Settling the Sample Complexity of Model-Based Offline Reinforcement Learning

Gen Li* UPenn | Laixi Shi† CMU | Yuxin Chen*‡ UPenn | Yuejie Chi† CMU | Yuting Wei* UPenn

April 2022; Revised: February 2024

Abstract

This paper is concerned with offline reinforcement learning (RL), which learns using pre-collected data without further exploration. Effective offline RL would be able to accommodate distribution shift and limited data coverage. However, prior algorithms or analyses either suffer from suboptimal sample complexities or incur high burn-in cost to reach sample optimality, thus posing an impediment to efficient offline RL in sample-starved applications.

We demonstrate that the model-based (or "plug-in") approach achieves minimax-optimal sample complexity without burn-in cost for tabular Markov decision processes (MDPs). Concretely, consider a γ-discounted infinite-horizon (resp. finite-horizon) MDP with S states and effective horizon $\frac{1}{1-\gamma}$ (resp. horizon H), and suppose the distribution shift of data is reflected by some single-policy clipped concentrability coefficient $C^{\star}_{\mathsf{clipped}}$. We prove that model-based offline RL yields ε-accuracy with a sample complexity of

$$\begin{cases} \frac{SC^{\star}_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2} & \text{(infinite-horizon MDPs)} \\ \frac{H^4SC^{\star}_{\mathsf{clipped}}}{\varepsilon^2} & \text{(finite-horizon MDPs)} \end{cases}$$

up to log factor, which is minimax optimal for the *entire* ε*-range*. The proposed algorithms are "pessimistic" variants of value iteration with Bernstein-style penalties, and do not require sophisticated variance reduction. Our analysis framework is established upon delicate leave-one-out decoupling arguments in conjunction with careful self-bounding techniques tailored to MDPs.

Keywords: offline reinforcement learning, model-based approach, minimax lower bounds, distribution shift, pessimism in the face of uncertainty

Contents

*Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.
†Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
‡Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA.

1 Introduction

Reinforcement learning (RL) has recently achieved superhuman performance in the gaming frontier, such as the game of Go (Silver et al., 2017), under the premise that vast amounts of training data can be obtained. However, limited capability of online data collection in other real-world applications — e.g., clinical trials and online advertising, where real-time data acquisition is expensive, high-stakes, and/or time-consuming, — presents a fundamental bottleneck for carrying such RL success over to broader scenarios. To circumvent this bottleneck, one plausible strategy is to make more effective use of data collected previously, given that such historical data might contain useful information that readily transfers to new tasks (for instance, the state transitions in a historical task might sometimes resemble what happen in new tasks). The potential of this data-driven approach has been explored and recognized in a diverse array of contexts including but not limited to robotic manipulation (Ebert et al., 2018), autonomous driving (Diehl et al., 2021), and healthcare (Tang and Wiens, 2021); see Levine et al. (2020); Prudencio et al. (2022) for overviews of recent development. Nowadays, the subfield of reinforcement learning using historical data, without further exploration of the environment, is commonly referred to as *offline RL* or *batch RL* (Lange et al., 2012; Levine et al., 2020). A desired offline RL algorithm would achieve the target statistical accuracy using as few samples as possible.

1.1 Challenges: distribution shift and limited data coverage

In contrast to online exploratory RL, offline RL has to deal with several critical issues resulting from the absence of active exploration. Below we single out two representative issues surrounding offline RL.

- *Distribution shift.* For the most part, the historical data is generated by a certain behavior policy that departs from the optimal one. A key challenge in offline RL thus stems from the shift of data distributions: how to leverage past data to the most effect, even though the distribution induced by the target policy differs from what we have available?
- *Limited data coverage.* Ideally, if the dataset contained sufficiently many data samples for every state-action pair, then there would be hope to simultaneously learn the performance of every policy. Such a uniform coverage requirement, however, is oftentimes not only unrealistic (given that we can no longer change the past data) but also unnecessary (given that we might only be interested in identifying a single optimal policy).

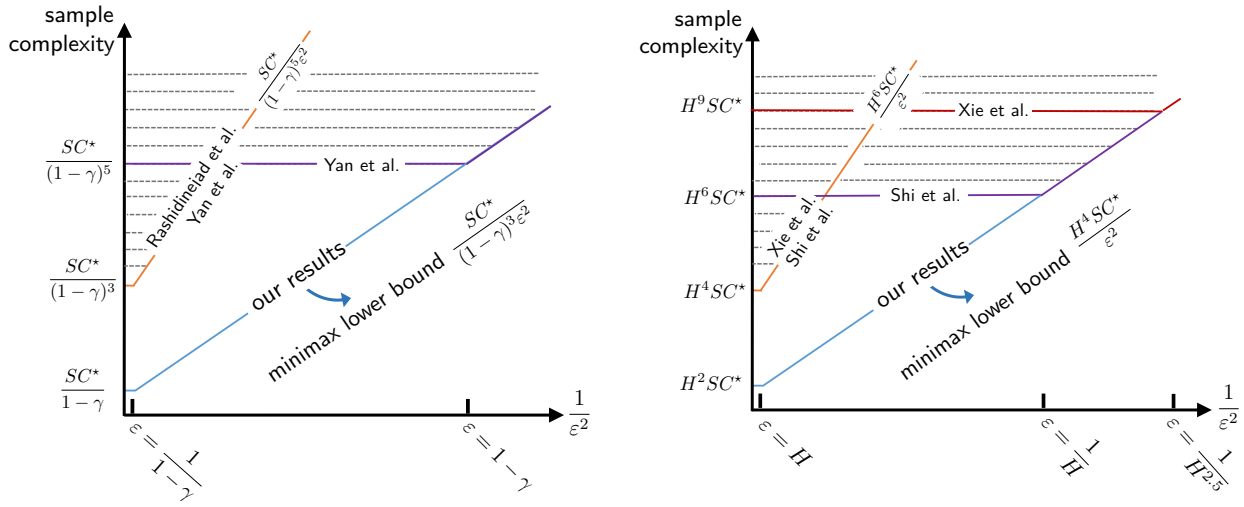
Whether one can effectively cope with distribution shift and insufficient data coverage becomes a major factor that governs the feasibility and statistical efficiency of offline RL.

In order to address the aforementioned issues, a recent strand of works put forward the *principle of pessimism or conservatism* (e.g., Buckman et al. (2020); Chen et al. (2021a); Cui and Du (2022); Jin et al. (2021); Kumar et al. (2020); Liu et al. (2020); Rashidinejad et al. (2022); Shi et al. (2022b); Uehara and Sun (2021); Xie et al. (2021); Yan et al. (2023); Yin and Wang (2021); Zanette et al. (2021); Zhong et al. (2022)). This is reminiscent of the optimism principle in the face of uncertainty for online exploration (Azar et al., 2017; Bourel et al., 2020; Jaksch et al., 2010; Jin et al., 2018; Lai and Robbins, 1985), but works for drastically different reasons (as we shall elucidate momentarily). One plausible idea of the pessimism principle, which has been incorporated into both model-based and model-free approaches, is to penalize value estimation of those state-action pairs that have been poorly covered. Informally speaking, insufficient coverage of a state-action pair inevitably results in low confidence and high uncertainty in the associated value estimation, and it is hence advisable to act cautiously by tuning down the corresponding value estimate. Proper use of pessimism amid uncertainty brings about several provable benefits (Rashidinejad et al., 2022; Xie et al., 2021): (i) it allows for a reduced sample size that adapts to the degree of distribution shift; (ii) as opposed to uniform data coverage, it only requires coverage of the part of the state-action space reachable by the target policy. Details to follow momentarily.

1.2 Inadequacy of prior works

In the present paper, we evaluate and compare the statistical performance of offline RL algorithms mainly through the lens of sample complexity — namely, the number of samples needed for an algorithm to output, with probability approaching one, a policy whose resultant value function is at most ε away from optimal (called "ε-accuracy" throughout). An ultimate goal is to design an algorithm to achieve the smallest possible sample complexity.

Despite extensive recent activities, however, existing statistical guarantees for the above paradigm remain inadequate, as we shall elaborate on below. For concreteness, our discussions focus on two widely studied Markov decision processes (MDPs) with S states and A actions (Bertsekas, 2017): (a) γ-discounted infinite-horizon MDPs, with effective horizon $\frac{1}{1-\gamma}$; (b) finite-horizon MDPs with horizon length H and nonstationary transition kernels. We shall bear in mind that all of these salient problem parameters (i.e., S, A, $\frac{1}{1-\gamma}$, H) could be enormous in modern RL applications. In addition, previous works have isolated an important parameter $C^\star \geq 1$ — called the single-policy concentrability coefficient (Rashidinejad et al., 2022; Xie et al., 2021) — that measures the mismatch of distributions induced by the target policy against the behavior policy; see Sections 3.1 and 2.1 for precise definitions. Naturally, the statistical performance of desirable algorithms would degrade gracefully as the distribution mismatch worsens (i.e., as $C^\star$ increases). In the sequel, we shall discuss two dinstinctive RL paradigms — model-based RL and model-free RL — separately. Throughout this paper, the standard notation $\widetilde{O}(\cdot)$ indicates the order of a function with all log terms in $S, A, \frac{1}{1-\gamma}, H, \frac{1}{\varepsilon}$, and $\frac{1}{\delta}$ (with $1-\delta$ the target success probability) hidden.



(a) discounted infinite-horizon MDPs (b) finite-horizon MDPs

Figure 1: An illustration of prior works, where (a) is about discounted infinite-horizon MDPs and (b) is about finite-horizon MDPs. To facilitate comparisons, we replace $C^\star_{\text{clipped}}$ with $C^\star$ in our results when drawing the plots given that $C^\star_{\text{clipped}} \le C^\star$. The shaded regions indicate the state-of-the-art achievability results. Our work manages to close the gaps between the prior achievable regions and the minimax lower bounds.

Model-based offline RL. Model-based algorithms — which can be interpreted as a "plug-in" statistical approach — start by computing an empirical model for the unknown MDP, and output a policy that is (near)-optimal in accordance with the empirical MDP. When coupled with the pessimism principle, the model-based approach has been shown to enjoy the following sample complexity bounds.

- By incorporating Hoeffding-style lower confidence bounds into value iteration, Rashidinejad et al. (2022); Xie et al. (2021) demonstrated that a sample complexity of

$$\begin{cases} \widetilde{O}\big(\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}\big) & \text{for infinite-horizon MDPs} \\ \widetilde{O}\big(\frac{H^6SC^\star}{\varepsilon^2}\big) & \text{for finite-horizon MDPs} \end{cases} \tag{1}$$

  suffices to yield ε-accuracy. Such a sample complexity bound, however, is a large factor of $\frac{1}{(1-\gamma)^2}$ (resp. H^2) above the minimax lower limit derived for infinite-horizon MDPs (resp. finite-horizon MDPs) (Rashidinejad et al., 2022; Xie et al., 2021; Yin and Wang, 2021).

- In an attempt to optimize the sample complexity, Xie et al. (2021) leveraged the idea of variance reduction — a powerful strategy originating from the stochastic optimization literature (Johnson and Zhang, 2013) — in model-based RL and obtained a strengthened sample complexity of

$$\widetilde{O}\Big(\frac{H^4SC^\star}{\varepsilon^2} + \frac{H^{6.5}SC^\star}{\varepsilon}\Big) \tag{2}$$

  for finite-horizon MDPs. This sample complexity bound approaches the minimax lower limit (i.e., the order of $\frac{H^4SC^\star}{\varepsilon^2}$) once the sample size exceeds the order of

$$\text{(burn-in cost)} \qquad H^9SC^\star; \tag{3}$$

  in other words, an enormous burn-in sample size is needed in order to attain sample optimality.

Model-free offline RL. The model-free approach forms a contrastingly different class of RL algorithms, which bypasses the model estimation stage and directly learns the optimal values. Noteworthily, Q-learning

and its variants (Watkins and Dayan, 1992), which apply stochastic approximation (Robbins and Monro, 1951) based on the Bellman optimality condition, are among the most widely used model-free paradigms. The principle of pessimism amid uncertainty has recently been integrated into model-free algorithms as well, with the state-of-the-art statistical guarantees listed below (Shi et al., 2022b; Yan et al., 2023).

- When Q-learning is implemented in conjunction with Hoeffding-style lower confidence bounds, it has been shown to achieve the same sample complexity as (1), which is suboptimal by a factor of either $\frac{1}{(1-\gamma)^2}$ or H^2.
- A variance-reduced variant of pessimistic Q-learning allows for further sample size benefits, achieving a sample complexity of

$$\begin{cases} \widetilde{O}\big(\frac{SC^\star}{(1-\gamma)^3\varepsilon^2} + \frac{SC^\star}{(1-\gamma)^4\varepsilon}\big) & \text{for infinite-horizon MDPs} \\ \widetilde{O}\big(\frac{H^4SC^\star}{\varepsilon^2} + \frac{H^5SC^\star}{\varepsilon}\big) & \text{for finite-horizon MDPs} \end{cases} \tag{4}$$

  for any target accuracy level ε. This means that the algorithm is guaranteed to be sample-optimal only after the total sample size exceeds the order of

$$\text{(burn-in cost)} \qquad \begin{cases} \frac{SC^\star}{(1-\gamma)^5} & \text{for infinite-horizon MDPs,} \\ H^6SC^\star & \text{for finite-horizon MDPs,} \end{cases} \tag{5}$$

  which again manifests itself as a significant burn-in cost for long-horizon problems.

Summary. As elucidated above, existing algorithms either suffer from suboptimal sample complexities, or require sophisticated techniques like variance reduction to approach minimax optimality. Even when variance reduction is employed, prior algorithms incur an enormous burn-in cost in order to work optimally, thus posing an impediment to achieving sample efficiency in data-starved applications. Table 1 summarizes quantitatively the previous results, whereas Figure 1 illustrates the gaps between the state-of-the-art upper bounds and the minimax lower bounds (as derived by Rashidinejad et al. (2022); Xie et al. (2021)). All this motivates the studies of the following natural questions:

> *Can we develop an offline RL algorithm that achieves near-optimal sample complexity without burn-in cost? If so, can we accomplish this goal by means of a simple algorithm without resorting to sophisticated schemes like variance reduction?*

The current paper answers these questions affirmatively by studying the model-based approach.

1.3 Main contributions

In this paper, we settle the sample complexity of model-based offline RL by studying a pessimistic variant of value iteration — called VI-LCB — applied to some empirical MDP. Encouragingly, for both discounted infinite-horizon and finite-horizon MDPs, the model-based algorithms provably achieve minimax-optimal sample complexities for any given target accuracy level ε — namely, any $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$ for discounted infinite-horizon MDPs and $\varepsilon \in (0, H]$ for finite-horizon MDPs.

To be more precise, we introduce a slightly modified version $C^\star_{\mathsf{clipped}}$ of the concentrability coefficient $C^\star$, which always satisfies $C^\star_{\mathsf{clipped}} \le C^\star$ and shall be termed the single-policy clipped concentrability coefficient (see Sections 2.1 and 3.1 for more details as well as the advantages of this coefficient). The introduction of this new parameter leads to slightly improved sample complexity compared to the one based on $C^\star$. The main contributions are summarized as follows.

- For γ-discounted infinite-horizon MDPs, we demonstrate that with high probability, the VI-LCB algorithm with Bernstein-style penalty finds an ε-optimal policy with a sample complexity of

$$\widetilde{O}\bigg(\frac{SC^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\bigg) \tag{6}$$

for any given accuracy level $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$ (see Theorem 1). Our algorithm reuses all samples across all iterations in order to achieve data efficiency, and our analysis builds upon a novel leave-one-out argument to decouple complicated statistical dependency across iterations. Moreover, the above sample complexity (6) remains valid if $C^\star_{\mathsf{clipped}}$ is replaced by $C^\star$.

- For finite-horizon MDPs with nonstationary transition kernels, we propose a variant of VI-LCB that adopts the Bernstein-style penalty to enforce pessimism in the face of uncertainty. We prove that for any given $\varepsilon \in (0, H]$, the proposed algorithm yields an ε-optimal policy using

$$\widetilde{O}\Big(\frac{H^4 S C^\star_{\mathsf{clipped}}}{\varepsilon^2}\Big) \tag{7}$$

samples with high probability (see Theorem 3). A key ingredient in the algorithm design is a two-fold subsampling trick that helps decouple the statistical dependency along the sample rollouts. Note that the above sample complexity result (7) continues to hold if one replaces $C^\star_{\mathsf{clipped}}$ with $C^\star$.

- To assess the tightness and optimality of our results, we further develop minimax lower bounds in Theorems 2 and 4, which match the above upper bounds modulo some logarithmic factors.

Remarkably, our algorithms do not require sophisticated variance reduction schemes, as long as suitable confidence bounds are adopted. Detailed theoretical comparisons with prior art can be found in Table 1 and are also illustrated in Figure 1. Finally, we have conducted a series of numerical experiments to evaluate the performance of the proposed algorithms in Section 4.

Statistical contributions: solving the most sample-hungry regime. The offline RL problem considered herein is statistical in nature, in that it seeks to learn from pre-collected data in the face of uncertainty. As far we know, our theory is the first to identify an offline algorithm that provably attains minimax-optimal statistical efficiency for the entire ε-range, which in turn makes clear that *no burn-in phase is needed* to achieve optimal statistical accuracy. Achieving this requires developing a new suite of statistical theory that works all the way to *the most data-hungry regime*. It is noteworthy that the existing statistical toolbox — not merely for offline RL, but for online exploratory RL as well (as we shall detail in Section 5) — is only guaranteed to work when the total sample size already exceeds a fairly large threshold, a (often unnecessary) requirement that substantially simplifies statistical analysis. In this sense, the regime we aim to solve is reminiscent of the subfield of high-dimensional statistics (Donoho et al., 2000; Wainwright, 2019a) that helps extend the frontier of classical statistics to the sample-starved regime, for which an enriched statistical toolbox is critically needed.

1.4 Notation

Throughout this paper, we adopt the convention that $0/0 = 0$. We use $\Delta(\mathcal{S})$ to indicate the probability simplex over the set $\mathcal{S}$, and denote by $[H]$ the set $\{1, \cdots, H\}$ for any positive integer H. We use $\mathbb{1}(\cdot)$ to represent the indicator function. For any vector $x = [x(s,a)]_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{SA}$, we overload the notation by letting $x^2 = [x(s,a)^2]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$. For two vectors $a = [a_i]_{1\le i\le n}$ and $b = [b_i]_{1\le i\le n}$, $a \circ b = [a_i b_i]_{1\le i\le n}$ denotes their Hadamard product, and $a \ge b$ (resp. $a \le b$) means $a_i \ge b_i$ (resp. $a_i \le b_i$) for all i. Following the convention in RL (Agarwal et al., 2021), the norm $\|\cdot\|_1$ of a matrix $P = [P_{ij}]$ is defined to be $\|P\|_1 := \max_i \sum_j |P_{ij}|$. For any probability vector $q \in \mathbb{R}^{1\times S}$ (which is a row vector) and any vector $V \in \mathbb{R}^S$, define

$$\mathsf{Var}_q(V) := q\big(V \circ V\big) - (qV)^2 \in \mathbb{R} \tag{8}$$

with $qV = \sum_i q_i V_i$, which corresponds to the variance of V w.r.t. the distribution q. The standard notation $O(\cdot)$ is adopted to represent the orderwise scaling of a function.

