \left(\frac{1}{2} - \theta_s\right)\mathbb{1}\{s' = -s\}\right] \\
&\qquad + \pi(-1 \,|\, s)\left(1 - \frac{1}{t_{\mathsf{mix}}}\right)\left[\left(\frac{1}{2} - \theta_s\right)\mathbb{1}\{s' = s\} + \left(\frac{1}{2} + \theta_s\right)\mathbb{1}\{s' = -s\}\right] + \frac{1}{t_{\mathsf{mix}}}\rho(s')\Bigg] \\
&= \left(1 - \frac{1}{t_{\mathsf{mix}}}\right)\left[\frac{1 + \theta_{s'}(2\pi(1 \,|\, s') - 1)}{2}\nu(s') + \frac{1 - \theta_{-s'}(2\pi(1 \,|\, -s') - 1)}{2}\nu(-s')\right] + \frac{1}{t_{\mathsf{mix}}}\rho(s') \\
&= \rho(s') + \left(1 - \frac{1}{t_{\mathsf{mix}}}\right)\left[\frac{\theta_{s'}(2\pi(1 \,|\, s') - 1)}{2}\nu(s') - \frac{\theta_{-s'}(2\pi(1 \,|\, -s') - 1)}{2}\nu(-s')\right]
\end{aligned}$$

for any $s' \in \mathcal{S}$. Here the last relation follows from $\nu(s) + \nu(-s) = 2\rho(s)$. Therefore for any $s \in \mathcal{S}$, we have

$$|(\nu P_\theta^\pi)(s) - \rho(s)| \le \frac{1}{2}\|\theta\|_\infty \nu(s) + \frac{1}{2}\|\theta\|_\infty \nu(-s) = \|\theta\|_\infty \rho(s).$$

We can then employ standard induction arguments to verify that for any $n \ge 0$,

$$[\rho(P_\theta^\pi)^n](s) + [\rho(P_\theta^\pi)^n](-s) = 2\rho(s).$$

This immediately gives

$$\begin{aligned}
d_\rho^\pi(s) &= (1-\gamma)\sum_{j=0}^\infty \gamma^j\left[\rho(P_\theta^\pi)^j\right](s) = \rho(s) + (1-\gamma)\sum_{j=0}^\infty \gamma^j\left[\rho(P_\theta^\pi)^j - \rho\right](s) \\
&\le \rho(s) + (1-\gamma)\sum_{j=0}^\infty \gamma^j \|\theta\|_\infty \rho(s) = (1 + \|\theta\|_\infty)\rho(s) \le \frac{3}{2}\rho(s).
\end{aligned}$$

F.2 Proof of Equations (F.9) and (F.10)

For any policy π, the Bellman equations assert that

$$\begin{aligned}
Q^\pi(s,a) &= r(s,a) + \gamma\left\{\left(1 - \frac{1}{t_{\mathsf{mix}}}\right)\left[\left(\frac{1}{2} + \theta_s a\right)V^\pi(s) + \left(\frac{1}{2} - \theta_s a\right)V^\pi(-s)\right] + \frac{1}{t_{\mathsf{mix}}}V^\pi(\rho)\right\} \\
V^\pi(s) &= \pi(1 \,|\, s)Q^\pi(s,1) + \pi(-1 \,|\, s)Q^\pi(s,-1)
\end{aligned}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. For any given $s > 0$, denote by $x = V^\pi(s)$ and $y = V^\pi(-s)$. Since $\theta_{-s} = 0$, we have $Q^\pi(-s,1) = Q^\pi(-s,-1) = V^\pi(-s) = y$. Then the above Bellman equations give