2 Algorithm and theory: discounted infinite-horizon MDPs

We begin by studying offline RL in discounted infinite-horizon Markov decision processes. In the following, we shall first introduce the models and assumptions for discounted infinite-horizon MDPs, followed by algorithm design and main results.

horizon	algorithm	sample complexity	ε-range to attain sample optimality	type
infinite	VI-LCB (Rashidinejad et al., 2022)	$\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}$	——	model-based
	Q-LCB (Yan et al., 2023)	$\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}$	——	model-free
	VR-Q-LCB (Yan et al., 2023)	$\frac{SC^\star}{(1-\gamma)^3\varepsilon^2} + \frac{SC^\star}{(1-\gamma)^4\varepsilon}$	$(0, 1-\gamma]$	model-free
	VI-LCB (this paper: Theorem 1)	$\frac{SC^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}$ $(\leq \frac{SC^\star}{(1-\gamma)^3\varepsilon^2})$	$\big(0, \frac{1}{1-\gamma}\big]$	model-based
	lower bound (this paper: Theorem 2)	$\frac{SC^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}$	——	——
finite	VI-LCB (Xie et al., 2021)	$\frac{H^6SC^\star}{\varepsilon^2}$	——	model-based
	VPVI (Yin and Wang, 2021)	$\frac{H^5SC^\star}{\varepsilon^2}$	——	model-based
	PEVI-Adv (Xie et al., 2021)	$\frac{H^4SC^\star}{\varepsilon^2} + \frac{H^{6.5}SC^\star}{\varepsilon}$	$\big(0, \frac{1}{H^{2.5}}\big]$	model-based
	LCB-Q-Advantage (Shi et al., 2022b)	$\frac{H^4SC^\star}{\varepsilon^2} + \frac{H^5SC^\star}{\varepsilon}$	$\big(0, \frac{1}{H}\big]$	model-free
	APVI/LCBVI (Yin and Wang, 2021)	$\frac{H^4SC^\star}{\varepsilon^2} + \frac{H^4}{d^{\mathsf{b}}_{\min}\varepsilon}$	$(0, SC^\star d^{\mathsf{b}}_{\min}]$	model-based
	VI-LCB (this paper: Theorem 3)	$\frac{H^4SC^\star_{\mathsf{clipped}}}{\varepsilon^2}$ $(\leq \frac{H^4SC^\star}{\varepsilon^2})$	$(0, H]$	model-based
	lower bound (this paper: Theorem 4)	$\frac{H^4SC^\star_{\mathsf{clipped}}}{\varepsilon^2}$	——	——

Table 1: Comparisons with prior results (up to log terms) regarding finding an ε-optimal policy in offline RL. The ε-range stands for the range of accuracy level ε for which the derived sample complexity is optimal. Here, one always has $C^\star_{\mathsf{clipped}} \leq C^\star$; and the parameter $d^{\mathsf{b}}_{\min} \coloneqq \frac{1}{\min_{s,a,h}\{d^{\mathsf{b}}_h(s,a):\, d^{\mathsf{b}}_h(s,a)>0\}}$ employed in Yin and Wang (2021) could be exceedingly small, with d^{b}_h the occupancy distribution of the dataset. While multiple algorithms are referred to as VI-LCB in the table, they correspond to different variants of VI-LCB. Our results are the first to achieve sample optimality for the full ε-range.

2.1 Models and assumptions

Let us begin with some preliminary concepts and notation of discounted infinite-horizon MDPs, followed by a concrete setting specific to offline RL. A more detailed introduction of discounted infinite-horizon MDPs can be found in classical textbooks like Bertsekas (2017).

Basics of discounted infinite-horizon MDPs. Consider a discounted infinite-horizon MDP (Bertsekas, 2017) represented by a tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, \gamma, r\}$. The key components of $\mathcal{M}$ are: (i) $\mathcal{S} = \{1, 2, \cdots, S\}$: a finite state space of size S; (ii) $\mathcal{A} = \{1, 2, \cdots, A\}$: an action space of size A; (iii) $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: the transition probability kernel of the MDP (i.e., $P(\cdot \,|\, s, a)$ denotes the transition probability from state s when action a is executed); (iv) $\gamma \in [0, 1)$: the discount factor, so that $\frac{1}{1-\gamma}$ represents the effective horizon; (v) $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$: the deterministic reward function (namely, $r(s, a)$ indicates the immediate reward received when the current state-action pair is (s, a)). Without loss of generality, the immediate rewards are normalized so that they are contained within the interval $[0, 1]$. Throughout this section, we introduce the convenient notation

$$P_{s,a} \coloneqq P(\cdot \,|\, s, a) \in \mathbb{R}^{1\times S}. \tag{9}$$

Policy, value function and Q-function. A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a possibly randomized action selection rule; that is, $\pi(a \,|\, s)$ represents the probability of choosing a in state s. When π is a deterministic policy, we abuse the notation by letting $\pi(s)$ represent the action chosen by the policy π in state s. A sample trajectory induced by the MDP under policy π can be written as $\{(s_t, a_t)\}_{t\geq 0}$, with s_t (resp. a_t) denoting the state (resp. action) of the trajectory at time t. To proceed, we shall also introduce the value function V^π and Q-value function Q^π associated with policy π. Specifically, the value function $V^\pi : \mathcal{S} \to \mathbb{R}$ of policy π is defined as the expected discounted cumulative reward as follows:

$$\forall s \in \mathcal{S}: \qquad V^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s; \pi\right], \tag{10}$$

where the expectation is taken over the sample trajectory $\{(s_t, a_t)\}_{t\geq 0}$ generated in a way that $a_t \sim \pi(\cdot \,|\, s_t)$ and $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ for all $t \geq 0$. Given that all immediate rewards lie within $[0, 1]$, it is easily verified that $0 \leq V^\pi(s) \leq \frac{1}{1-\gamma}$ for any policy π. The Q-function (or action-state function) of policy π can be defined analogously as follows:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \qquad Q^\pi(s, a) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a; \pi\right], \tag{11}$$

which differs from (10) in that it is also conditioned on $a_0 = a$.

Let $\rho \in \Delta(\mathcal{S})$ be a given state distribution. If the initial state is randomly drawn from ρ, then we can define the following weighted value function of policy π:

$$V^\pi(\rho) \coloneqq \mathop{\mathbb{E}}_{s\sim\rho}\left[V^\pi(s)\right]. \tag{12}$$

In addition, we introduce the *discounted occupancy distributions* associated with policy π as follows:

$$\forall s \in \mathcal{S}: \quad d^\pi(s; \rho) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \mid s_0 \sim \rho; \pi), \tag{13}$$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \qquad d^\pi(s, a; \rho) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a \mid s_0 \sim \rho; \pi), \tag{14}$$

where we consider the randomness over a sample trajectory that starts from an initial state $s_0 \sim \rho$ and that follows policy π (i.e., $a_t \sim \pi(\cdot \,|\, s_t)$ and $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ for all $t \geq 0$).

It is known that there exists at least one deterministic policy — denoted by $\pi^\star$ — that simultaneously maximizes $V^\pi(s)$ and $Q^\pi(s, a)$ for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ (Bertsekas, 2017). We use the following shorthand notation to represent respectively the resulting optimal value and optimal Q-function:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad V^\star(s) \coloneqq V^{\pi^\star}(s) \quad \text{and} \quad Q^\star(s, a) \coloneqq Q^{\pi^\star}(s, a). \tag{15}$$

Correspondingly, we adopt the notation of the discounted occupancy distributions associated with $\pi^\star$ as:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad d^\star(s) \coloneqq d^{\pi^\star}(s; \rho) \quad \text{and} \quad d^\star(s, a) \coloneqq d^{\pi^\star}(s, a; \rho) = d^\star(s)\, \mathbb{1}\left(a = \pi^\star(s)\right), \tag{16}$$

where the last equality is valid since $\pi^\star$ is assumed to be deterministic.

Offline/batch data. Let us work with an independent sampling model as studied in the prior work Rashidinejad et al. (2022). To be precise, imagine that we observe a batch dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{1\leq i\leq N}$ containing N sample transitions. These samples are independently generated based on a distribution $d^{\mathsf{b}} \in \Delta(\mathcal{S} \times \mathcal{A})$ and the transition kernel P of the MDP, namely,

$$(s_i, a_i) \overset{\text{ind.}}{\sim} d^{\mathsf{b}} \qquad \text{and} \qquad s_i' \overset{\text{ind.}}{\sim} P(\cdot \,|\, s_i, a_i), \qquad 1 \leq i \leq N. \tag{17}$$

In addition, it is assumed that the learner is aware of the reward function.

In order to capture the distribution shift between the desired occupancy measure and the data distribution, we introduce a key quantity previously introduced in Rashidinejad et al. (2022).

Definition 1 (Single-policy concentrability for infinite-horizon MDPs). *The single-policy concentrability coefficient of a batch dataset $\mathcal{D}$ is defined as*

$$C^\star := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d^\star(s,a)}{d^{\mathsf{b}}(s,a)}. \tag{18}$$

Clearly, one necessarily has $C^\star \geq 1$.

In words, $C^\star$ measures the distribution mismatch in terms of the maximum density ratio. The batch dataset can be viewed as expert data when $C^\star$ approaches 1, meaning that the batch dataset is close to the target policy in terms of the induced distributions. Moreover, this coefficient $C^\star$ is referred to as the "single-policy" concentrability coefficient since it is concerned with a single policy $\pi^\star$; this is clearly a much weaker assumption compared to the all-policy concentrability assumption (as adopted in, e.g., Chen and Jiang (2019); Fan et al. (2019); Farahmand et al. (2010); Munos (2007); Ren et al. (2021); Xie and Jiang (2021)), the latter of which assumes a uniform density-ratio bound over all policies and requires the dataset to be highly exploratory.

In the current paper, we also introduce a slightly improved version of $C^\star$ as follows.

Definition 2 (Single-policy clipped concentrability for infinite-horizon MDPs). *The single-policy clipped concentrability coefficient of a batch dataset $\mathcal{D}$ is defined as*

$$C^\star_{\mathsf{clipped}} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\min\left\{d^\star(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)}. \tag{19}$$

Remark 1. A direct comparison of Conditions (18) and (19) implies that for a given batch dataset $\mathcal{D}$,

$$C^\star_{\mathsf{clipped}} \leq C^\star. \tag{20}$$

As we shall see later, while our sample complexity upper bounds will be mainly stated in terms of $C^\star_{\mathsf{clipped}}$, all of them remain valid if $C^\star_{\mathsf{clipped}}$ is replaced with $C^\star$. Additionally, in contrast to $C^\star$ that is always lower bounded by 1, we have a smaller lower bound as follows (directly from the definition (19))

$$C^\star_{\mathsf{clipped}} \geq 1/S, \tag{21}$$

which is nearly tight.[1] This attribute could lead to sample size saving in some cases, to be detailed shortly.

Let us take a moment to further interpret the coefficient in Definition 2, which says that

$$d^{\mathsf{b}}(s,a) \geq \begin{cases} \frac{1}{C^\star_{\mathsf{clipped}}} d^\star(s,a), & \text{if } d^\star(s,a) \leq 1/S \\ \frac{1}{C^\star_{\mathsf{clipped}} S}, & \text{if } d^\star(s,a) > 1/S \end{cases} \tag{22}$$

holds for any pair (s,a). Consider, for instance, the case where $C^\star_{\mathsf{clipped}} = O(1)$: if a state-action pair is infrequently (or rarely) visited by the optimal policy, then it is fine for the associated density in the batch data to be very small (e.g., a density proportional to that of the optimal policy); by contrast, if a state-action pair is visited fairly often by the optimal policy, then Definition 2 might only require $d^{\mathsf{b}}(s,a)$ to exceed the order of $1/S$. In other words, the required level of $d^{\mathsf{b}}(s,a)$ is clipped at the level $\frac{1}{C^\star_{\mathsf{clipped}} S}$ regardless of the value of $d^\star(s,a)$.

[1] As a concrete example, suppose that $d^\star(s) = \begin{cases} 1 - \frac{S-1}{S^3} & \text{if } s = 1 \\ \frac{1}{S^3} & \text{else} \end{cases}$ and $d^{\mathsf{b}}(s,a) = \begin{cases} 1 - \frac{S-1}{S^2} & \text{if } a = \pi^\star(s) \text{ and } s = 1, \\ \frac{1}{S^2} & \text{if } a = \pi^\star(s) \text{ and } s \neq 1, \\ 0, & \text{else.} \end{cases}$ Then it can be easily verified that $C^\star_{\mathsf{clipped}} = \frac{1}{S-1+\frac{1}{S}}$. Nonetheless, caution should be exercised that an exceedingly small $C^\star_{\mathsf{clipped}}$ requires highly compressible structure of $d^\star$, and the real-world data often do not fall within this benign range of $C^\star_{\mathsf{clipped}}$.

Goal. Armed with the batch dataset $\mathcal{D}$, the objective of offline RL in this case is to find a policy $\widehat{\pi}$ that attains near-optimal value functions — with respect to a given test state distribution $\rho \in \Delta(\mathcal{S})$ — in a sample-efficient manner. To be precise, for a prescribed accuracy level ε, we seek to identify an ε-optimal policy $\widehat{\pi}$ satisfying

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon \tag{23}$$

with high probability, using a batch dataset $\mathcal{D}$ (cf. (17)) containing as few samples as possible. Particular emphasis is placed on achieving minimal sample complexity for the entire range of accuracy levels (namely, for any $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$).

2.2 Algorithm: VI-LCB for infinite-horizon MDPs

In this subsection, we introduce a model-based offline RL algorithm that incorporates lower concentration bounds in value estimation. The algorithm, called VI-LCB, applies value iteration (based on some pessimistic Bellman operator) to the empirical MDP, with the key ingredients described below.

The empirical MDP. Recall that we are given N independent sample transitions $\{(s_i, a_i, s_i')\}_{i=1}^{N}$ in the dataset $\mathcal{D}$. For any given state-action pair (s, a), we denote by

$$N(s,a) \coloneqq \sum_{i=1}^{N} \mathbb{1}\left((s_i, a_i) = (s, a)\right) \tag{24}$$

the number of samples transitions from (s, a). We then construct an empirical transition matrix $\widehat{P}$ such that

$$\widehat{P}(s' \,|\, s, a) = \begin{cases} \frac{1}{N(s,a)} \sum_{i=1}^{N} \mathbb{1}\big\{(s_i, a_i, s_i') = (s, a, s')\big\}, & \text{if } N(s,a) > 0 \\ \frac{1}{S}, & \text{else} \end{cases} \tag{25}$$

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

The pessimistic Bellman operator. Our offline algorithm is developed based on finding the fixed point of some variant of the classical Bellman operator. Let us first introduce this key operator and eludicate how the pessimism principle is enforced. Recall that the Bellman operator $\mathcal{T}(\cdot) : \mathbb{R}^{SA} \to \mathbb{R}^{SA}$ w.r.t. the transition kernel P is defined such that for any vector $Q \in \mathbb{R}^{SA}$,

$$\mathcal{T}(Q)(s,a) \coloneqq r(s,a) + \gamma P_{s,a} V \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{26}$$

where $V = [V(s)]_{s \in \mathcal{S}}$ with $V(s) \coloneqq \max_a Q(s,a)$. We propose to penalize the original Bellman operator w.r.t. the empirical kernel $\widehat{P}$ as follows:

$$\widehat{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a) \coloneqq \max\Big\{r(s,a) + \gamma \widehat{P}_{s,a} V - b(s,a;V), 0\Big\} \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{27}$$

where $b(s,a;V)$ denotes the penalty term employed to enforce pessimism amid uncertainty. As one can anticipate, the properties of the fixed point of $\widehat{\mathcal{T}}_{\mathsf{pe}}(\cdot)$ relies heavily upon the choice of the penalty terms $\{b_h(s,a;V)\}$, often derived based on certain concentration bounds. In this paper, we focus on the following Bernstein-style penalty to exploit the importance of certain variance statistics:

$$b(s,a;V) \coloneqq \min\left\{ \max\left\{ \sqrt{\frac{c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)} \mathsf{Var}_{\widehat{P}_{s,a}}(V)}, \frac{2c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} \right\}, \frac{1}{1-\gamma} \right\} + \frac{5}{N} \tag{28}$$

for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $c_{\mathsf{b}} > 0$ is some numerical constant (e.g., $c_{\mathsf{b}} = 144$), and $\delta \in (0,1)$ is some given quantity (in fact, $1-\delta$ is the target success probability). Here, for any vector $V \in \mathbb{R}^S$, we recall that $\mathsf{Var}_{\widehat{P}_{s,a}}(V)$ is the variance of V w.r.t. the distribution $\widehat{P}_{s,a}$ (see (8)).

We immediately isolate several useful properties as follows, whose proof is postponed to Appendix A.1.

Algorithm 1: Offline value iteration with LCB (VI-LCB) for discounted infinite-horizon MDPs

input: dataset $\mathcal{D}$; reward function r; target success probability $1-\delta$; max iteration number $\tau_{\max}$.
initialization: $\widehat{Q}_0 = 0$, $\widehat{V}_0 = 0$.
construct the empirical transition kernel $\widehat{P}$ according to (25).
for $\tau = 1, 2, \cdots, \tau_{\max}$ **do**
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    compute the penalty term $b\big(s,a;\widehat{V}_{\tau-1}\big)$ according to (28).
    set $\widehat{Q}_\tau(s,a) = \max\big\{r(s,a) + \gamma \widehat{P}_{s,a}\widehat{V}_{\tau-1} - b\big(s,a;\widehat{V}_{\tau-1}\big), 0\big\}$.
  for $s \in \mathcal{S}$ **do**
    set $\widehat{V}_\tau(s) = \max_a \widehat{Q}_\tau(s,a)$.
output: $\widehat{\pi}$ s.t. $\widehat{\pi}(s) \in \arg\max_a \widehat{Q}_{\tau_{\max}}(s,a)$ for any $s \in \mathcal{S}$.

Lemma 1. *For any $\gamma \in [\frac{1}{2}, 1)$, the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}(\cdot)$ (cf. (27)) with the Bernstein-style penalty (28) is a γ-contraction w.r.t. $\|\cdot\|_\infty$, that is,*

$$\big\|\widehat{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}(Q_2)\big\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty \tag{29}$$

for any $Q_1, Q_2 \in \mathbb{R}^{S\times A}$ obeying $Q_1(s,a), Q_2(s,a) \in \big[0, \frac{1}{1-\gamma}\big]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. In addition, there exists a unique fixed point $\widehat{Q}^\star_{\mathsf{pe}}$ of the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}(\cdot)$, which also obeys $0 \le \widehat{Q}^\star_{\mathsf{pe}}(s,a) \le \frac{1}{1-\gamma}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

In words, even though $\widehat{\mathcal{T}}_{\mathsf{pe}}(\cdot)$ integrates the penalty terms, it still preserves the γ-contraction property and admits a unique fixed point, thereby resembling the classical Bellman operator (26).

The VI-LCB algorithm. We are now positioned to introduce the VI-LCB algorithm, which can be regarded as classical value iteration applied in conjunction with pessimism. Specifically, the algorithm applies the Bernstein-style pessimistic operator $\widehat{\mathcal{T}}_{\mathsf{pe}}$ (cf. (27)) iteratively in order to find its fixed point:

$$\widehat{Q}_\tau(s,a) = \widehat{\mathcal{T}}_{\mathsf{pe}}\big(\widehat{Q}_{\tau-1}\big)(s,a) = \max\Big\{r(s,a) + \gamma \widehat{P}_{s,a}\widehat{V}_{\tau-1} - b\big(s,a;\widehat{V}_{\tau-1}\big), 0\Big\}, \qquad \tau = 1, 2, \cdots \tag{30}$$

We shall initialize it to $\widehat{Q}_0 = 0$, implement (30) for $\tau_{\max}$ iterations, and output $\widehat{Q} = \widehat{Q}_{\tau_{\max}}$ as the final Q-estimate. The final policy estimate $\widehat{\pi}$ is chosen on the basis of $\widehat{Q}$ as follows:

$$\widehat{\pi}(s) \in \arg\max_a \widehat{Q}(s,a) \qquad \text{for all } s \in \mathcal{S}, \tag{31}$$

with the whole algorithm summarized in Algorithm 1.

Let us pause to explain the rationale of the pessimism principle on a high level. If a pair (s,a) has been insufficiently visited in $\mathcal{D}$ (i.e., $N(s,a)$ is small), then the resulting Q-estimate $\widehat{Q}_\tau(s,a)$ could suffer from high uncertainty and become unreliable, which might in turn mislead value estimation. By enforcing suitable penalization $b(s,a;\widehat{V}_{\tau-1})$ based on certain lower confidence bounds, we can suppress the negative influence of such poorly visited state-action pairs. Fortunately, suppressing these state-action pairs might not result in significant bias in value estimation when $C^\star_{\mathsf{clipped}}$ is small; for instance, when the behavior policy π^{b} resembles $\pi^\star$, the poorly visited state-action pairs correspond primarily to suboptimal actions (as they are not selected by $\pi^\star$), making it acceptable to neglect these pairs.

Interestingly, Algorithm 1 is guaranteed to converge rapidly. In view of the γ-contraction property in Lemma 1, the iterates $\{\widehat{Q}_\tau\}_{\tau\ge 0}$ converge linearly to the fixed point $\widehat{Q}^\star_{\mathsf{pe}}$, as asserted below.

Lemma 2. *Suppose $\widehat{Q}_0 = 0$. Then the iterates of Algorithm 1 obey*

$$\widehat{Q}_\tau \le \widehat{Q}^\star_{\mathsf{pe}} \qquad \text{and} \qquad \big\|\widehat{Q}_\tau - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \le \frac{\gamma^\tau}{1-\gamma} \qquad \text{for all } \tau \ge 0, \tag{32}$$

where $\widehat{Q}^{\star}_{\mathsf{pe}}$ *is the unique fixed point of* $\widehat{\mathcal{T}}_{\mathsf{pe}}$*. As a consequence, by choosing* $\tau_{\max} \geq \frac{\log \frac{N}{1-\gamma}}{\log(1/\gamma)}$ *one fulfills*

$$\left\|\widehat{Q}_{\tau_{\max}} - \widehat{Q}^{\star}_{\mathsf{pe}}\right\|_{\infty} \leq 1/N. \tag{33}$$

The proof of this lemma is deferred to Appendix A.2.

Algorithmic comparison with Rashidinejad et al. (2022). VI-LCB has been studied in the prior work Rashidinejad et al. (2022). The difference between our algorithm and the version therein is two-fold:

- *Sample reuse vs.* $\widetilde{O}\big(\frac{1}{1-\gamma}\big)$*-fold sample splitting.* Our algorithm reuses the same set of samples across all iterations, which is in sharp contrast to Rashidinejad et al. (2022) that employs fresh samples in each of the $\widetilde{O}\big(\frac{1}{1-\gamma}\big)$ iterations. This results in considerably better usage of available information.
- *Bernstein-style vs. Hoeffding-style penalty.* Our algorithm adopts the Bernstein-type penalty, as opposed to the Hoeffding-style penalty in Rashidinejad et al. (2022). This choice leads to more effective exploitation of the variance structure across time.

Pessimism vs. optimism in the face of uncertainty. The careful reader might also notice the similarity between the pessimism principle and the optimism principle utilized in online RL. A well-developed paradigm that balances exploration and exploitation in online RL is optimistic exploration based on uncertainty quantification (Lai and Robbins, 1985). The earlier work Jaksch et al. (2010) put forward an algorithm called UCRL2 that computes an optimistic policy with the aid of Hoeffding-style confidence regions for the probability transition kernel. Later on, Azar et al. (2017) proposed to build upper confidence bounds (UCB) for the optimal values instead, which leads to significantly improved sample complexity; see, e.g., He et al. (2021); Wang et al. (2019) for the application of this strategy to discounted infinite-horizon MDPs. Note, however, that the rationales behind optimism and pessimism are remarkably different. In offline RL (which does not allow further data collection), the uncertainty estimates are employed to identify, and then rule out, poorly-visited actions; this stands in sharp contrast to the online counterpart where poorly-visited actions might be more favored during exploration.

2.3 Performance guarantees

When the Bernstein-style concentration bound (28) is adopted, the VI-LCB algorithm in Algorithm 1 yields ε-accuracy with a near-minimal number of samples, as stated below.

Theorem 1. *Suppose* $\gamma \in [\frac{1}{2}, 1)$*, and consider any* $0 < \delta < 1$ *and* $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$*. Suppose that the total number of iterations exceeds* $\tau_{\max} \geq \frac{1}{1-\gamma} \log \frac{N}{1-\gamma}$*. With probability at least* $1 - 2\delta$*, the policy* $\widehat{\pi}$ *returned by Algorithm 1 obeys*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon, \tag{34}$$

provided that c_{b} *(cf. the Bernstein-style penalty term in (28)) is some sufficiently large numerical constant and the total sample size exceeds*

$$N \geq \frac{c_1 S C^{\star}_{\mathsf{clipped}} \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3 \varepsilon^2} \tag{35}$$

for some large enough numerical constant $c_1 > 0$*, where* $C^{\star}_{\mathsf{clipped}}$ *is introduced in Definition 2. In addition, the above result continues to hold if* $C^{\star}_{\mathsf{clipped}}$ *is replaced with* $C^{\star}$ *(introduced in Definition 1).*

Remark 2. Regarding the numerical constants in Theorem 1, a conservative yet concrete sufficient condition is that $c_{\mathsf{b}} \geq 144$ and $c_1 = 21000 c_{\mathsf{b}}$, which we shall rigorize in the proof.

The proof of this theorem is postponed to Section 6. In general, the total sample size characterized by Theorem 1 could be far smaller than the ambient dimension (i.e., S^2A) of the transition kernel P, thus precluding one from estimating P in a reliable fashion. As a crucial insight from Theorem 1, the model-based (or plug-in) approach enables reliable offline learning even when model estimation is completely off.

Before discussing key implications of Theorem 1, we develop matching minimax lower bounds that help confirm the efficacy of the proposed model-based algorithm, whose proof can be found in Appendix C.2.

Theorem 2. *For any* $(\gamma, S, C^\star_{\mathsf{clipped}}, \varepsilon)$ *obeying* $\gamma \in \left[\frac{2}{3}, 1\right)$, $S \geq 2$, $C^\star_{\mathsf{clipped}} \geq \frac{8\gamma}{S}$, *and* $\varepsilon \leq \frac{1}{42(1-\gamma)}$, *one can construct two MDPs* $\mathcal{M}_0, \mathcal{M}_1$, *an initial state distribution* ρ, *and a batch dataset with* N *independent samples and single-policy clipped concentrability coefficient* $C^\star_{\mathsf{clipped}}$ *such that*

$$\inf_{\widehat{\pi}} \max\left\{\mathbb{P}_0\big(V^\star(\rho) - V^{\widehat{\pi}}(\rho) > \varepsilon\big),\, \mathbb{P}_1\big(V^\star(\rho) - V^{\widehat{\pi}}(\rho) > \varepsilon\big)\right\} \geq \frac{1}{8},$$

provided that

$$N \leq \frac{c_2 S C^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}$$

for some numerical constant $c_2 > 0$. *Here, the infimum is over all estimator* $\widehat{\pi}$, *and* $\mathbb{P}_0$ *(resp.* $\mathbb{P}_1$*) denotes the probability when the MDP is* $\mathcal{M}_0$ *(resp.* $\mathcal{M}_1$*).*

Remark 3. As a more concrete (yet conservative) condition for c_2, Theorem 2 is valid when $c_2 = 1/25088$.

Implications. In the following, we take a moment to interpret the above two theorems and single out several key implications about the proposed model-based algorithm.

- *Optimal sample complexities.* In the presence of the Bernstein-style penalty, the total number of samples needed for our algorithm to yield ε-accuracy is

$$\widetilde{O}\left(\frac{SC^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\right). \tag{36}$$

  This taken together with the minimax lower bound asserted in Theorem 2 confirms the optimality of the proposed model-based approach (up to some logarithmic factor). In comparison, the sample complexity derived in Rashidinejad et al. (2022) exhibits a worse dependency on the effective horizon (i.e., $\frac{1}{(1-\gamma)^5}$). Theorem 2 also enhances the lower bound developed in Rashidinejad et al. (2022) to accommodate the scenario where $C^\star_{\mathsf{clipped}}$ can be much smaller than $C^\star$, i.e., $C^\star_{\mathsf{clipped}} = O(1/S)$.

- *No burn-in cost.* The fact that the sample size bound (35) holds for the full ε-range (i.e., any given $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$) means that there is no burn-in cost required to achieve sample optimality. This not only drastically improves upon, but in fact eliminates, the burn-in cost of the best-known sample-optimal result (cf. (5)), the latter of which required a burn-in cost at least on the order of $\frac{SC^\star}{(1-\gamma)^5}$. Accomplishing this requires one to tackle the sample-hungry regime, which is statistically challenging to cope with.

- *No need of sample splitting.* It is noteworthy that prior works typically required sample splitting. For instance, Rashidinejad et al. (2022) analyzed the VI-LCB algorithm with fresh samples employed in each iteration, which effectively split the data into $\widetilde{O}\left(\frac{1}{1-\gamma}\right)$ disjoint subsets. In contrast, the algorithm studied herein permits the reuse of all samples across all iterations. This is an important feature in sample-starved applications to effectively maximize information utilization, and is a crucial factor that assists in improving the sample complexity compared to Rashidinejad et al. (2022).

- *Sample size saving when* $C^\star_{\mathsf{clipped}} < 1$. In view of Theorem 1, the sample complexity of the proposed algorithm can be as low as

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

  when $C^\star_{\mathsf{clipped}}$ is on the order of $1/S$. This might seem somewhat surprising at first glance, given that the minimax sample complexity for policy evaluation is at least $\widetilde{O}\left(\frac{S}{(1-\gamma)^3\varepsilon^2}\right)$ even in the presence of a simulator (Azar et al., 2013). To elucidate this, we note that the condition $C^\star_{\mathsf{clipped}} = O(1/S)$ implicitly imposes special — in fact, highly compressible — structure on the MDP that enables sample size reduction. As we shall see from the lower bound construction in Theorem 2, the case with $C^\star_{\mathsf{clipped}} = O(1/S)$ might require $d^\star(s,a)$ to concentrate on one or a small number of important states, with exceedingly small probability assigned to the remaining ones. If this occurs, then it often suffices to focus on what happens on these important states, thus requiring much fewer samples.

Comparisons with prior statistical analysis. Before concluding this section, we highlight the innovations of our statistical analysis compared to past theory when it comes to discounted infinite-horizon MDPs. To begin with, our sample size improvement over Rashidinejad et al. (2022) stems from the two algorithmic differences mentioned in Section 2.2: the sample-reuse feature allows one to improve a factor of $\frac{1}{1-\gamma}$, while the use of Bernstein-style penalty yields an additional gain of $\frac{1}{1-\gamma}$. In addition, while the design of data-driven Bernstein-style bounds has been extensively studied in online RL in discounted MDPs (e.g., He et al. (2021); Zhang et al. (2021b)), all of these past results were either sample-suboptimal, or required a huge burn-in sample size (e.g., $\frac{S^3A^2}{(1-\gamma)^4}$ in He et al. (2021)). In other words, sample optimality was not previously achieved in the most data-hungry regime. In comparison, our theory ensures optimality of our algorithm even for the most sample-constrained scenario, which relies on much more delicate statistical tools. In a nutshell, our statistical analysis is built upon at least two ideas: (i) a leave-one-out analysis framework that allows to decouple complicated statistical dependency across iterations without losing statistical tightness; (ii) a delicate self-bounding trick that allows us to simultaneously control multiple crucial statistical quantities (e.g., empirical variance) in the most sample-starved regime.

3 Algorithm and theory: episodic finite-horizon MDPs

In this section, we turn attention to the studies of offline RL for episodic finite-horizon MDPs.

3.1 Models and assumptions

As before, we briefly state some preliminaries about finite-horizon MDPs, before moving on to the sampling model and the goal. The readers can consult Bertsekas (2017) for more details about finite-horizon MDPs.

Basics of finite-horizon MDPs. Consider the setting of a finite-horizon Markov decision process, as denoted by $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, P, r\}$. It consists of the following key components: (i) $\mathcal{S} = \{1, \cdots, S\}$: a state space of size S; (ii) $\mathcal{A} = \{1, \cdots, A\}$: an action space of size A; (iii) H: the horizon length; (iv) $P = \{P_h\}_{1\le h\le H}$, with $P_h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ denoting the probability transition kernel at step h (namely, $P_h(\cdot \,|\, s, a)$ stands for the transition probability of the MDP at step h when the current state-action pair is (s, a)); (v) $r = \{r_h\}_{1\le h\le H}$, with $r_h : \mathcal{S}\times\mathcal{A} \to [0, 1]$ denoting the reward function at step h (namely, $r_h(s, a)$ indicates the immediate reward gained at step h when the current state-action pair is (s, a)). It is assumed without loss of generality that the immediate rewards fall within the interval $[0, 1]$ and are deterministic. Conveniently, we introduce the following S-dimensional row vector

$$P_{h,s,a} \coloneqq P_h(\cdot \,|\, s, a) \tag{37}$$

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

A (possibly randomized) policy $\pi = \{\pi_h\}_{1\le h\le H}$ with $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ is an action selection rule, such that $\pi_h(a \,|\, s)$ specifies the probability of choosing action a when in state s and step h. When π is a deterministic policy, we overload the notation and let $\pi_h(s)$ represent the action selected by π in state s at step h. We can generate a sample trajectory $\{(s_h, a_h)\}_{1\le h\le H}$ by implementing policy π in the MDP $\mathcal{M}$, where s_h and a_h denote the state and the action in step h, respectively. We then introduce the value function $V^\pi = \{V_h^\pi\}_{1\le h\le H}$ and the Q-function $Q^\pi = \{Q_h^\pi\}_{1\le h\le H}$ associated with policy π; specifically, the value function $V_h : \mathcal{S} \to \mathbb{R}$ of policy π at step h is defined to the be the expected cumulative reward from step h on as a result of policy π, namely,

$$\forall s \in \mathcal{S}: \qquad V_h^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \mid s_h = s; \pi\right], \tag{38}$$

where the expectation is taken over the randomness over the sample trajectory $\{(s_t, a_t)\}_{t=h}^H$ when policy π is implemented (i.e., $a_t \sim \pi_t(\cdot \,|\, s_t)$ and $s_{t+1} \sim P_t(\cdot \,|\, s_t, a_t)$ for all $t \ge h$). Correspondingly, the Q-function of policy π at step h is defined to be

$$\forall (s, a) \in \mathcal{S}\times\mathcal{A}: \qquad Q_h^\pi(s, a) \coloneqq \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \mid s_h = s, a_h = a; \pi\right] \tag{39}$$

when conditioned on the state-action pair (s,a) at step h. If the initial state is drawn from a distribution $\rho \in \Delta(\mathcal{S})$, we find it convenient to define the following weighted value function of policy π:

$$V_1^\pi(\rho) \coloneqq \mathop{\mathbb{E}}_{s\sim\rho}\big[V_1^\pi(s)\big]. \tag{40}$$

Additionally, we introduce the following *occupancy distributions* associated with policy π at step h:

$$d_h^\pi(s;\rho) \coloneqq \mathbb{P}\big(s_h = s \mid s_1 \sim \rho;\pi\big), \tag{41a}$$
$$d_h^\pi(s,a;\rho) \coloneqq \mathbb{P}\big(s_h = s, a_h = a \mid s_1 \sim \rho;\pi\big) = d_h^\pi(s;\rho)\pi(a\,|\,s), \tag{41b}$$

which are conditioned on the initial state distribution $s_1 \sim \rho$ and the event that all actions are selected according to π. In particular, it is self-evident that

$$d_1^\pi(s;\rho) = \rho(s) \qquad \text{for any policy } \pi \text{ and any state } s \in \mathcal{S}. \tag{42}$$

It is well known that there exists at least one deterministic policy that simultaneously maximizes the value function and the Q-function for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ (Bertsekas, 2017). In light of this, we shall denote by $\pi^\star = \{\pi_h^\star\}_{1\le h\le H}$ an *optimal deterministic* policy throughout this paper; this allows us to employ $\pi_h^\star(s)$ to indicate the corresponding optimal action chosen in state s at step h. The resulting optimal value function and optimal Q-function are denoted respectively by $V^\star = \{V_h^\star\}_{1\le h\le H}$ and $Q^\star = \{Q_h^\star\}_{1\le h\le H}$:

$$\forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]: \qquad V_h^\star \coloneqq V_h^{\pi^\star} \qquad \text{and} \qquad Q_h^\star \coloneqq Q_h^{\pi^\star}.$$

Furthermore, we adopt the following notation for convenience:

$$\forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]: \quad d_h^\star(s) \coloneqq d_h^{\pi^\star}(s;\rho) \quad \text{and} \quad d_h^\star(s,a) \coloneqq d_h^{\pi^\star}(s,a;\rho) = d_h^\star(s)\mathbb{1}\{a=\pi^\star(s)\}, \tag{43}$$

where the last identity holds given that $\pi^\star$ is assumed to be deterministic.

Offline/batch data. Suppose that we have access to a batch dataset (or historical dataset) $\mathcal{D}$, which comprises a collection of K i.i.d. sample trajectories generated by a behavior policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{1\le h\le H}$. More specifically, the k-th sample trajectory $(1\le k\le K)$ consists of a data sequence

$$\big(s_1^k, a_1^k, s_2^k, a_2^k, \ldots, s_H^k, a_H^k, s_{H+1}^k\big), \tag{44}$$

which is generated by the MDP $\mathcal{M}$ under the behavior policy π^{b} in the following manner:

$$s_1^k \sim \rho^{\mathsf{b}}, \qquad a_h^k \sim \pi_h^{\mathsf{b}}(\cdot\,|\,s_h^k) \qquad \text{and} \qquad s_{h+1}^k \sim P_h(\cdot\,|\,s_h^k, a_h^k), \qquad 1\le h\le H. \tag{45}$$

Here and throughout, ρ^{b} stands for some predetermined initial state distribution associated with the batch dataset. In addition to the above dataset (cf. (44) for all $1\le k\le K$), the learner also has access to the reward function. For notational simplicity, we introduce the following short-hand notation for the occupancy distribution w.r.t. the behavior policy π^{b}:

$$\forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]: \qquad d_h^{\mathsf{b}}(s) \coloneqq d_h^{\pi^{\mathsf{b}}}(s;\rho^{\mathsf{b}}) \quad \text{and} \quad d_h^{\mathsf{b}}(s,a) \coloneqq d_h^{\pi^{\mathsf{b}}}(s,a;\rho^{\mathsf{b}}). \tag{46}$$

In particular, it is easily seen that $d_1^{\mathsf{b}}(s) = \rho^{\mathsf{b}}(s)$ for all $s\in\mathcal{S}$. Note that the initial state distribution ρ^{b} of the batch dataset might not coincide with the test state distribution ρ.

Akin to Definition 1, prior works (e.g., Xie et al. (2021)) have introduced the following concentrability coefficient to capture the distribution shift between the desired distribution and the one induced by the behavior policy.

Definition 3 (Single-policy concentrability for finite-horizon MDPs)**.** *The single-policy concentrability coefficient of a batch dataset $\mathcal{D}$ is defined as*

$$C^\star \coloneqq \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{d_h^\star(s,a)}{d_h^{\mathsf{b}}(s,a)}, \tag{47}$$

which clearly satisfies $C^\star \ge 1$.