$$\begin{aligned}
x &= \pi(1 \,|\, s)\left[1 + \gamma\left\{\left(1 - \frac{1}{t_{\mathsf{mix}}}\right)\left[\left(\frac{1}{2} + \theta_s\right)x + \left(\frac{1}{2} - \theta_s\right)y\right] + \frac{1}{t_{\mathsf{mix}}}V^\pi(\rho)\right\}\right] \\
&\quad + \pi(-1 \,|\, s)\left[1 + \gamma\left\{\left(1 - \frac{1}{t_{\mathsf{mix}}}\right)\left[\left(\frac{1}{2} - \theta_s\right)x + \left(\frac{1}{2} + \theta_s\right)y\right] + \frac{1}{t_{\mathsf{mix}}}V^\pi(\rho)\right\}\right] \\
&= \gamma\left(1 - \frac{1}{t_{\mathsf{mix}}}\right)\left\{\left[\frac{1}{2} + (2\pi(1 \,|\, s) - 1)\theta_s\right]x + \left[\frac{1}{2} - (2\pi(1 \,|\, s) - 1)\theta_s\right]y\right\} + 1 + \frac{\gamma}{t_{\mathsf{mix}}}V^\pi(\rho) \qquad \text{(F.11a)}
\end{aligned}$$

and

$$y=\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)\left(\frac{1}{2}x+\frac{1}{2}y\right)+\frac{\gamma}{t_{\mathsf{mix}}}V^{\pi}\left(\rho\right). \tag{F.11b}$$

Solving the system of equations (F.11) with respect to x and y gives

$$V^{\pi}\left(s\right)=\left[\frac{2-\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}{2-2\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}\right]\frac{1}{1-\theta_{s}\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)\left(2\pi\left(1\,|\,s\right)-1\right)}+\frac{1}{1-\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}\frac{\gamma}{t_{\mathsf{mix}}}V^{\pi}\left(\rho\right)$$

and

$$V^{\pi}\left(-s\right)=\frac{\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}{2-2\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}\frac{1}{1-\theta_{s}\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)\left(2\pi\left(1\,|\,s\right)-1\right)}+\frac{1}{1-\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}\frac{\gamma}{t_{\mathsf{mix}}}V^{\pi}\left(\rho\right).$$

Similarly, we can also derive

$$V^{\pi}\left(0\right)=\frac{1}{1-\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}\frac{\gamma}{t_{\mathsf{mix}}}V^{\pi}\left(\rho\right).$$

Then it boils down to determining $V^{\pi}(\rho)$, for which we have

$$\begin{aligned}V^{\pi}\left(\rho\right)&=\mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\pi}\left(s\right)\right]=\left(1-\frac{1}{\tau_{\mathsf{mix}}}\right)V^{\pi}\left(0\right)+\sum_{s\neq0}\frac{1}{(S-1)\tau_{\mathsf{mix}}}V^{\pi}\left(s\right)\\&=\frac{1-\frac{1}{\tau_{\mathsf{mix}}}}{1-\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}\frac{\gamma}{t_{\mathsf{mix}}}V^{\pi}\left(\rho\right)+\sum_{s>0}\frac{1}{(S-1)\tau_{\mathsf{mix}}}\frac{1}{1-\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}\frac{1}{1-\theta_{s}\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)\left(2\pi\left(1\,|\,s\right)-1\right)}.\end{aligned}$$

This further leads to

$$V^{\pi}\left(\rho\right)=\frac{1}{1-\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)-\left(1-\frac{1}{\tau_{\mathsf{mix}}}\right)\frac{\gamma}{t_{\mathsf{mix}}}}\cdot\frac{1}{(S-1)\tau_{\mathsf{mix}}}\cdot\sum_{s>0}\frac{1}{1-\theta_{s}\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)\left(2\pi\left(1\,|\,s\right)-1\right)}. \tag{F.13}$$

By taking together (F.12) and (F.13), one can check that the optimal policy $\pi_{\theta}^{\star}$ is given by

$$\pi_{\theta}^{\star}\left(s\right)=\mathsf{sign}(\theta_{s})$$

for any $s>0$. The choice of $\pi_{\theta}^{\star}(s)$ for $s\leq0$ does not affect the value function and can be chosen aribtrarily.