Similar to the discounted infinite-horizon counterpart, $C^\star$ employs the largest density ratio (using the occupancy distributions defined above) to measure the distribution mismatch; it concerns the behavior policy vs. a single policy $\pi^\star$, and does not require uniform coverage of the state-action space (namely, it suffices to cover the part reachable by $\pi^\star$). As before, we further introduce a slightly modified version of $C^\star$ as follows.

Definition 4 (Single-policy clipped concentrability for finite-horizon MDPs)**.** *The single-policy clipped concentrability coefficient of a batch dataset $\mathcal{D}$ is defined as*

$$C^\star_{\mathsf{clipped}} := \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{\min\left\{d^\star_h(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}_h(s,a)}. \tag{48}$$

From the definition above, it holds trivially that

$$C^\star_{\mathsf{clipped}} \le C^\star \qquad \text{and} \qquad C^\star_{\mathsf{clipped}} \ge \frac{1}{S}. \tag{49}$$

As we shall see shortly, while all sample complexity upper bounds developed herein remain valid if we replace $C^\star_{\mathsf{clipped}}$ with $C^\star$, the use of $C^\star_{\mathsf{clipped}}$ might yield some sample size reduction when $C^\star_{\mathsf{clipped}}$ drops below 1.

Goal. With the above batch dataset $\mathcal{D}$ in hand, our aim is to compute, in a sample-efficient fashion, a policy $\widehat{\pi}$ that results in near-optimal values w.r.t. a given test state distribution $\rho \in \Delta(\mathcal{S})$. Formally speaking, the current paper focuses on achieving

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \le \varepsilon$$

with high probability using as few samples as possible, where ε stands for the target accuracy level. We seek to achieve sample optimality for the full ε-range, i.e., for any $\varepsilon \in (0, H]$.

3.2 A model-based offline RL algorithm: VI-LCB

Suppose for the moment that we have access to a dataset $\mathcal{D}_0$ containing N sample transitions $\{(s_i, a_i, h_i, s_i')\}_{i=1}^N$, where (s_i, a_i, h_i, s_i') denotes the transition from state s_i at step h_i to state s_i' in the next step when action a_i is taken. We now describe a pessimistic variant of the model-based approach on the basis of $\mathcal{D}_0$.

Empirical MDP. For each $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, we denote by

$$N_h(s,a) := \sum_{i=1}^N \mathbb{1}\big\{(s_i,a_i,h_i) = (s,a,h)\big\} \tag{50a}$$

$$N_h(s) := \sum_{i=1}^N \mathbb{1}\big\{(s_i,h_i) = (s,h)\big\} \tag{50b}$$

the total number of sample transitions at step h that transition from (s,a) and from s, respectively. We can then compute the empirical estimate $\widehat{P} = \{\widehat{P}_h\}_{1\le h\le H}$ of the transition kernel P as follows:

$$\widehat{P}_h(s' \mid s,a) = \begin{cases} \frac{1}{N_h(s,a)} \sum_{i=1}^N \mathbb{1}\big\{(s_i,a_i,h_i,s_i') = (s,a,h,s')\big\}, & \text{if } N_h(s,a) > 0 \\ \frac{1}{S}, & \text{else} \end{cases} \tag{51}$$

for each $(s,a,h,s') \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}$.

The VI-LCB algorithm. With this estimated model in place, the VI-LCB algorithm (i.e., value iteration with lower confidence bounds) maintains the value function estimate $\{\widehat{V}_h\}$ and Q-function estimate $\{\widehat{Q}_h\}$, and works backward from $h = H$ to $h = 1$ as in classical dynamic programming with the terminal value $\widehat{V}_{H+1} = 0$ (Jin et al., 2021; Xie et al., 2021). Specifically, the algorithm adopts the following update rule:

$$\widehat{Q}_h(s,a) = \max\Big\{r_h(s,a) + \widehat{P}_{h,s,a}\widehat{V}_{h+1} - b_h(s,a), 0\Big\}, \tag{52}$$

Algorithm 2: Offline value iteration with LCB (VI-LCB) for finite-horizon MDPs.

input: dataset $\mathcal{D}_0$; reward function r; target success probability $1-\delta$.
initialization: $\widehat{V}_{H+1} = 0$.
for $h = H, \cdots, 1$ **do**
  compute the empirical transition kernel $\widehat{P}_h$ according to (51).
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    compute the penalty term $b_h(s,a)$ according to (55).
    set $\widehat{Q}_h(s,a) = \max\big\{r_h(s,a) + \widehat{P}_{h,s,a}\widehat{V}_{h+1} - b_h(s,a), 0\big\}$.
  for $s \in \mathcal{S}$ **do**
    set $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a)$ and $\widehat{\pi}_h(s) \in \arg\max_a \widehat{Q}_h(s,a)$.
output: $\widehat{\pi} = \{\widehat{\pi}_h\}_{1\le h\le H}$.

where $\widehat{P}_{h,s,a}$ is the empirical estimate of $P_{h,s,a}$ (cf. (37)),

$$\widehat{V}_{h+1}(s) = \max_a \widehat{Q}_{h+1}(s,a), \tag{53}$$

and $b_h(s,a) \ge 0$ denotes some penalty term that is a decreasing function in $N_h(s,a)$ (as we shall specify momentarily). In addition, the policy $\widehat{\pi}$ is selected greedily in accordance to the Q-estimate:

$$\forall (s,h) \in \mathcal{S}\times[H]: \qquad \widehat{\pi}_h(s) \in \arg\max_a \widehat{Q}_h(s,a). \tag{54}$$

In a nutshell, the VI-LCB algorithm — as summarized in Algorithm 2 — applies the classical value iteration approach to the empirical model $\widehat{P}$, and in addition, implements the principle of pessimism via certain lower confidence penalty terms $\{b_h(s,a)\}$.

The Bernstein-style penalty terms. As before, we adopt Bernstein-style penalty in order to better capture the variance structure over time; that is,

$$\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]: \ b_h(s,a) = \min\left\{\sqrt{\frac{c_{\mathsf{b}}\log\frac{NH}{\delta}}{N_h(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}\big(\widehat{V}_{h+1}\big)} + c_{\mathsf{b}}H\frac{\log\frac{NH}{\delta}}{N_h(s,a)}, H\right\} \tag{55}$$

for some universal constant $c_{\mathsf{b}} > 0$ (e.g., $c_{\mathsf{b}} = 16$). Here, $\mathsf{Var}_{\widehat{P}_{h,s,a}}\big(\widehat{V}_{h+1}\big)$ corresponds to the variance of $\widehat{V}_{h+1}$ w.r.t. the distribution $\widehat{P}_{h,s,a}$ (see the definition (8)). Note that we choose $\widehat{P}$ as opposed to P (i.e., $\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)$) in the variance term, mainly because we have no access to the true transition kernel P.

Finally, it is worth noting that the Bernstein-style uncertainty estimates have been widely studied when performing online exploration in episodic finite-horizon MDPs (e.g., Azar et al. (2017); Fruit et al. (2020); Jin et al. (2018); Li et al. (2021); Talebi and Maillard (2018); Zhang et al. (2020)). Once again, the main purpose therein is to encourage exploration of the insufficiently visited states/actions, a mechanism that is not applicable to offline RL due to the absence of further data collection.

3.3 VI-LCB with two-fold subsampling

Given that the batch dataset $\mathcal{D}$ is composed of several sample trajectories each of length H, the sample transitions in $\mathcal{D}$ cannot be viewed as being independently generated (as the sample transitions at step h might influence the sample transitions in the subsequent steps). As one can imagine, the presence of such temporal statistical dependency considerably complicates analysis.

In order to circumvent this technical difficulty, we propose a two-fold subsampling trick that allows one to exploit the desired statistical independence. Informally, we propose the following steps:

- First of all, we randomly split the dataset into two halves $\mathcal{D}^{\mathsf{main}}$ and $\mathcal{D}^{\mathsf{aux}}$, where $\mathcal{D}^{\mathsf{main}}$ consists of $N_h^{\mathsf{main}}(s)$ sample transitions from state s at step h.

Algorithm 3: Subsampled VI-LCB for episodic finite-horizon MDPs

input: a dataset $\mathcal{D}$; reward function r.

subsampling: run the following procedure to generate the subsampled dataset $\mathcal{D}^{\mathsf{trim}}$.

1) *Data splitting.* Split $\mathcal{D}$ into two halves: $\mathcal{D}^{\mathsf{main}}$ (which contains the first $K/2$ trajectories), and $\mathcal{D}^{\mathsf{aux}}$ (which contains the remaining $K/2$ trajectories); we let $N_h^{\mathsf{main}}(s)$ (resp. $N_h^{\mathsf{aux}}(s)$) denote the number of sample transitions in $\mathcal{D}^{\mathsf{main}}$ (resp. $\mathcal{D}^{\mathsf{aux}}$) that transition from state s at step h.

2) *Lower bounding* $\{N_h^{\mathsf{main}}(s)\}$ *using* $\mathcal{D}^{\mathsf{aux}}$. For each $s \in \mathcal{S}$ and $1 \le h \le H$, compute

$$N_h^{\mathsf{trim}}(s) \coloneqq \max\left\{N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s)\log\frac{HS}{\delta}},\, 0\right\}; \tag{56}$$

3) *Random subsampling.* Let $\mathcal{D}^{\mathsf{main}'}$ be the set of all sample transitions (i.e., the quadruples taking the form (s,a,h,s')) from $\mathcal{D}^{\mathsf{main}}$. Subsample $\mathcal{D}^{\mathsf{main}'}$ to obtain $\mathcal{D}^{\mathsf{trim}}$, such that for each $(s,h) \in \mathcal{S}\times[H]$, $\mathcal{D}^{\mathsf{trim}}$ contains $\min\{N_h^{\mathsf{trim}}(s), N_h^{\mathsf{main}}(s)\}$ sample transitions randomly drawn from $\mathcal{D}^{\mathsf{main}'}$.

run VI-LCB: set $\mathcal{D}_0 = \mathcal{D}^{\mathsf{trim}}$; run Algorithm 2 to compute a policy $\widehat{\pi}$.

- For each $(s,h) \in \mathcal{S}\times[H]$, we use the dataset $\mathcal{D}^{\mathsf{aux}}$ to construct a high-probability lower bound $N_h^{\mathsf{trim}}(s)$ on $N_h^{\mathsf{main}}(s)$, and then subsample $N_h^{\mathsf{trim}}(s)$ sample transitions w.r.t. (s,h) from $\mathcal{D}^{\mathsf{main}}$; this results in a new subsampled dataset $\mathcal{D}^{\mathsf{trim}}$.
- Run VI-LCB on the subsampled dataset $\mathcal{D}^{\mathsf{trim}}$ (i.e., Algorithm 2).

The whole procedure is detailed in Algorithm 3. A few important features are worth highlighting, under the assumption that the sample trajectories in $\mathcal{D}$ are independently generated from the same distribution.

- Given that $\{N_h^{\mathsf{trim}}(s)\}$ are computed on the basis of the dataset $\mathcal{D}^{\mathsf{aux}}$ and that $\mathcal{D}^{\mathsf{trim}}$ is subsampled from another dataset $\mathcal{D}^{\mathsf{main}}$, one can clearly see that $\{N_h^{\mathsf{trim}}(s)\}$ are statistically independent from the sample transitions in $\mathcal{D}^{\mathsf{trim}}$.
- As we shall justify in the analysis (i.e., Section 7.2), the samples in $\mathcal{D}^{\mathsf{trim}}$ can almost be treated as being statistically independent, a key attribute resulting from the subsampling trick.
- The proposed algorithm only splits the data into two subsets, which is in stark contrast to prior variants of VI-LCB that perform H-fold sample splitting (e.g., Xie et al. (2021)). Eliminating the H-fold splitting requirement plays a crucial role in enabling optimal sample complexity.

Before proceeding, we formally justify that $N_h^{\mathsf{trim}}(s)$ — as computed in (56) — is a valid lower bound on $N_h^{\mathsf{main}}(s)$. Here and below, we denote by $N_h^{\mathsf{trim}}(s,a)$ the number of sample transitions in $\mathcal{D}^{\mathsf{trim}}$ that are associated with the state-action pair (s,a) at step h. The proof of this lemma can be found in Appendix B.1.

Lemma 3. *Suppose that the K trajectories in $\mathcal{D}$ are generated in an i.i.d. fashion (see Section 3.1). With probability at least $1-8\delta$, the quantities constructed in (56) obey*

$$N_h^{\mathsf{trim}}(s) \le N_h^{\mathsf{main}}(s), \tag{57a}$$

$$N_h^{\mathsf{trim}}(s,a) \ge \frac{Kd_h^{\mathsf{b}}(s,a)}{8} - 5\sqrt{Kd_h^{\mathsf{b}}(s,a)\log\frac{KH}{\delta}} \tag{57b}$$

simultaneously for all $1 \le h \le H$ and all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

3.4 Performance guarantees

In what follows, we characterize the sample complexity of Algorithm 3, as formalized below.

Theorem 3. *Consider any $\varepsilon \in (0, H]$ and any $0 < \delta < 1$. With probability exceeding $1 - 12\delta$, the policy $\widehat{\pi}$ returned by Algorithm 3 obeys*

$$V_1^{\star}(\rho) - V_1^{\widehat{\pi}}(\rho) \le \varepsilon \tag{58}$$

as long as the penalty terms are chosen according to the Bernstein-style quantity (55) for some large enough numerical constant $c_{\mathsf{b}} > 0$, and the total number of sample trajectories exceeds

$$K \ge \frac{c_{\mathsf{k}} H^3 S C^{\star}_{\mathsf{clipped}} \log \frac{KH}{\delta}}{\varepsilon^2} \tag{59}$$

for some sufficiently large numerical constant $c_{\mathsf{k}} > 0$, where $C^{\star}_{\mathsf{clipped}}$ is introduced in Definition 4. Additionally, the above result continues to hold if $C^{\star}_{\mathsf{clipped}}$ is replaced with $C^{\star}$ (introduced in Definition 3).

Remark 4. One concrete yet conservative requirement on c_{b} and c_{k} for Theorem 3 to hold is: $c_{\mathsf{b}} \ge 16$ and $c_{\mathsf{k}} = 12800 c_{\mathsf{b}}$, as we shall solidify in the proof of Theorem 3.

The proof of this result is postponed to Section 7. In general, the total sample size characterized by Theorem 3 could be far smaller than the ambient dimension (i.e., S^2AH) of the probability transition kernel P, thus precluding one from estimating P in a reliable fashion. As a crucial insight from Theorem 3, the model-based (or plug-in) approach enables reliable policy learning even when model estimation is completely off. Our analysis of Theorem 3 relies heavily on (i) suitable decoupling of complicated statistical dependency via subsampling, and (ii) careful control of the variance terms in the presence of Bernstein-style penalty.

In order to help assess the tightness and optimality of Theoerem 3, we further develop a minimax lower bound as follows; the proof can be found in Appendix C.3.

Theorem 4. *For any $(H, S, C^{\star}_{\mathsf{clipped}}, \varepsilon)$ obeying $H \ge 12$, $C^{\star}_{\mathsf{clipped}} \ge 8/S$ and $\varepsilon \le c_3 H$, one can construct a collection of MDPs $\{\mathcal{M}_\theta \mid \theta \in \Theta\}$, an initial state distribution ρ, and a batch dataset with K independent sample trajectories each of length H, such that*

$$\inf_{\widehat{\pi}} \max_{\theta \in \Theta} \mathbb{P}_\theta \left\{ V_1^{\star}(\rho) - V_1^{\widehat{\pi}}(\rho) \ge \varepsilon \right\} \ge \frac{1}{4}, \tag{60}$$

provided that the total sample size

$$N = KH \le \frac{c_4 C^{\star}_{\mathsf{clipped}} S H^4}{\varepsilon^2}. \tag{61}$$

Here, $c_3, c_4 > 0$ are some small enough numerical constants, the infimum is over all estimator $\widehat{\pi}$, and $\mathbb{P}_\theta$ denotes the probability when the MDP is $\mathcal{M}_\theta$.

Remark 5. More concretely, one (conservative) condition regarding c_3 and c_4 that is sufficient for the validity of Theorem 4 is: $c_3 = 1/2^{14}$ and $c_4 = 1/2^{36}$, as we shall see in the proof.

Implications. In what follows, let us take a moment to discuss several other key implications of Theorem 3.

- *Near-optimal sample complexities.* In the presence of the Bernstein-style penalty, the total number of samples (i.e., KH) needed for our algorithm to yield ε-accuracy is

$$\widetilde{O}\left(\frac{H^4 S C^{\star}_{\mathsf{clipped}}}{\varepsilon^2} \right). \tag{62}$$

  This confirms the optimality of the proposed model-based approach (up to some logarithmic term) when Bernstein-style penalty is employed, since Theorem 4 reveals that at least $\frac{H^4 S C^{\star}_{\mathsf{clipped}}}{\varepsilon^2}$ samples are needed regardless of the algorithm in use.

- *Full ε-range and no burn-in cost.* The sample complexity bound (59) stated in Theorem 3 holds for an arbitrary $\varepsilon \in (0, H]$. In other words, no burn-in cost is needed for the algorithm to work sample-optimally. This improves substantially upon the state-of-the-art results for model-based and model-free offline algorithms, both of which require a significant level of burn-in sample size ($H^9 SC^{\star}$ and $H^6 SC^{\star}$, respectively).

- *Sample reduction and model compressibility when* $C^\star_{\mathsf{clipped}} < 1$. Given that $C^\star_{\mathsf{clipped}}$ might drop below 1, the sample complexity of our algorithm might be as low as $\widetilde{O}\big(\frac{H^4 S}{\varepsilon^2}\big)$. In fact, recognizing that $C^\star_{\mathsf{clipped}}$ can be as small as $\frac{1+o(1)}{S}$, we see that the sample complexity can sometimes be reduced to

$$\widetilde{O}\bigg(\frac{H^4}{\varepsilon^2}\bigg), \tag{63}$$

resulting in significant sample size saving compared to prior works. Caution needs to be exercised, however, that this sample size improvement is made possible as a result of certain *model compressibility* implied by a small $C^\star_{\mathsf{clipped}}$. For instance, $C^\star_{\mathsf{clipped}} = O(1/S)$ might happen when a small number of states accounts for a dominant fraction of probability mass in $d^\star_h(s)$, with the remaining states exhibiting vanishingly small occupancy probability (see also the lower bound construction in the proof of Theorem 4); if this happens, then it often suffices to focus on learning those dominant states.

(In)-feasibility of estimating $C^\star_{\mathsf{clipped}}$. With the sample complexity (62) in mind, one natural question arises as to whether it is possible to estimate $C^\star_{\mathsf{clipped}}$ from the batch dataset. Unfortunately, this is in general infeasible, as demonstrated by the following example.

- (A hard example) Consider an MDP with horizon $H = 2$. In step $h = 1$, we have a singleton state space $\mathcal{S}_1 = \{0\}$ and an action space $\mathcal{A}_1 = \{0, 1\}$, whereas in step $h = 2$, we have a state space $\mathcal{S}_2 = \{0, 1\}$ and a singleton action space $\mathcal{A}_2 = \{0\}$. The reward function and the transition kernel are given by:

$$\begin{gathered} r_1(0,0) = 0, \quad r_1(0,1) = 0, \quad r_2(0,0) = 0, \quad r_2(1,0) = 1 \\ P_1(0\,|\,0,0) = 0.5, \quad P_1(1\,|\,0,0) = 0.5, \quad P_1(0\,|\,0,1) = p, \quad P_1(1\,|\,0,1) = 1-p \end{gathered}$$

for some unknown parameter $p \in (0, 1)$. We have K independent trajectories as usual, and let

$$d_1^{\mathsf{b}}(0,0) = 1 - \frac{1}{K} \qquad \text{and} \qquad d_1^{\mathsf{b}}(0,1) = \frac{1}{K}. \tag{64}$$

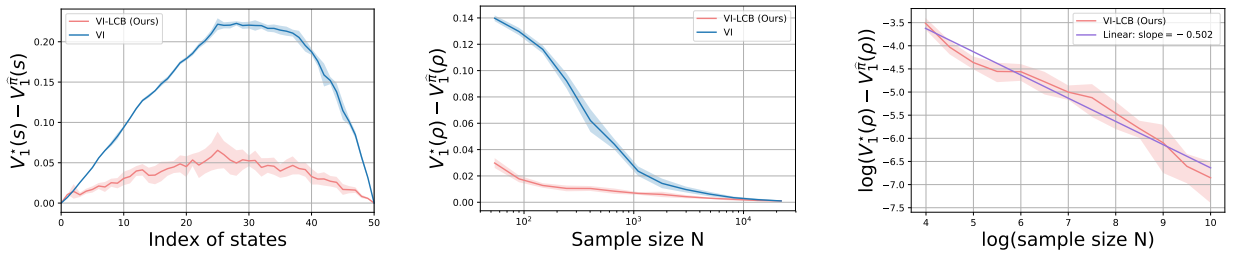
Elementary calculation then reveals that: $C^\star_{\mathsf{clipped}} = K$ when $p < \frac{1}{2}$, and $C^\star_{\mathsf{clipped}} = 1 + \frac{1}{K-1}$ when $p > \frac{1}{2}$. Such a remarkable difference in $C^\star_{\mathsf{clipped}}$ depends on the value of p, which is only reflected in $(s, a) = (0, 1)$ at step 1. However, by construction, there is nonvanishing probability (i.e., $\big(1 - d_1^{\mathsf{b}}(0,1)\big)^K \approx 1/e$ for large K) such that the dataset does not visit $(s, a) = (0, 1)$ in step $h = 1$ at all, which in turn precludes one from distinguishing $C^\star_{\mathsf{clipped}} = 1 + \frac{1}{K-1}$ from $C^\star_{\mathsf{clipped}} = K$ given only the available dataset.

Fortunately, implementing our algorithm does not require prior knowledge of $C^\star_{\mathsf{clipped}}$ at all, and the algorithm succeeds once the task becomes feasible. On the other hand, we won't be able to tell how large a sample size is enough *a priori*, but this is in general information-theoretically infeasible as illustrated by the above example.

Towards instance optimality. While the primary focus of the current paper is minimax-optimal algorithm design, the theoretical framework developed herein enables instance-dependent analysis as well. Take episodic finite-horizon MDPs for example: our analysis framework directly leads to the following instance-dependent guarantee for Algorithm 3:

$$\begin{aligned} V_h^\star(\rho) - V_h^{\widehat{\pi}}(\rho) &= \big\langle d_1^\star, V_1^\star - V_1^{\widehat{\pi}} \big\rangle \\ &\le 12 \sum_{j=h}^{H} \sum_s d_j^\star(s) \sqrt{\frac{c_{\mathsf{b}} \log \frac{NH}{\delta}}{K d_j^{\mathsf{b}}\big(s, \pi_j^\star(s)\big)} \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(V_{j+1}^\star)} + \bigg(\frac{100 c_{\mathsf{b}} H^3 S C^\star \log \frac{NH}{\delta}}{K}\bigg)^{3/4}, \end{aligned} \tag{65}$$

with the proviso that $K \ge 100 c_{\mathsf{b}} H S C^\star \log \frac{NH}{\delta}$. Encouragingly, the dominate term (i.e., the first term in the bound (65)) matches the instance-dependent lower bound established in (Yin and Wang, 2021, Theorem 4.3), thus confirming the instance optimality of the proposed algorithm for a large enough sample size. The proof of (65) can be found in Appendix B.2.



(a) Value gap versus states (b) Value gap versus sample size (c) Value gap dependency w.r.t. N

Figure 2: The performances of the proposed method VI-LCB and the baseline value iteration (VI) in the gambler's problem. It shows that VI-LCB outperforms VI by taking advantage of the pessimism principle and achieves approximately $1/\sqrt{N}$ sample complexity dependency w.r.t. the sample size N.

Comparisons with prior statistical analysis. We now briefly discuss the novelty of our statistical analysis compared with past theory. Perhaps the most related prior work is Xie et al. (2021), which proposed two algorithms. The first algorithm therein is VI-LCB with H-fold sample splitting and Hoeffding-style penalty, and each of these two features adds an H factor to the total sample complexity. The second algorithm therein combines VI-LCB with variance reduction, which leads to optimal sample complexity for sufficiently small ε (i.e., a large burn-in cost is required). Note, however, that none of the existing statistical tools for variance reduction is able to work without imposing a large burn-in cost, regardless of the sampling mechanism in use (e.g., generative model, offline RL, online RL) (Li et al., 2021; Sidford et al., 2018; Xie et al., 2021; Zhang et al., 2020). In contrast, our theory makes apparent that variance reduction is unnecessary, which leads to both simpler algorithm and tighter analysis. Additionally, while Bernstein-style confidence bounds have been deployed in online RL for finite-horizon MDPs (Azar et al., 2017; Fruit et al., 2020; Jin et al., 2018; Zhang et al., 2020), none of these works was able to yield optimal sample complexity without a large burn-in cost (e.g., Azar et al. (2017) incurred a burn-in cost as large as S^3AH^6). This in turn underscores the power of our statistical analysis when coping with the most data-hungry regime.

4 Numerical experiments

To confirm the practical applicability of the proposed VI-LCB algorithm, we evaluate its performance in the gambler's problem (Panaganti and Kalathil, 2022; Shi and Chi, 2022; Sutton and Barto, 2018; Zhou et al., 2021). The code can be accessed at:

https://github.com/Laixishi/Model-based-VI-LCB.

Gambler's problem. We start by introducing the formulation of the gambler's problem and its underlying MDP. An agent plays a gambling game in which she bets on a sequence of random coin flips, winning when the coins are heads and losing when they are tails. To bet on each random clip, the agent's policy chooses an integer number of dollars based on an initial balance. If the number of bets hits the maximum length H, or if the agent reaches 50 dollars (win) or 0 dollars (lose), the game ends. Without loss of generality, the problem can be formulated as an episodic finite-horizon MDP. Here, $\mathcal{S}$ is the state space $\{0, 1, \cdots, 50\}$ and the associated accessible actions obey $a \in \big\{0, 1, \cdots, \min\{s, 50-s\}\big\}$, $H = 100$ is the horizon length, the reward is set to 0 for all other states unless $s = 50$. For the transition kernel, we fix the probability of heads as $p_{\text{head}} = 0.45$ at all steps $h \in [H]$ in the episode. Moreover, the initial state/balance distribution of the agent ρ is taken as a uniform distribution over $\mathcal{S}$. The offline historical dataset is constructed by collecting N independent samples drawn randomly over each state-action pair and time step.

Evaluation results. First, we evaluate the performance of our proposed method VI-LCB (cf. Algorithm 2) with comparisons to the well-known value iteration (VI) method without the pessimism principle. To begin with, Fig. 2(a) shows the average and standard derivations of the performance gap $V_1^\star(s) - V_1^{\widehat{\pi}}(s)$ over

all states $s \in \mathcal{S}$, over 10 independent experiments with a fixed sample size $N = 50$. The results indicate that the proposed VI-LCB method outperforms the baseline VI method uniformly over the entire state space, showing that pessimism brings significant advantages in this sample-scarce regime. Secondly, we evaluate the performance gap $V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho)$ with varying sample size $N \in \{54, 90, 148, \cdots, 22026\} \approx \{e^4, e^{4.5}, e^5, \cdots, e^{10}\}$, over 10 independent trials. Note that throughout the experiments, we fix the parameter $c_{\mathsf{b}} = 0.05$, which determines the level of the pessimism penalty of VI-LCB (cf. (55)). Fig. 2(b) shows the average and standard derivations of the performance gap $V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho)$ with respect to the sample size N. Clearly, as the sample size increases, both our method VI-LCB and the baseline VI method perform better. Moreover, our VI-LCB method consistently outperforms the baseline VI method over the entire range of the sample size N, especially in the sample-starved regime. In addition, to corroborate the scaling of the sample size on the performance gap, we plot the sub-optimality performance gap of VI-LCB w.r.t. the sample size on a log-log scale in Fig. 2(c). Fitting using linear regression leads to a slope estimate of -0.502, with the corresponding fitted line plotted in Fig. 2(c) as well. This nicely matches the finding of Theorem 3, which says the performance gap of VI-LCB scales as $N^{-1/2}$.

5 Related works

In this section, we provide further discussions about prior art, with an emphasis on settings that are most relevant to the current paper.

Off-policy evaluation and offline RL. Broadly speaking, at least two families of problems have been investigated in the literature that tackle offline batch data: off-policy evaluation, where the goal is to estimate the value function of a target policy that deviates from the behavior policy used in data collection; and offline policy learning, where the goal is to identify a near-optimal policy (or at least an improved one compared to the behavior policy). Our work falls under the second category. A topic of its own interest, off-policy evaluation has been extensively studied in the recent literature; we excuse ourselves from enumerating the works in that space but only provide pointers to a few examples including Duan et al. (2020, 2021); Jiang and Huang (2020); Jiang and Li (2016); Kallus and Uehara (2020); Li et al. (2014); Ren et al. (2021); Thomas and Brunskill (2016); Uehara et al. (2020); Xu et al. (2021); Yang et al. (2020).

Offline RL with the pessimism principle. The prior works that are the most relevant to this paper are Jin et al. (2021); Rashidinejad et al. (2022); Shi et al. (2022b); Xie et al. (2021); Yan et al. (2023); Yin and Wang (2021), which incorporated lower confidence bounds into value estimation in order to avoid overly uncertain regions not covered by the target policy. In addition to the ones discussed in Section 1.2 that focus on minimax performance, the recent works Yin et al. (2022); Yin and Wang (2021) further developed instance-dependent statistical guarantees for the pessimistic model-based approach. The results in Yin and Wang (2021), however, required a large burn-in sample size $\frac{H^4}{SC^\star(d_{\min}^{\mathsf{b}})^2}$ (since $d_{\min}^{\mathsf{b}}$ could be exceedingly small), thus preventing it from attaining minimax optimality for the entire ε-range. It is noteworthy that the principle of pessimism has been incorporated into policy optimization and actor-critic methods as well by searching for some least-favorable models (e.g., Uehara and Sun (2021); Zanette et al. (2021)), which is quite different from the approach studied herein. On the empirical side, model-based algorithms (Kidambi et al., 2020; Yu et al., 2020) have been shown to achieve superior performance than their model-free counterpart for offline RL. In addition, a number of recent works studied offline RL under various function approximation assumptions, e.g., Jin et al. (2021); Nguyen-Tang et al. (2021); Uehara and Sun (2021); Uehara et al. (2022); Yin et al. (2022); Zanette et al. (2021); Zhan et al. (2022), which are beyond the scope of the current paper. Recently, the insights gleaned from the studies of offline RL have inspired improved algorithm designs for online and hybrid RL as well (Li et al., 2023, 2024c).

Online RL and the optimism principle. The optimism principle in the face of uncertainty has received widespread adoption from bandits to online RL (Agarwal et al., 2021; Lai and Robbins, 1985; Lattimore and Szepesvári, 2020). In the context of online RL, Jaksch et al. (2010) constructed confidence regions for the probability transition kernel to help select optimistic policies in the setting of weakly communicating MDPs, based on a variant (called UCRL2) of the UCRL algorithm originally proposed in Auer and Ortner

(2006); see also Bourel et al. (2020); Filippi et al. (2010); Talebi and Maillard (2018) for other variants of UCRL. When applied to episodic finite-horizon MDPs, the regret bound in Jaksch et al. (2010) was suboptimal by a factor of at least $\sqrt{H^2S}$; see discussion in Azar et al. (2017); Jin et al. (2018). Fruit et al. (2020) developed an improved regret bound for UCRL2 by using empirical Bernstein-style bounds, which however was still suboptimal by a factor of at least $\sqrt{HS}$ when specialized to episodic finite-horizon MDPs. In comparison, a more sample-efficient paradigm is to build Bernstein-style UCBs for the optimal values to help select exploration policies, which has been recently adopted in both model-based (Azar et al., 2017; Zhang et al., 2023) and model-free algorithms (Jin et al., 2018). Note that Bernstein-style uncertainty estimation alone is not enough to ensure regret optimality in model-free algorithms, thereby motivating the design of more sophisticated variance reduction strategies (Li et al., 2021; Zhang et al., 2020). Finally, the optimism principle has been studied in undiscounted infinite-horizon MDPs too (e.g., Qian et al. (2019)), which is beyond the scope of this paper.

Model-based RL. The algorithms studied herein fall under the category of model-based RL, which decouples the model estimation and the planning. This popular paradigm has been deployed and studied under various data collection mechanisms beyond offline RL, including but not limited to the generative model (or simulator) setting (Agarwal et al., 2020; Azar et al., 2013; Li et al., 2024b, 2020) and the online exploratory setting (Azar et al., 2017; Jin et al., 2020; Zhang et al., 2023, 2021a). The leave-one-out analysis (and the construction of absorbing MDPs) adopted in the proof of Theorem 1 has been inspired by several recent works Agarwal et al. (2020); Cui and Yang (2021); Li et al. (2024b); Pananjady and Wainwright (2020), and has recently been shown to be effective for multi-agent offline RL as well (Yan et al., 2022).

Model-free RL. Another widely used paradigm is model-free RL, which attempts to learn the optimal value function without explicit construction of the model. Arguably the most famous example of model-free RL is Q-learning, which applies the stochastic approximation paradigm to find the fixed point of the Bellman operator (Beck and Srikant, 2012; Chen et al., 2020; Even-Dar and Mansour, 2003; Li et al., 2024a, 2022; Murphy, 2005; Qu and Wierman, 2020; Shi et al., 2022a; Szepesvári, 1998; Watkins and Dayan, 1992; Xiong et al., 2020). It is worth noting that the asynchronous Q-learning, which aims to learn the optimal Q-function from a data trajectory collected by following a certain behavior policy, shares some similarity with offline RL; note that prior results on vanilla asynchronous Q-learning require a strong uniform coverage requirement (Chen et al., 2021b; Li et al., 2024a; Qu and Wierman, 2020), which is stronger than the single-policy concentrability considered herein. Moreover, Q-learning alone is known to be sub-optimal in terms of the sample complexity in various settings (Bai et al., 2019; Jin et al., 2018; Li et al., 2024a; Shi et al., 2022b; Wainwright, 2019b). This motivates the incorporation of the variance reduction in order to further improve the sample complexity (Du et al., 2017; Li et al., 2021, 2022; Shi et al., 2022b; Wainwright, 2019c; Yan et al., 2023; Zhang et al., 2020, 2021b). Note, however, variance-reduced model-free RL typically requires a large burn-in cost in order to operate in a sample-optimal fashion, and is hence outperformed by the model-based approach under multiple sampling mechanisms.

6 Analysis: discounted infinite-horizon MDPs

This section is devoted to establishing Theorem 1. Towards this end, we claim that it is sufficient to prove the following theorem.

Theorem 5. *Consider any $0 < \delta < 1$ and any $\gamma \in [\frac{1}{2}, 1)$. Suppose that the penalty terms are set to be (28) for any numerical constant $c_{\mathsf{b}} \geq 144$. Then with probability exceeding $1 - 2\delta$, for any estimate $\widehat{Q}$ obeying $\big\|\widehat{Q} - \widehat{Q}^{\star}_{\mathsf{pe}}\big\|_{\infty} \leq 1/N$ one has*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq 120\sqrt{\frac{c_{\mathsf{b}} S C^{\star}_{\mathsf{clipped}} \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3 N}} + \frac{3464 c_{\mathsf{b}} S C^{\star}_{\mathsf{clipped}} \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^2 N}, \tag{66}$$

where $\widehat{\pi}(s) \in \arg\max_a \widehat{Q}(s,a)$ for any $s \in \mathcal{S}$.

As we have demonstrated in Lemma 2, the output of Algorithm 1 satisfies $\big\|\widehat{Q}-\widehat{Q}^{\star}_{\mathsf{pe}}\big\|_{\infty}\leq 1/N$ once the iteration number exceeds $\tau_{\max}\geq \frac{\log\frac{N}{1-\gamma}}{\log(1/\gamma)}$, thus making Theorem 5 applicable. Taking the right-hand side of (66) to be no larger than ε reveals that: $V^{\star}(\rho)-V^{\widehat{\pi}}(\rho)\leq\varepsilon$ holds as long as N exceeds

$$N\geq \frac{21000c_{\mathsf{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3\varepsilon^2}, \tag{67}$$

given that $\varepsilon\in\big(0,\frac{1}{1-\gamma}\big]$.

The remainder of this section is thus dedicated to establishing Theorem 5. Throughout the proof, it suffices to focus on the case where

$$N\geq \frac{c_3SC^{\star}_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{1-\gamma} \tag{68}$$

for some large constant $c_3\geq 2880000$; otherwise the claim (66) follows directly since $V^{\star}(\rho)-V^{\widehat{\pi}}(\rho)\leq\frac{1}{1-\gamma}$.

6.1 Preliminary facts

Before embarking on the proof, we collect a couple of preliminary facts that will be used multiple times.

Properties of $N(s,a)$. To begin with, the quantity $N(s,a)$ — the total number of sample transitions from (s,a) — can be bounded through the following lemma; the proof is provided in Appendix A.3.

Lemma 4. *Consider any $\delta\in(0,1)$. With probability at least $1-\delta$, the quantities $\{N(s,a)\}$ in (24) obey*

$$\max\Big\{N(s,a),\frac{2}{3}\log\frac{N}{\delta}\Big\}\geq\frac{Nd^{\mathsf{b}}(s,a)}{12} \tag{69}$$

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

Properties about $\widehat{V}$ and $\widehat{V}^{\star}_{\mathsf{pe}}$. First of all, note that the assumption

$$\big\|\widehat{Q}-\widehat{Q}^{\star}_{\mathsf{pe}}\big\|_{\infty}\leq\frac{1}{N} \tag{70}$$

has the following direct consequence:

$$\big\|\widehat{V}-\widehat{V}^{\star}_{\mathsf{pe}}\big\|_{\infty}=\max_s\Big|\max_a\widehat{Q}(s,a)-\max_a\widehat{Q}^{\star}_{\mathsf{pe}}(s,a)\Big|\leq\big\|\widehat{Q}-\widehat{Q}^{\star}_{\mathsf{pe}}\big\|_{\infty}\leq\frac{1}{N}. \tag{71}$$

Given the proximity of $\widehat{V}$ and $\widehat{V}^{\star}_{\mathsf{pe}}$, we can bound the difference of the corresponding variance terms as follows:

$$\begin{aligned}
\Big|\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big)-\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)\Big| &\overset{\text{(i)}}{=}\Big|\widehat{P}_{s,a}\big(\widehat{V}^{\star}_{\mathsf{pe}}\circ\widehat{V}^{\star}_{\mathsf{pe}}\big)-\widehat{P}_{s,a}\big(\widehat{V}\circ\widehat{V}\big)+\big(\widehat{P}_{s,a}\widehat{V}\big)^2-\big(\widehat{P}_{s,a}\widehat{V}^{\star}_{\mathsf{pe}}\big)^2\Big|\\
&=\Big|\widehat{P}_{s,a}\Big(\big(\widehat{V}+\widehat{V}^{\star}_{\mathsf{pe}}\big)\circ\big(\widehat{V}^{\star}_{\mathsf{pe}}-\widehat{V}\big)\Big)+\Big(\widehat{P}_{s,a}\big(\widehat{V}+\widehat{V}^{\star}_{\mathsf{pe}}\big)\Big)\Big(\widehat{P}_{s,a}\big(\widehat{V}^{\star}_{\mathsf{pe}}-\widehat{V}\big)\Big)\Big|\\
&\leq\big\|\widehat{P}_{s,a}\big\|_1\big\|\widehat{V}+\widehat{V}^{\star}_{\mathsf{pe}}\big\|_{\infty}\big\|\widehat{V}^{\star}_{\mathsf{pe}}-\widehat{V}\big\|_{\infty}+\big\|\widehat{P}_{s,a}\big\|_1^2\big\|\widehat{V}+\widehat{V}^{\star}_{\mathsf{pe}}\big\|_{\infty}\big\|\widehat{V}^{\star}_{\mathsf{pe}}-\widehat{V}\big\|_{\infty}\\
&\leq\Big(\big\|\widehat{P}_{s,a}\big\|_1+\big\|\widehat{P}_{s,a}\big\|_1^2\Big)\Big(2\big\|\widehat{V}\big\|_{\infty}+\big\|\widehat{V}^{\star}_{\mathsf{pe}}-\widehat{V}\big\|_{\infty}\Big)\big\|\widehat{V}^{\star}_{\mathsf{pe}}-\widehat{V}\big\|_{\infty}\\
&\leq\frac{2}{N}\left(\frac{2}{1-\gamma}+\frac{1}{N}\right)\leq\frac{6}{(1-\gamma)N}.
\end{aligned} \tag{72}$$

Here, (i) follows from the definition (8), the penultimate inequality follows from (71) and the basic facts $\|\widehat{P}_{s,a}\|_1=1$ and $\|\widehat{V}\|_{\infty}\leq\frac{1}{1-\gamma}$, while the last line relies on (68).