Let

$$C_{0}\coloneqq\frac{2-\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}{2-2\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}=\frac{1}{2}+\frac{1}{2}\frac{1}{1-\gamma+\frac{\gamma}{t_{\mathsf{mix}}}}\geq\frac{5}{11\left(1-\gamma\right)}, \tag{F.14}$$

where the last step holds since $t_{\mathsf{mix}}\geq10/(1-\gamma)$. Then (F.12) implies that

$$V^{\pi}\left(s\right)=\frac{C_{0}}{1-\theta_{s}\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)\left(2\pi\left(1\,|\,s\right)-1\right)}+R^{\pi}\qquad\text{where}\qquad0\leq R^{\pi}\leq\frac{1}{\left(1-\gamma\right)^{2}t_{\mathsf{mix}}}.$$

It is straightforward to compute

$$\mathop{\mathbb{E}}_{\mathcal{M}_{\theta}:\theta\sim\mathcal{U}}\left[V_{\theta}^{\star}\left(s\right)\right]=\frac{C_{0}}{1-\frac{1}{2}\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)}+\mathop{\mathbb{E}}_{\mathcal{M}_{\theta}:\theta\sim\mathcal{U}}\left[R^{\pi_{\theta}^{\star}}\left(s\right)\right].$$

For any policy π that is independent of θ_s, we can also compute

$$\mathop{\mathbb{E}}_{\mathcal{M}_{\theta}:\theta\sim\mathcal{U}}\left[V^{\pi}\left(s\right)\right]=\frac{1}{2}\frac{C_{0}}{1-\frac{1}{2}\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)\left(2\pi\left(1\,|\,s\right)-1\right)}+\frac{1}{2}\frac{C_{0}}{1+\frac{1}{2}\gamma\left(1-\frac{1}{t_{\mathsf{mix}}}\right)\left(2\pi\left(1\,|\,s\right)-1\right)}+\mathop{\mathbb{E}}_{\mathcal{M}_{\theta}:\theta\sim\mathcal{U}}\left[R^{\pi}\left(s\right)\right]$$

$$
\begin{aligned}
&= \frac{C_0}{1 - \frac{1}{4}\gamma^2 \left(1 - \frac{1}{t_{\mathsf{mix}}}\right)^2 \left(2\pi\left(1 \mid s\right) - 1\right)^2} + \mathop{\mathbb{E}}_{\mathcal{M}_\theta : \theta \sim \mathcal{U}} \left[R^\pi\left(s\right)\right] \\
&\le \frac{C_0}{1 - \frac{1}{4}\gamma^2 \left(1 - \frac{1}{t_{\mathsf{mix}}}\right)^2} + \mathop{\mathbb{E}}_{\mathcal{M}_\theta : \theta \sim \mathcal{U}} \left[R^\pi\left(s\right)\right].
\end{aligned}
$$

Therefore, we can conclude that

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{M}_\theta : \theta \sim \mathcal{U}} \left[V^\star_\theta\left(s\right) - V^\pi\left(s\right)\right] &\ge C_0 \frac{\frac{1}{2}\gamma\left(1 - \frac{1}{t_{\mathsf{mix}}}\right)}{1 - \frac{1}{2}\gamma\left(1 - \frac{1}{t_{\mathsf{mix}}}\right)} + \mathop{\mathbb{E}}_{\mathcal{M}_\theta : \theta \sim \mathcal{U}} \left[R^{\pi^\star_\theta}\left(s\right)\right] - \mathop{\mathbb{E}}_{\mathcal{M}_\theta : \theta \sim \mathcal{U}} \left[R^\pi\left(s\right)\right] \\
&\ge \frac{45}{341\left(1-\gamma\right)} - \frac{1}{\left(1-\gamma\right)^2 t_{\mathsf{mix}}} \ge \frac{1}{35\left(1-\gamma\right)},
\end{aligned}
$$

where the last two relations follow from (F.14) and the assumptions that $\gamma \le 1/2$ and $t_{\mathsf{mix}} \ge 10/(1-\gamma)$.