Armed with (72), one can further control the difference of the associated penalty terms. Note that the definition of $b(s,a;V)$ in (28) tells us that

$$\begin{aligned}\left|b\big(s,a;\widehat{V}^{\star}_{\mathsf{pe}}\big)-b\big(s,a;\widehat{V}\big)\right| = \Bigg|&\min\left\{\max\left\{\sqrt{\frac{c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big)},\frac{2c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\},\frac{1}{1-\gamma}\right\}\\ &-\min\left\{\max\left\{\sqrt{\frac{c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)},\frac{2c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\},\frac{1}{1-\gamma}\right\}\Bigg|. \end{aligned} \tag{73}$$

If at least one of the variance terms is not too small in the sense that

$$\max\left\{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big),\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)\right\}\geq\frac{4c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2N(s,a)}, \tag{74}$$

then (73) implies that

$$\begin{aligned}(73)&\leq\sqrt{\frac{c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}}\left|\sqrt{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big)}-\sqrt{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)}\right|=\sqrt{\frac{c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}}\frac{\left|\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big)-\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)\right|}{\sqrt{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big)}+\sqrt{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)}}\\ &\overset{\text{(i)}}{\leq}\frac{1-\gamma}{2}\left|\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big)-\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)\right|\overset{\text{(ii)}}{\leq}\frac{3}{N}, \end{aligned} \tag{75}$$

where (i) results from (74), and (ii) holds due to (72). On the other hand, if (74) is not satisfied, then one clearly has $b\big(s,a;\widehat{V}^{\star}_{\mathsf{pe}}\big)=b\big(s,a;\widehat{V}\big)$. In conclusion, in all cases we have

$$\left|b\big(s,a;\widehat{V}^{\star}_{\mathsf{pe}}\big)-b\big(s,a;\widehat{V}\big)\right|\leq\frac{3}{N}. \tag{76}$$

6.2 Proof of Theorem 5

Armed with the preceding preliminary facts, we can readily turn to the proof of Theorem 5. By virtue of Lemma 4, our proof shall — unless otherwise noted — operate on the high-probability event that

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}:\qquad\max\left\{N(s,a),\frac{2}{3}\log\frac{SN}{\delta}\right\}\geq\frac{Nd^{\mathsf{b}}(s,a)}{12}. \tag{77}$$

In addition, from the sampling model (17), the sample transitions employed to form $\widehat{P}$ are statistically independent conditional on $\{N(s,a)\}$. Our proof consists of four steps as detailed below.

Step 1: Bernstein-style inequalities and leave-one-out decoupling argument. We are in need of tight control of the size of $\big(\widehat{P}_{s,a}-P_{s,a}\big)\widehat{V}$. However, this becomes challenging due to the statistical dependency between $\widehat{P}$ and the value estimate $\widehat{V}$ (given that we reuse samples in all iterations of Algorithm 1). In order to circumvent this difficulty, we resort to a leave-one-out argument to decouple the statistical dependency, as motivated by Agarwal et al. (2020); Li et al. (2024b). The result stated below establishes Bernstein-style inequalities despite the complicated dependency.

Lemma 5. *Suppose that $\gamma\in[\frac{1}{2},1)$, and consider any $\delta\in(0,1)$. With probability at least $1-\delta$, we have*

$$\left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\widetilde{V}\right|\leq12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widetilde{V}\big)}+\frac{74\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}, \tag{78a}$$

$$\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widetilde{V}\big)\leq2\mathsf{Var}_{P_{s,a}}\big(\widetilde{V}\big)+\frac{41\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)^2N(s,a)} \tag{78b}$$

simultaneously for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and all $\widetilde{V}$ with $\big\|\widetilde{V}-\widehat{V}^{\star}_{\mathsf{pe}}\big\|_{\infty}\leq\frac{1}{N}$ and $\|\widetilde{V}\|_{\infty}\leq\frac{1}{1-\gamma}$.

High-level proof ideas. In short, the proof consists of contructing a finite collection of auxiliary MDPs $\{\widehat{\mathcal{M}}^{s,u}\}$ for each state s obeying the following properties: (i) each $\widehat{\mathcal{M}}^{s,u}$ is constructed without using any sample transition that comes from state s, and is hence statistically independent from $\widehat{P}_{s,a}$ for all $a \in \mathcal{A}$ (instead, the useful information is embedded into the corresponding immediate reward, which is a low-dimensional object and easier to control); (ii) at least one of the MDPs in $\{\widehat{\mathcal{M}}^{s,u}\}$ is extremely close to the true MDP in terms of the resulting value function. With the aid of these leave-one-out auxiliary MDPs, one can control $\big(\widehat{P}_{s,a} - P_{s,a}\big)\widetilde{V}$ by first exploiting the statistical independence between $\widehat{P}_{s,a}$ and $\{\widehat{\mathcal{M}}^{s,u}\}$ and then transferring the concentration bound back to the original MDP using the proximity property (ii). The construction of these auxiliary MDPs and the proof details can be found in Appendix A.4. □

Note that (78a) has been derived only for those pairs (s,a) with $N(s,a) > 0$. For every (s,a) with $N(s,a) = 0$, one can directly obtain

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widetilde{V}\Big| = \big|P_{s,a}\widetilde{V}\big| \le \big\|P_{s,a}\big\|_1 \big\|\widetilde{V}\big\|_\infty \le \frac{1}{1-\gamma}.$$

Putting these bounds together with the definition (28) of $b(s,a;V)$ reveals that

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widetilde{V}\Big| + \frac{5}{N} \le b\big(s,a;\widetilde{V}\big) \qquad \text{for all } (s,a) \in \mathcal{S}\times\mathcal{A} \tag{79}$$

for all $\widetilde{V}$ obeying $\big\|\widetilde{V} - \widehat{V}^\star_{\mathsf{pe}}\big\|_\infty \le \frac{1}{N}$ and $\|\widetilde{V}\|_\infty \le \frac{1}{1-\gamma}$, provided that the constant c_{b} is sufficiently large. The remainder of the proof should then also operate on the high-probability events (79) and (78b), in addition to assuming that the event (77) occurs.

Step 2: showing that $\widehat{Q}(s,a)$ is a lower bound on $Q^{\widehat{\pi}}(s,a)$. We now justify that $\widehat{Q}(s,a)$ (resp. $\widehat{V}(s)$) is a "pessimistic" estimate of $Q^{\widehat{\pi}}(s,a)$ (resp. $V^{\widehat{\pi}}(s)$); this is enabled by the pessimism principle (so that the algorithm effectively seeks lower estimates of the value iteration) and the Bernstein-style bounds in Lemma 5 (so that the penalty term always dominates the uncertainty incurred by using the empirical MDP).

To begin with, recall that $\widehat{Q}^\star_{\mathsf{pe}}(s,a)$ is the unique fixed point of the pessimistic Bellman operator that obeys

$$\widehat{Q}^\star_{\mathsf{pe}}(s,a) = \max\Big\{r(s,a) + \gamma\widehat{P}_{s,a}\widehat{V}^\star_{\mathsf{pe}} - b\big(s,a;\widehat{V}^\star_{\mathsf{pe}}\big), 0\Big\}. \tag{80}$$

In the sequel, we divide the set of state-action pairs (s,a) into two types.

- *Case 1:* $\widehat{Q}^\star_{\mathsf{pe}}(s,a) = 0$. Given that $\widehat{Q}_0 = 0$, Lemma 2 tells us that

  $$\widehat{Q}(s,a) = \widehat{Q}_{\tau_{\max}}(s,a) \le \widehat{Q}^\star_{\mathsf{pe}}(s,a) = 0.$$

  This combined with the basic fact $Q^{\widehat{\pi}} \ge 0$ immediately yields $0 = \widehat{Q}(s,a) \le Q^{\widehat{\pi}}(s,a)$.

- *Case 2:* $\widehat{Q}^\star_{\mathsf{pe}}(s,a) = r(s,a) + \gamma\widehat{P}_{s,a}\widehat{V}^\star_{\mathsf{pe}} - b\big(s,a;\widehat{V}^\star_{\mathsf{pe}}\big) > 0$. It is first observed that

$$\begin{aligned}
\widehat{Q}(s,a) &\overset{\text{(i)}}{\le} \widehat{Q}^\star_{\mathsf{pe}}(s,a) + \frac{1}{N} \overset{\text{(ii)}}{=} r(s,a) - b\big(s,a;\widehat{V}^\star_{\mathsf{pe}}\big) + \gamma\widehat{P}_{s,a}\widehat{V}^\star_{\mathsf{pe}} + \frac{1}{N} \\
&\le r(s,a) - b\big(s,a;\widehat{V}^\star_{\mathsf{pe}}\big) + \gamma\widehat{P}_{s,a}\widehat{V} + \frac{1}{N} + \gamma\big\|\widehat{P}_{s,a}\big\|_1\big\|\widehat{V} - \widehat{V}^\star_{\mathsf{pe}}\big\|_\infty \\
&\overset{\text{(iii)}}{\le} r(s,a) - b\big(s,a;\widehat{V}^\star_{\mathsf{pe}}\big) + \gamma\widehat{P}_{s,a}\widehat{V} + \frac{2}{N} \\
&\le r(s,a) - b\big(s,a;\widehat{V}\big) + \gamma P_{s,a}\widehat{V} + \frac{2}{N} + \gamma\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widehat{V}\Big| + \Big|b\big(s,a;\widehat{V}^\star_{\mathsf{pe}}\big) - b\big(s,a;\widehat{V}\big)\Big| \\
&\overset{\text{(iv)}}{\le} r(s,a) + \gamma P_{s,a}\widehat{V}.
\end{aligned} \tag{81}$$

Here, (i) and (iii) arise from the assumption (70), (ii) relies on the fact that $\widehat{Q}^{\star}_{\mathsf{pe}}$ is the fixed point of the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}$, whereas (iv) takes advantage of (76) and (79). Combining (81) with the Bellman equation $Q^{\widehat{\pi}} = r + \gamma P V^{\widehat{\pi}}$ results in

$$Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a) \geq r(s,a) + \gamma P_{s,a} V^{\widehat{\pi}} - \big(r(s,a) + \gamma P_{s,a}\widehat{V}\big) = \gamma P_{s,a}\big(V^{\widehat{\pi}} - \widehat{V}\big). \tag{82}$$

Suppose for the moment that there exists some (s,a) obeying $Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a) < 0$ (which clearly cannot happen in Case 1), then $\arg\min_{s,a}\big[Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a)\big]$ must belong to Case 2. Thus, taking the minimum over (s,a) and using the above inequality (82) give

$$\begin{aligned}\min_{s,a}\big[Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a)\big] &\geq \min_{s,a}\big[\gamma P_{s,a}\big(V^{\widehat{\pi}} - \widehat{V}\big)\big] \overset{\text{(i)}}{\geq} \gamma \min_{s}\big[V^{\widehat{\pi}}(s) - \widehat{V}(s)\big] \\ &= \gamma \min_{s}\big[Q^{\widehat{\pi}}\big(s,\widehat{\pi}(s)\big) - \widehat{Q}\big(s,\widehat{\pi}(s)\big)\big] \geq \gamma \min_{s,a}\big[Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a)\big],\end{aligned} \tag{83}$$

where (i) holds since $P_{s,a} \in \Delta(\mathcal{S})$. Given that $1 > \gamma > 0$, inequality (83) holds only when $\min_{s,a}\big[Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a)\big] \geq 0$. We therefore conclude that in this case, one also has $Q^{\widehat{\pi}}(s,a) \geq \widehat{Q}(s,a)$.

With the arguments for the above two cases in place, we arrive at

$$Q^{\widehat{\pi}}(s,a) \geq \widehat{Q}(s,a) \qquad \text{for all } (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{84}$$

and evidently,

$$V^{\star}(s) \geq V^{\widehat{\pi}}(s) = Q^{\widehat{\pi}}\big(s,\widehat{\pi}(s)\big) \geq \widehat{Q}\big(s,\widehat{\pi}(s)\big) = \max_{a}\widehat{Q}(s,a) = \widehat{V}(s) \qquad \text{for all } s \in \mathcal{S}. \tag{85}$$

Step 3: bounding $V^{\star}(s) - V^{\widehat{\pi}}(s)$. Recall that the Bellman optimality equation gives

$$V^{\star}(s) = r\big(s,\pi^{\star}(s)\big) + \gamma P_{s,\pi^{\star}(s)} V^{\star}. \tag{86}$$

Before continuing, we make note of the following lower bound on $\widehat{V}$:

$$\begin{aligned}\widehat{V}(s) = \max_{a}\widehat{Q}(s,a) \geq \widehat{Q}\big(s,\pi^{\star}(s)\big) &\overset{\text{(i)}}{\geq} \widehat{Q}^{\star}_{\mathsf{pe}}\big(s,\pi^{\star}(s)\big) - \frac{1}{N} \\ &\overset{\text{(ii)}}{\geq} r\big(s,\pi^{\star}(s)\big) - b\big(s,\pi^{\star}(s);\widehat{V}^{\star}_{\mathsf{pe}}\big) + \gamma\widehat{P}_{s,\pi^{\star}(s)}\widehat{V}^{\star}_{\mathsf{pe}} - \frac{1}{N} \\ &= r\big(s,\pi^{\star}(s)\big) - b\big(s,\pi^{\star}(s);\widehat{V}^{\star}_{\mathsf{pe}}\big) + \gamma\widehat{P}_{s,\pi^{\star}(s)}\widehat{V} - \frac{1}{N} - \gamma\widehat{P}_{s,\pi^{\star}(s)}\big(\widehat{V} - \widehat{V}^{\star}_{\mathsf{pe}}\big) \\ &\overset{\text{(iii)}}{\geq} r\big(s,\pi^{\star}(s)\big) - b\big(s,\pi^{\star}(s);\widehat{V}^{\star}_{\mathsf{pe}}\big) + \gamma\widehat{P}_{s,\pi^{\star}(s)}\widehat{V} - \frac{2}{N} \\ &\geq r\big(s,\pi^{\star}(s)\big) - b\big(s,\pi^{\star}(s);\widehat{V}\big) + \gamma P_{s,\pi^{\star}(s)}\widehat{V} - \frac{2}{N} - \gamma\Big|\big(\widehat{P}_{s,\pi^{\star}(s)} - P_{s,\pi^{\star}(s)}\big)\widehat{V}\Big| \\ &\qquad - \Big|b\big(s,\pi^{\star}(s);\widehat{V}^{\star}_{\mathsf{pe}}\big) - b\big(s,\pi^{\star}(s);\widehat{V}\big)\Big| \\ &\overset{\text{(iv)}}{\geq} r\big(s,\pi^{\star}(s)\big) - 2b\big(s,\pi^{\star}(s);\widehat{V}\big) + \gamma P_{s,\pi^{\star}(s)}\widehat{V}.\end{aligned} \tag{87}$$

Here, (i) results from the assumption (70), (ii) relies on (80), (iii) is valid since $\widehat{P}_{s,\pi^{\star}(s)}\big(\widehat{V} - \widehat{V}^{\star}_{\mathsf{pe}}\big) \leq \big\|\widehat{P}_{s,\pi^{\star}(s)}\big\|_1\big\|\widehat{V} - \widehat{V}^{\star}_{\mathsf{pe}}\big\|_{\infty} \leq 1/N$, whereas (iv) holds by virtue of (76) and (79). Armed with the results in (86) and (87), we can readily show that

$$\begin{aligned}\big\langle \rho, V^{\star} - \widehat{V}\big\rangle &= \sum_{s\in\mathcal{S}}\rho(s)\big(V^{\star}(s) - \widehat{V}(s)\big) \\ &\leq \sum_{s\in\mathcal{S}}\rho(s)\Big\{r\big(s,\pi^{\star}(s)\big) + \gamma P_{s,\pi^{\star}(s)}V^{\star} - \Big(r\big(s,\pi^{\star}(s)\big) - 2b\big(s,\pi^{\star}(s);\widehat{V}\big) + \gamma P_{s,\pi^{\star}(s)}\widehat{V}\Big)\Big\}\end{aligned}$$

$$\leq \gamma \sum_{s\in\mathcal{S}} \rho(s) P_{s,\pi^\star(s)}\big(V^\star - \widehat{V}\big) + 2\sum_{s\in\mathcal{S}} \rho(s) b\big(s,\pi^\star(s);\widehat{V}\big). \tag{88}$$

For notational convenience, let us introduce a matrix $P^\star \in \mathbb{R}^{S\times S}$ and a vector $b^\star \in \mathbb{R}^{S\times 1}$ whose s-th row are given respectively by

$$\big[P^\star\big]_{s,\cdot} \coloneqq P_{s,\pi^\star(s)} \qquad \text{and} \qquad b^\star(s) \coloneqq b\big(s,\pi^\star(s);\widehat{V}\big) \qquad \text{for all } s\in\mathcal{S}. \tag{89}$$

This allows us to rewrite (88) in the following matrix/vector form:

$$\rho^\top\big(V^\star - \widehat{V}\big) \leq \gamma\rho^\top P^\star\big(V^\star - \widehat{V}\big) + 2\rho^\top b^\star. \tag{90}$$

Note that this relation holds for any arbitrary ρ. Apply it recursively to arrive at

$$\begin{aligned}
\rho^\top\big(V^\star - \widehat{V}\big) &\leq \big(\gamma\rho^\top P^\star\big)\big(V^\star - \widehat{V}\big) + 2\rho^\top b^\star \\
&\leq \gamma\big(\gamma\rho^\top P^\star\big)P^\star\big(V^\star - \widehat{V}\big) + 2\big(\gamma\rho^\top P^\star\big)b^\star + 2\rho^\top b^\star \\
&= \gamma^2\rho^\top\big(P^\star\big)^2\big(V^\star - \widehat{V}\big) + 2\gamma\rho^\top P^\star b^\star + 2\rho^\top b^\star \\
&\leq \cdots \leq \Big\{\lim_{i\to\infty}\gamma^i\rho^\top\big(P^\star\big)^i\big(V^\star - \widehat{V}\big)\Big\} + 2\rho^\top\Big\{\sum_{i=0}^{\infty}\gamma^i\big(P^\star\big)^i\Big\}b^\star \\
&\overset{\text{(i)}}{=} 2\rho^\top\Big\{\sum_{i=0}^{\infty}\gamma^i\big(P^\star\big)^i\Big\}b^\star = 2\rho^\top\big(I-\gamma P^\star\big)^{-1}b^\star \\
&= \frac{2}{1-\gamma}\langle d^\star, b^\star\rangle,
\end{aligned} \tag{91}$$

where (i) holds since $\lim_{i\to\infty}\gamma^i\rho^\top\big(P^\star\big)^i\big(V^\star - \widehat{V}\big) = 0$ (given that $\lim_{i\to\infty}\gamma^i = 0$ and $\|\rho^\top\big(P^\star\big)^i\|_1 = 1$ for any $i\geq 0$), and the last equality results from the definition of $d^\star$ (see (16)) expressed in the following matrix/vector form:

$$\big(d^\star\big)^\top = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\rho^\top\big(P^\star\big)^t = (1-\gamma)\rho^\top\big(I-\gamma P^\star\big)^{-1}. \tag{92}$$

Combine the above inequality with (85) to reach

$$\big\langle\rho, V^\star - V^{\widehat{\pi}}\big\rangle \leq \big\langle\rho, V^\star - \widehat{V}\big\rangle \leq \frac{2\big\langle d^\star, b^\star\big\rangle}{1-\gamma}. \tag{93}$$

Step 4: using concentrability to control $\big\langle d^\star, b^\star\big\rangle$. We shall control $\big\langle d^\star, b^\star\big\rangle$ by dividing the state set $\mathcal{S}$ into the following two disjoint subsets:

$$\mathcal{S}^{\mathsf{small}} \coloneqq \left\{s\in\mathcal{S} \mid Nd^{\mathsf{b}}\big(s,\pi^\star(s)\big) \leq 8\log\frac{NS}{(1-\gamma)\delta}\right\}; \tag{94a}$$

$$\mathcal{S}^{\mathsf{large}} \coloneqq \left\{s\in\mathcal{S} \mid Nd^{\mathsf{b}}\big(s,\pi^\star(s)\big) > 8\log\frac{NS}{(1-\gamma)\delta}\right\}. \tag{94b}$$

- To begin with, consider any state $s\in\mathcal{S}^{\mathsf{small}}$. Applying Definition 2 and the definition of $\mathcal{S}^{\mathsf{small}}$ yields

$$\min\left\{d^\star(s), \frac{1}{S}\right\} \leq C^\star_{\mathsf{clipped}}d^{\mathsf{b}}\big(s,\pi^\star(s)\big) \leq \frac{8C^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{N} < \frac{1}{S}, \tag{95}$$

  provided that $N > 8SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}$ (see (68)). This inequality necessarily implies that

$$d^\star(s) \leq \frac{8C^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{N} < \frac{1}{S}. \tag{96}$$

Combining the preceding inequality with the following fact (see the definition (28))

$$b^\star(s) \coloneqq b\big(s, \pi^\star(s); \widehat{V}\big) \le \frac{1}{1-\gamma} + \frac{5}{N}, \tag{97}$$

we arrive at

$$\sum_{s\in\mathcal{S}^{\mathsf{small}}} d^\star(s) b^\star(s) \le \sum_{s\in\mathcal{S}^{\mathsf{small}}} \left(\frac{8C^\star_{\mathsf{clipped}} \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + d^\star(s)\frac{5}{N} \right) \le \frac{8SC^\star_{\mathsf{clipped}} \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + \frac{5}{N}. \tag{98}$$

- Next, we turn to any state $s \in \mathcal{S}^{\mathsf{large}}$. Using the definition (28) of $b(s,a;V)$, we obtain

$$\begin{aligned} b^\star(s) = b\big(s,\pi^\star(s);\widehat{V}\big) &\le \sqrt{\frac{c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N\big(s,\pi^\star(s)\big)} \mathsf{Var}_{\widehat{P}_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \frac{2c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N\big(s,\pi^\star(s)\big)} + \frac{5}{N} \\ &\overset{\text{(i)}}{\le} \sqrt{\frac{c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N\big(s,\pi^\star(s)\big)} \left(2\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) + \frac{41 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)^2 N\big(s,\pi^\star(s)\big)} \right)} + \frac{2c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N\big(s,\pi^\star(s)\big)} + \frac{5}{N} \\ &\overset{\text{(ii)}}{\le} \sqrt{\frac{2c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N\big(s,\pi^\star(s)\big)} \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) + \frac{4c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N\big(s,\pi^\star(s)\big)}}, \end{aligned} \tag{99}$$

where (i) arises from Lemma 5 and (71), (ii) applies the elementary inequality $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for any $x, y \ge 0$ and the fact $N \ge N(s,a)$, in addition to assuming that c_{b} is large enough. To continue, we observe that

$$\frac{1}{N\big(s,\pi^\star(s)\big)} \overset{\text{(i)}}{\le} \frac{12}{Nd^{\mathsf{b}}\big(s,\pi^\star(s)\big)} \overset{\text{(ii)}}{\le} \frac{12C^\star_{\mathsf{clipped}}}{N \min\big\{d^\star(s), \frac{1}{S}\big\}} \le \frac{12C^\star_{\mathsf{clipped}}}{N}\left(\frac{1}{d^\star(s)} + S\right), \tag{100}$$

where (i) follows from the assumption (77) and the definition of $\mathcal{S}^{\mathsf{large}}$, and (ii) results from Assumption 2. Substitution into (99) yields

$$b^\star(s) \le \underbrace{\sqrt{\frac{24c_{\mathsf{b}}C^\star_{\mathsf{clipped}} \log \frac{N}{(1-\gamma)\delta}}{N} \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} \left(\frac{1}{\sqrt{d^\star(s)}} + \sqrt{S}\right)}_{=:\alpha_1(s)} + \underbrace{\frac{48c_{\mathsf{b}}C^\star_{\mathsf{clipped}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}\left(\frac{1}{d^\star(s)} + S\right)}_{=:\alpha_2(s)}, \tag{101}$$

where the last line comes from the elementary inequality $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for any $x, y \ge 0$.

To proceed, observe that the sum of the first terms in (101) satisfies

$$\begin{aligned} &\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\alpha_1(s) \\ &= \sqrt{\frac{24c_{\mathsf{b}}C^\star_{\mathsf{clipped}} \log \frac{N}{(1-\gamma)\delta}}{N}} \left(\sum_{s\in\mathcal{S}^{\mathsf{large}}} \sqrt{d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \sum_{s\in\mathcal{S}^{\mathsf{large}}} \sqrt{d^\star(s)}\sqrt{Sd^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} \right) \\ &\overset{\text{(i)}}{\le} \sqrt{\frac{24c_{\mathsf{b}}C^\star_{\mathsf{clipped}} \log \frac{N}{(1-\gamma)\delta}}{N}} \left(\sqrt{S} \cdot \sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} Sd^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} \right) \\ &= \sqrt{\frac{96c_{\mathsf{b}}SC^\star_{\mathsf{clipped}} \log \frac{N}{(1-\gamma)\delta}}{N}} \sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)}, \end{aligned} \tag{102}$$

where (i) arises from the Cauchy-Schwarz inequality and the fact $\sum_s d^\star(s) = 1$. In addition, it is easily verified that the sum of the second terms in (101) obeys

$$\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\alpha_2(s) \le \frac{96c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}, \tag{103}$$

which also makes use of the identity $\sum_s d^\star(s) = 1$. Combining (102) and (103) with (101) gives

$$\begin{aligned}\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)b^\star\big(s,\pi^\star(s)\big) &\le \sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\alpha_1(s) + \sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\alpha_2(s)\\ &\le \sqrt{\frac{96c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{N}{(1-\gamma)\delta}}{N}}\sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \frac{96c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}.\end{aligned} \tag{104}$$

The above results (98) and (104) taken collectively give

$$\begin{aligned}\big\langle d^\star, b^\star\big\rangle &= \sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)b^\star(s) + \sum_{s\in\mathcal{S}^{\mathsf{small}}} d^\star(s)b^\star(s)\\
&\le \sqrt{\frac{96c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{N}{(1-\gamma)\delta}}{N}}\sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \frac{96c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}\\
&\quad + \frac{8SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + \frac{5}{N}\\
&\overset{\text{(i)}}{\le} \sqrt{\frac{96c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{N}}\sqrt{\sum_{s\in\mathcal{S}} d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \frac{98c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}\\
&\overset{\text{(ii)}}{\le} \frac{2}{\gamma}\sqrt{\frac{96c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}\big\langle d^\star, b^\star\big\rangle} + \frac{1}{\gamma}\sqrt{\frac{192c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}} + \frac{98c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N},\\
&\overset{\text{(iii)}}{\le} 4\sqrt{\frac{96c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}\big\langle d^\star, b^\star\big\rangle} + 2\sqrt{\frac{192c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}} + \frac{98c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N},\\
&\overset{\text{(iv)}}{\le} \frac{1}{2}\big\langle d^\star, b^\star\big\rangle + \frac{768c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + \sqrt{\frac{768c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}} + \frac{98c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}.\end{aligned}$$

Here, (i) follows when c_{b} is sufficiently large and $C^\star_{\mathsf{clipped}} \ge 1/S$ (see (21)), (ii) would hold as long as the following inequality could be established:

$$\sum_{s\in\mathcal{S}} d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) \le \frac{2}{\gamma^2(1-\gamma)} + \frac{4}{\gamma^2(1-\gamma)}\big\langle d^\star, b^\star\big\rangle; \tag{105}$$

(iii) is valid since $\gamma \in [\frac{1}{2}, 1)$, and (iv) follows from the elementary inequality $2xy \le x^2 + y^2$. Rearranging terms, we are left with

$$\big\langle d^\star, b^\star\big\rangle \le \sqrt{\frac{3072c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}} + \frac{1732c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}, \tag{106}$$

which combined with (93) yields

$$\big\langle \rho, V^\star - V^{\widehat{\pi}}\big\rangle \le \frac{2\big\langle d^\star, b^\star\big\rangle}{1-\gamma} \le 120\sqrt{\frac{c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3N}} + \frac{3464c_{\mathsf{b}}SC^\star_{\mathsf{clipped}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^2N}. \tag{107}$$

This concludes the proof, as long as the inequality (105) can be established.

Proof of inequality (105). To begin with, we make the observation that

$$\begin{aligned}\big(\widehat{V}\circ\widehat{V}\big)-\big(\gamma P^\star\widehat{V}\big)\circ\big(\gamma P^\star\widehat{V}\big) &= \big(\widehat{V}-\gamma P^\star\widehat{V}\big)\circ\big(\widehat{V}+\gamma P^\star\widehat{V}\big)\\ &\overset{\text{(i)}}{\le}\big(\widehat{V}-\gamma P^\star\widehat{V}+2b^\star\big)\circ\big(\widehat{V}+\gamma P^\star\widehat{V}\big)\\ &\overset{\text{(ii)}}{\le}\frac{2}{1-\gamma}\big(\widehat{V}-\gamma P^\star\widehat{V}+2b^\star\big),\end{aligned}\tag{108}$$

where (i) holds since $b^\star\ge 0$ and $\widehat{V}+\gamma P^\star\widehat{V}\ge 0$, (ii) follows from the basic property $\|\widehat{V}+\gamma P^\star\widehat{V}\|_\infty\le 2\|\widehat{V}\|_\infty\le\frac{2}{1-\gamma}$ and the fact $\widehat{V}-\gamma P^\star\widehat{V}+2b^\star\ge 0$, the latter of which has been verified in (87). Armed with this fact, one can deduce that

$$\begin{aligned}\sum_s d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) &\overset{\text{(i)}}{=}\Big\langle d^\star, P^\star\big(\widehat{V}\circ\widehat{V}\big)-\big(P^\star\widehat{V}\big)\circ\big(P^\star\widehat{V}\big)\Big\rangle\\ &\overset{\text{(ii)}}{\le}\Big\langle d^\star, P^\star\big(\widehat{V}\circ\widehat{V}\big)-\frac{1}{\gamma^2}\widehat{V}\circ\widehat{V}+\frac{2}{\gamma^2(1-\gamma)}\big(\widehat{V}-\gamma P^\star\widehat{V}+2b^\star\big)\Big\rangle\\ &\overset{\text{(iii)}}{\le}\Big\langle d^\star, P^\star\big(\widehat{V}\circ\widehat{V}\big)-\frac{1}{\gamma}\widehat{V}\circ\widehat{V}+\frac{2}{\gamma^2(1-\gamma)}(I-\gamma P^\star)\widehat{V}+\frac{4}{\gamma^2(1-\gamma)}b^\star\Big\rangle\\ &=\Big\langle d^\star, \frac{1}{\gamma}\big(\gamma P^\star-I\big)\big(\widehat{V}\circ\widehat{V}\big)+\frac{2}{\gamma^2(1-\gamma)}(I-\gamma P^\star)\widehat{V}+\frac{4}{\gamma^2(1-\gamma)}b^\star\Big\rangle\\ &=d^{\star\top}\big(I-\gamma P^\star\big)\Big\{-\frac{1}{\gamma}\widehat{V}\circ\widehat{V}+\frac{2}{\gamma^2(1-\gamma)}\widehat{V}\Big\}+\frac{4}{\gamma^2(1-\gamma)}\langle d^\star,b^\star\rangle\\ &\overset{\text{(iv)}}{\le}(1-\gamma)\rho^\top\Big\{-\frac{1}{\gamma}\widehat{V}\circ\widehat{V}+\frac{2}{\gamma^2(1-\gamma)}\widehat{V}\Big\}+\frac{4}{\gamma^2(1-\gamma)}\langle d^\star,b^\star\rangle\\ &\le\frac{2}{\gamma^2}\rho^\top\widehat{V}+\frac{4}{\gamma^2(1-\gamma)}\langle d^\star,b^\star\rangle\\ &\overset{\text{(v)}}{\le}\frac{2}{\gamma^2(1-\gamma)}+\frac{4}{\gamma^2(1-\gamma)}\langle d^\star,b^\star\rangle.\end{aligned}$$

Here, (i) follows by invoking the definition (8), (ii) holds due to (108), (iii) is valid since $\gamma<1$, (iv) is a direct consequence of (92), while (v) comes from the basic facts $\|\rho^\top\|_1=1$ and $\|\widehat{V}\|_\infty\le\frac{1}{1-\gamma}$.

7 Analysis: episodic finite-horizon MDPs

7.1 Preliminary facts and notation

We first collect a few preliminary facts that are useful for the analysis. The first fact determines the range of our estimates $\widehat{Q}_h$ and $\widehat{V}_h$.

Lemma 6. *The iterates of Algorithm 2 obey*

$$0\le\widehat{Q}_h(s,a)\le H-h+1\quad\text{and}\quad 0\le\widehat{V}_h(s)\le H-h+1\qquad\text{for all }(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H].\tag{109}$$

Proof. The non-negativity of $\widehat{Q}_h$ (and hence $\widehat{V}_h$) follows directly from the update rule (52). Regarding the upper bound, we suppose for the moment that $\widehat{V}_{h+1}(s)\le H-h$ for step $h+1$. Then (52) tells us that

$$\widehat{Q}_h(s,a)\le 1+\big\|\widehat{V}_{h+1}\big\|_\infty\le 1+H-h,$$

which together with $\widehat{V}_h(s)=\max_a\widehat{Q}_h(s,a)$ justifies the claim (109) for step h as well. Taking this together with the base case $\widehat{V}_{H+1}=0$ and the standard induction argument concludes the proof. □

The second fact is concerned with the vector $d_h^\star := [d_h^\star(s)]_{s\in\mathcal{S}} \in \mathbb{R}^S$. For any $h \in [H]$, denote by $P_h^\star \in \mathbb{R}^{S\times S}$ a matrix whose s-th row is given by $P_h\big(\cdot \,|\, s, \pi_h^\star(s)\big)$. Then the Markovian property of the MDP indicates that: for any $j > h$, one has

$$\big(d_j^\star\big)^\top = \big(d_h^\star\big)^\top P_h^\star \cdots P_{j-1}^\star. \tag{110}$$

Notation. We remind the reader that $P_{h,s,a} \in \mathbb{R}^{1\times S}$ represents the probability transition vector $P_h(\cdot \,|\, s, a)$, and the associated variance parameter $\mathsf{Var}_{P_{h,s,a}}(V)$ is defined to be the (h,s,a)-th row of $\mathsf{Var}_P(V)$ (cf. (8)), namely,

$$\mathsf{Var}_{P_{h,s,a}}(V) := \sum_{s'\in S} P_h(s' \,|\, s, a)\big(V(s')\big)^2 - \bigg(\sum_{s'\in S} P_h(s' \,|\, s, a)V(s')\bigg)^2 \tag{111}$$

for any given vector $V \in \mathbb{R}^S$. The vector $\widehat{P}_{h,s,a} \in \mathbb{R}^{1\times S}$ and the variance parameter $\mathsf{Var}_{\widehat{P}_{h,s,a}}(V)$ are defined analogously.

7.2 A crucial statistical independence property

This subsection demonstrates that the subsampling trick introduced in Section 3.3 leads to some crucial statistical independence property. To be precise, let us consider the following two data-generating mechanisms; here and below, a sample transition refers to a quadruple (s, a, h, s') that indicates a transition from state s to state s' when action a is taken at step h.

- **Model 1 (augmented dataset).** Augment $\mathcal{D}^{\mathsf{trim}}$ to yield a dataset $\mathcal{D}^{\mathsf{trim,aug}}$ via the following steps. For every $(s,h) \in \mathcal{S}\times[H]$:
  1) Add to $\mathcal{D}^{\mathsf{trim,aug}}$ all $N_h^{\mathsf{trim}}(s)$ sample transitions in $\mathcal{D}^{\mathsf{trim}}$ that transition from state s at step h;
  2) If $N_h^{\mathsf{trim}}(s) > N_h^{\mathsf{main}}(s)$, then we further add to $\mathcal{D}^{\mathsf{trim,aug}}$ another set of $N_h^{\mathsf{trim}}(s) - N_h^{\mathsf{main}}(s)$ independent sample transitions $\big\{\big(s, a_{h,s}^{(i)}, h, s_{h,s}^{\prime(i)}\big)\big\}$ obeying

$$a_{h,s}^{(i)} \overset{\text{i.i.d.}}{\sim} \pi_h^{\mathsf{b}}(\cdot \,|\, s), \qquad s_{h,s}^{\prime(i)} \overset{\text{i.i.d.}}{\sim} P_h\big(\cdot \,|\, s, a_{h,s}^{(i)}\big), \qquad N_h^{\mathsf{main}}(s) < i \le N_h^{\mathsf{trim}}(s). \tag{112}$$

- **Model 2 (independent dataset).** For every $(s,h) \in \mathcal{S}\times[H]$, generate $N_h^{\mathsf{trim}}(s)$ independent sample transitions $\big\{\big(s, a_{h,s}^{(i)}, h, s_{h,s}^{\prime(i)}\big)\big\}$ as follows:

$$a_{h,s}^{(i)} \overset{\text{i.i.d.}}{\sim} \pi_h^{\mathsf{b}}(\cdot \,|\, s), \qquad s_{h,s}^{\prime(i)} \overset{\text{i.i.d.}}{\sim} P_h(\cdot \,|\, s, a), \qquad 1 \le i \le N_h^{\mathsf{trim}}(s). \tag{113}$$

  This forms the following dataset:

$$\mathcal{D}^{\text{i.i.d.}} := \Big\{\big(s, a_{h,s}^{(i)}, h, s_{h,s}^{\prime(i)}\big) \mid s \in \mathcal{S}, 1 \le h \le H, 1 \le i \le N_h^{\mathsf{trim}}(s)\Big\}. \tag{114}$$

In words, the dataset $\mathcal{D}^{\mathsf{trim,aug}}$ generated in Model 1 differs from $\mathcal{D}^{\mathsf{trim}}$ only if $N_h^{\mathsf{trim}}(s) > N_h^{\mathsf{main}}(s)$ occurs; this data generating mechanism ensures that $\mathcal{D}^{\mathsf{trim,aug}}$ comprises exactly $N_h^{\mathsf{trim}}(s)$ sample transitions from state s at step h. Two key features are: (a) the samples in $\mathcal{D}^{\mathsf{trim,aug}}$ are statistically independent, and (b) $\mathcal{D}^{\mathsf{trim,aug}}$ is essentially equivalent to $\mathcal{D}^{\mathsf{trim}}$ with high probability, as asserted below.