References

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

Agarwal, A., Kakade, S., and Yang, L. F. (2020b). 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.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient q-learning with low switching cost. *Advances in Neural Information Processing Systems*, 32.

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

Bertsekas, D. P. and Tsitsiklis, J. N. (2003). Parallel and distributed computation: numerical methods.

Bowen, W., Huaqing, X., Lin, Z., Yingbin, L., and Wei, Z. (2021). Finite-time theory for momentum Q-learning. In *Uncertainty in Artificial Intelligence*, pages 665–674. PMLR.

Chang, J., Uehara, M., Sreenivas, D., Kidambi, R., and Sun, W. (2021). Mitigating covariate shift in imitation learning via offline data with partial coverage. *Advances in Neural Information Processing Systems*, 34:965–979.

Chen, Z., Clarke, J. P., and Maguluri, S. T. (2022). Target network and truncation overcome the deadly triad in q-learning. *arXiv preprint arXiv:2203.02628*.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2021). A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567*.

Dong, K., Wang, Y., Chen, X., and Wang, L. (2019). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. (2017). Stochastic variance reduction methods for policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1049–1058. JMLR. org.

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

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., and Jordan, M. I. (2021a). Is temporal difference learning optimal? an instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1013–1040.

Khamaru, K., Xia, E., Wainwright, M. J., and Jordan, M. I. (2021b). Instance-optimality in optimal value estimation: Adaptivity via variance-reduced q-learning. *arXiv preprint arXiv:2106.14352*.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc.

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2023). Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022). Settling the sample complexity of model-based offline reinforcement learning. *in preparation*.

Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. (2021a). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 33.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021b). Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2020). Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274.

Ménard, P., Domingues, O. D., Shang, X., and Valko, M. (2021). Ucb momentum q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pages 7609–7618. PMLR.

Munos, R. (2003). Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567.

Munos, R. (2007). Performance bounds in ℓ_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561.

Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20.

Pazis, J., Parr, R. E., and How, J. P. (2016). Improving pac exploration using the median of means. *Advances in Neural Information Processing Systems*, 29.

Qu, G. and Wierman, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. *Conference on Learning Theory*, pages 3185–3205.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Neural Information Processing Systems (NeurIPS)*.

Shah, D. and Xie, Q. (2018). Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 3111–3121.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pages 19967–20025. PMLR.

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. In *Advances in Neural Information Processing Systems*, pages 5186–5196.

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.

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. (2006). PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888.

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

Tsybakov, A. B. and Zaiats, V. (2009). *Introduction to nonparametric estimation*, volume 11. Springer.

Uehara, M. and Sun, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*.

Wai, H.-T., Hong, M., Yang, Z., Wang, Z., and Tang, K. (2019). Variance reduced policy evaluation with smooth function approximation. *Advances in Neural Information Processing Systems*, 32:5784–5795.

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*.

Wang, B., Yan, Y., and Fan, J. (2021). Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. *Advances in Neural Information Processing Systems*, 34.

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

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407.

Xiong, H., Zhao, L., Liang, Y., and Zhang, W. (2020). Finite-time analysis for double Q-learning. *Advances in Neural Information Processing Systems*, 33.

Xu, T., Wang, Z., Zhou, Y., and Liang, Y. (2019). Reanalysis of variance reduced temporal difference learning. In *International Conference on Learning Representations*.

Yang, K., Yang, L., and Du, S. (2021). Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. 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.

Yin, M., Bai, Y., and Wang, Y.-X. (2021a). Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34:7677–7688.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. (2021b). Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *International Conference on Learning Representations*.

Yin, M. and Wang, Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34.

Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., and Finn, C. (2021a). Conservative data sharing for multi-task offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:11501–11516.

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. (2021b). Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. (2020). Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. (2022). Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5757–5773. PMLR.

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33.

Zhang, Z., Zhou, Y., and Ji, X. (2021). Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, pages 12653–12662. PMLR.