Lemma 7. *The above two datasets $\mathcal{D}^{\mathsf{trim,aug}}$ and $\mathcal{D}^{\text{i.i.d.}}$ have the same distributions. In addition, with probability exceeding $1-8\delta$, $\mathcal{D}^{\mathsf{trim,aug}} = \mathcal{D}^{\mathsf{trim}}$.*

Proof. Both $\mathcal{D}^{\mathsf{trim,aug}}$ and $\mathcal{D}^{\text{i.i.d.}}$ contain exactly $N_h^{\mathsf{trim}}(s)$ sample transitions from state s at step h. where $\{N_h^{\mathsf{trim}}(s)\}$ are statistically independent from the randomness of the samples. It is easily seen that: given $\{N_h^{\mathsf{trim}}(s)\}$, the sample transitions in $\mathcal{D}^{\mathsf{trim,aug}}$ across different steps are statistically independent. As a result, $\mathcal{D}^{\mathsf{trim}}$ and $\mathcal{D}^{\text{i.i.d.}}$ both consist of independent samples and are of the same distribution.

Furthermore, Lemma 3 tells us that with probability at least $1-8\delta$, $N_h^{\mathsf{trim}}(s) \le N_h^{\mathsf{main}}(s)$ holds for all $(s,h) \in \mathcal{S}\times[H]$, implying that that all data in $\mathcal{D}^{\mathsf{trim,aug}}$ come from $\mathcal{D}^{\mathsf{main}}$ and hence $\mathcal{D}^{\mathsf{trim,aug}} = \mathcal{D}^{\mathsf{trim}}$. □

7.3 Proof of Theorem 3

We first demonstrate that Theorem 3 is valid as long as the following theorem can be established.

Theorem 6. *Consider the dataset $\mathcal{D}_0$ described in Section 3.2, and any $0 < \delta < 1$. Suppose that $\mathcal{D}_0$ contains N sample transitions, and that the non-negative integers $\{N_h(s,a)\}$ defined in (50) obey*

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]: \qquad N_h(s,a) \geq \frac{K d_h^{\mathsf{b}}(s,a)}{8} - 5\sqrt{K d_h^{\mathsf{b}}(s,a) \log \frac{NH}{\delta}}, \tag{115}$$

with K some quantity obeying $K \geq 3872 H S C^{\star}_{\mathsf{clipped}} \log \frac{NH}{\delta}$. Assume that conditional on $\{N_h(s,a)\}$, the sample transitions $\{(s,a,h,s'_{(i)}) \,|\, 1 \leq i \leq N_h(s,a), (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]\}$ are statistically independent. The penalty terms are taken to be (55), where $c_{\mathsf{b}} \geq 16$ is chosen to be some constant. Then with probability at least $1-4\delta$, one has

$$\sum_s d_h^{\star}(s)\big(V_h^{\star}(s) - V_h^{\widehat{\pi}}(s)\big) \leq 80\sqrt{\frac{2c_{\mathsf{b}} H^3 S C^{\star}_{\mathsf{clipped}} \log \frac{NH}{\delta}}{K}}, \qquad 1 \leq h \leq H. \tag{116}$$

By construction, $\{N_h^{\mathsf{trim}}(s,a)\}$ are computed using $\mathcal{D}^{\mathsf{aux}}$, and hence are independent from the empirical model $\widehat{P}_h$ generated based on $\mathcal{D}^{\mathsf{trim}}$. Additionally, Lemma 7 permits us to treat the samples in $\mathcal{D}^{\mathsf{trim}}$ as being statistically independent. Recalling that the lower bound (57b) holds with probability at least $1-8\delta$, we can readily invoke Theorem 6 by taking $N_h(s,a) = N_h^{\mathsf{trim}}(s,a)$ and the property (42) to show that

$$\sum_{s\in\mathcal{S}} \rho(s)\big(V_1^{\star}(s) - V_1^{\widehat{\pi}}(s)\big) = \sum_{s\in\mathcal{S}} d_1^{\star}(s)\big(V_1^{\star}(s) - V_1^{\widehat{\pi}}(s)\big) \leq 80\sqrt{\frac{2c_{\mathsf{b}} H^3 S C^{\star}_{\mathsf{clipped}} \log \frac{NH}{\delta}}{K}} \tag{117}$$

with probability at least $1-12\delta$, provided that $K \geq 3872 H S C^{\star}_{\mathsf{clipped}} \log \frac{KH}{\delta}$. Setting the right-hand side of (117) to be smaller than ε immediately concludes the proof of Theorem 3, where we have used the fact that $N \leq KH$ in $\mathcal{D}_0$. As a consequence, it suffices to establish Theorem 6. In the sequel, we shall assume without loss of generality that we are working on the high-probability event (57).

7.3.1 Proof of Theorem 6

Step 1: showing that $\widehat{Q}_h(s,a) \leq Q_h^{\widehat{\pi}}(s,a)$. This part relies crucially on the following lemma.

Lemma 8. *Consider any $1 \leq h \leq H$, and any vector $V \in \mathbb{R}^S$ independent of $\widehat{P}_h$ obeying $\|V\|_\infty \leq H$. With probability at least $1 - 4\delta/H$, one has*

$$\big|(\widehat{P}_{h,s,a} - P_{h,s,a})V\big| \leq \sqrt{\frac{48 \mathsf{Var}_{\widehat{P}_{h,s,a}}(V) \log \frac{NH}{\delta}}{N_h(s,a)}} + \frac{48H \log \frac{NH}{\delta}}{N_h(s,a)} \tag{118}$$

$$\mathsf{Var}_{\widehat{P}_{h,s,a}}(V) \leq 2\mathsf{Var}_{P_{h,s,a}}(V) + \frac{5H^2 \log \frac{NH}{\delta}}{3N_h(s,a)} \tag{119}$$

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

Proof. The proof follows from exactly the same argument as that of Lemma 9, except that the assumed upper bound on $\|V\|_\infty$ is now H (as opposed to $\frac{1}{1-\gamma}$) and δ is replaced with δ/H. We thus omit the proof details for brevity. □

Additionally, we make note of the crude bound $\big|(\widehat{P}_{h,s,a} - P_{h,s,a})\widehat{V}_{h+1}\big| \leq \|\widehat{V}_{h+1}\|_\infty \leq H$. Also, given that Algorithm 2 works backwards, the iterate $\widehat{V}_{h+1}$ does not use $\widehat{P}_h$, and is hence statistically independent from $\widehat{P}_h$. Thus, we can readily apply Lemma 8 to obtain

$$\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]: \quad \big|(\widehat{P}_{h,s,a} - P_{h,s,a})\widehat{V}_{h+1}\big| \leq b_h(s,a) \tag{120}$$

in the presence of the Bernstein-style penalty (55), provided that the constant $c_{\mathsf{b}} > 0$ is sufficiently large.

In the sequel, we shall work with the high-probability events (120) and (119), in addition to (57). We intend to prove the following relation

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]: \quad \widehat{Q}_h(s,a) \le Q_h^{\widehat{\pi}}(s,a) \quad \text{and} \quad \widehat{V}_h(s) \le V_h^{\widehat{\pi}}(s) \tag{121}$$

hold with probability exceeding $1-4\delta$. Note that the latter assertion concerning $\widehat{V}_h$ is implied by the former, according to the following relation:

$$\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a) = \widehat{Q}_h\big(s, \widehat{\pi}_h(s)\big) \le Q_h^{\widehat{\pi}}\big(s, \widehat{\pi}_h(s)\big) = V_h^{\widehat{\pi}}(s). \tag{122}$$

Therefore, we focus on the first assertion and will show it by induction. First of all, the claim (121) holds trivially for the base case with $h = H+1$, given that $\widehat{Q}_{H+1}(s,a) = Q_{H+1}^{\widehat{\pi}}(s,a) = 0$. Next, suppose that $\widehat{Q}_{h+1}(s,a) \le Q_{h+1}^{\widehat{\pi}}(s,a)$ holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and some step $h+1$. We would like to show that the claimed inequality holds for step h as well. If $\widehat{Q}_h(s,a) = 0$, then the claim holds trivially; otherwise, our update rule (52) reveals that

$$\begin{aligned}
\widehat{Q}_h(s,a) &= r_h(s,a) + \widehat{P}_{h,s,a}\widehat{V}_{h+1} - b_h(s,a) \\
&= r_h(s,a) + P_{h,s,a}\widehat{V}_{h+1} + \big(\widehat{P}_{h,s,a} - P_{h,s,a}\big)\widehat{V}_{h+1} - b_h(s,a) \\
&\overset{\text{(i)}}{\le} r_h(s,a) + P_{h,s,a}V_{h+1}^{\widehat{\pi}} \overset{\text{(ii)}}{=} Q_h^{\widehat{\pi}}(s,a),
\end{aligned}$$

with probability at least $1-\delta/2$, where (i) results from (120) and (122) (i.e., $\widehat{V}_{h+1}(s) \le V_{h+1}^{\widehat{\pi}}(s)$), and (ii) arises from the Bellman equation. We have thus established (121) via a standard induction argument.

Step 2: bounding $V_h^\star(s) - V_h^{\widehat{\pi}}(s)$. In view of (122), we make the observation that

$$0 \le V_h^\star(s) - V_h^{\widehat{\pi}}(s) \le V_h^\star(s) - \widehat{V}_h(s) \le Q_h^\star\big(s, \pi_h^\star(s)\big) - \widehat{Q}_h\big(s, \pi_h^\star(s)\big), \tag{123}$$

where the last inequality holds true since $V_h^\star(s) = Q_h^\star(s, \pi_h^\star(s))$ and $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a) \ge \widehat{Q}_h(s, \pi_h^\star(s))$. Recognizing that

$$\begin{aligned}
Q_h^\star\big(s, \pi_h^\star(s)\big) &= r\big(s, \pi_h^\star(s)\big) + P_{h,s,\pi_h^\star(s)}V_{h+1}^\star, \\
\widehat{Q}_h\big(s, \pi_h^\star(s)\big) &= \max\Big\{r\big(s, \pi_h^\star(s)\big) + \widehat{P}_{h,s,\pi_h^\star(s)}\widehat{V}_{h+1} - b_h\big(s, \pi_h^\star(s)\big),\, 0\Big\},
\end{aligned}$$

we can continue the derivation of (123) to obtain

$$\begin{aligned}
V_h^\star(s) - \widehat{V}_h(s) &\le r\big(s, \pi_h^\star(s)\big) + P_{h,s,\pi_h^\star(s)}V_{h+1}^\star - \Big\{r\big(s, \pi_h^\star(s)\big) + \widehat{P}_{h,s,\pi_h^\star(s)}\widehat{V}_{h+1} - b_h\big(s, \pi_h^\star(s)\big)\Big\} \\
&= P_{h,s,\pi_h^\star(s)}V_{h+1}^\star - \widehat{P}_{h,s,\pi_h^\star(s)}\widehat{V}_{h+1} + b_h\big(s, \pi_h^\star(s)\big) \\
&= P_{h,s,\pi_h^\star(s)}\big(V_{h+1}^\star - \widehat{V}_{h+1}\big) - \Big(\widehat{P}_{h,s,\pi_h^\star(s)} - P_{h,s,\pi_h^\star(s)}\Big)\widehat{V}_{h+1} + b_h\big(s, \pi_h^\star(s)\big) \\
&\le P_{h,s,\pi_h^\star(s)}\big(V_{h+1}^\star - \widehat{V}_{h+1}\big) + 2b_h\big(s, \pi_h^\star(s)\big)
\end{aligned} \tag{124}$$

with probability at least $1-\delta$, where the last inequality is valid due to (120). For notational convenience, let us introduce a sequence of matrices $P_h^\star \in \mathbb{R}^{S\times S}$ $(1 \le h \le H)$ and vectors $b_h^\star \in \mathbb{R}^S$ $(1 \le h \le H)$, with their s-th rows given by

$$\big[P_h^\star\big]_{s,\cdot} := P_{h,s,\pi_h^\star(s)} \qquad \text{and} \qquad b_h^\star(s) := b_h\big(s, \pi_h^\star(s)\big). \tag{125}$$

This allows us to rewrite (124) in matrix/vector form as follows:

$$0 \le V_h^\star - \widehat{V}_h \le P_h^\star\big(V_{h+1}^\star - \widehat{V}_{h+1}\big) + 2b_h^\star. \tag{126}$$

The inequality (126) plays a key role in the analysis since it establishes a connection between the value estimation errors in step h and step $h+1$.

Given that $b_h^\star$, $P_h^\star$ and $V_h^\star - \widehat{V}_h$ are all non-negative, applying (126) recursively with the boundary condition $V_{H+1}^\star = \widehat{V}_{H+1} = 0$ leads to

$$\begin{aligned} 0 \le V_h^\star - \widehat{V}_h &\le P_h^\star\big(V_{h+1}^\star - \widehat{V}_{h+1}\big) + 2b_h^\star \\ &\le P_h^\star P_{h+1}^\star\big(V_{h+2}^\star - \widehat{V}_{h+2}\big) + 2P_h^\star b_{h+1}^\star + 2b_h^\star \le \cdots \\ &\le 2\sum_{j=h}^{H}\Big(\prod_{k=h}^{j-1} P_k^\star\Big) b_j^\star, \end{aligned}$$

where we adopt the following notation for convenience (note the order of the product)

$$\prod_{k=h}^{h-1} P_k^\star = I \qquad \text{and} \qquad \prod_{k=h}^{j-1} P_k^\star = P_h^\star \cdots P_{j-1}^\star \quad \text{if } j > h.$$

With this inequality in mind, we can let $d_h^\star := [d_h^\star(s)]_{s\in\mathcal{S}}$ be a S-dimensional vector and derive

$$\begin{aligned} \big\langle d_h^\star, V_h^\star - V_h^{\widehat{\pi}}\big\rangle \le \big\langle d_h^\star, V_h^\star - \widehat{V}_h\big\rangle &\le \left\langle d_h^\star, 2\sum_{j=h}^{H}\Big(\prod_{k=h}^{j-1} P_k^\star\Big) b_j^\star \right\rangle \\ &= 2\sum_{j=h}^{H}\big(d_h^\star\big)^\top\Big(\prod_{k=h}^{j-1} P_k^\star\Big) b_j^\star = 2\sum_{j=h}^{H}\big\langle d_j^\star, b_j^\star\big\rangle, \end{aligned} \tag{127}$$

where we have made use of (123) and the elementary identity (110).

Step 3: using concentrability to bound $\langle d_j^\star, b_j^\star\rangle$. To finish up, we need to make use of the concentrability coefficient. In what follows, we look at two cases separately.

- *Case 1:* $Kd_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big) \le 4c_{\mathsf{b}}\log\frac{NH}{\delta}$. Given that $b_h(s,a) \le H$ (cf. (55)), we necessarily have

$$b_j^\star(s) \le H \le H\cdot\frac{4c_{\mathsf{b}}\log\frac{NH}{\delta}}{Kd_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)} \le \frac{4c_{\mathsf{b}}C^\star_{\mathsf{clipped}}H\log\frac{NH}{\delta}}{K\min\big\{d_j^\star(s),\frac{1}{S}\big\}} \tag{128}$$

  in this case, where the last inequality arises from Definition 4.

- *Case 2:* $Kd_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big) > 4c_{\mathsf{b}}\log\frac{NH}{\delta}$. It follows from the assumption (115) that

$$\begin{aligned} N_j\big(s,\pi_j^\star(s)\big) &\ge \frac{Kd_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}{8} - 5\sqrt{Kd_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)\log\frac{N}{\delta}} \ge \frac{Kd_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}{16} \\ &\ge \frac{K\min\big\{d_j^\star\big(s,\pi_j^\star(s)\big),\frac{1}{S}\big\}}{16C^\star_{\mathsf{clipped}}} = \frac{K\min\big\{d_j^\star(s),\frac{1}{S}\big\}}{16C^\star_{\mathsf{clipped}}}, \end{aligned} \tag{129}$$

  as long as $c_{\mathsf{b}} > 0$ is sufficiently large. Here, the last line results from Definition 4 and the assumption that $\pi^\star$ is a deterministic policy (so that $d_j^\star(s) = d_j^\star\big(s,\pi_j^\star(s)\big)$).

  This further leads to

$$\begin{aligned} b_j^\star(s) &\le \sqrt{\frac{c_{\mathsf{b}}\log\frac{NH}{\delta}}{N_j\big(s,\pi_j^\star(s)\big)}\mathsf{Var}_{\widehat{P}_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} + c_{\mathsf{b}}H\frac{\log\frac{NH}{\delta}}{N_j\big(s,\pi_j^\star(s)\big)} \\ &\overset{\text{(i)}}{\le} \sqrt{\frac{2c_{\mathsf{b}}\log\frac{NH}{\delta}}{N_j\big(s,\pi_j^\star(s)\big)}\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} + 3c_{\mathsf{b}}H\frac{\log\frac{NH}{\delta}}{N_j\big(s,\pi_j^\star(s)\big)} \end{aligned}$$

$$\overset{\text{(ii)}}{\leq} \sqrt{\frac{32c_{\mathsf{b}}C^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K\min\left\{d_j^{\star}(s),\frac{1}{S}\right\}}\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})} + 48c_{\mathsf{b}}C^{\star}_{\mathsf{clipped}}H\frac{\log\frac{NH}{\delta}}{K\min\left\{d_j^{\star}(s),\frac{1}{S}\right\}}.$$

Here, (i) comes from (119) and the elementary inequality $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for any $x,y\geq 0$, provided that c_{b} is large enough; and (ii) relies on (129).

Putting the above two cases together, we arrive at

$$\begin{aligned}\sum_s d_j^{\star}(s)b_j^{\star}(s) &\leq \sum_s d_j^{\star}(s)\sqrt{\frac{32c_{\mathsf{b}}C^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K\min\left\{d_j^{\star}(s),\frac{1}{S}\right\}}\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})} + 48c_{\mathsf{b}}H\sum_s d_j^{\star}(s)\frac{C^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K\min\left\{d_j^{\star}(s),\frac{1}{S}\right\}}\\ &\leq \sum_s d_j^{\star}(s)\sqrt{\frac{32c_{\mathsf{b}}C^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K\min\left\{d_j^{\star}(s),\frac{1}{S}\right\}}\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})} + \frac{96c_{\mathsf{b}}HSC^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K},\end{aligned} \tag{130}$$

where the last inequality holds since

$$\sum_s \frac{d_j^{\star}(s)}{\min\left\{d_j^{\star}(s),\frac{1}{S}\right\}} \leq \sum_s d_j^{\star}(s)\left\{\frac{1}{d_j^{\star}(s)}+\frac{1}{1/S}\right\} \leq \sum_s 1 + S\sum_s d_j^{\star}(s) \leq 2S.$$

In addition, we make the observation that

$$\begin{aligned}&\sum_{j=h}^{H}\sum_s d_j^{\star}(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})}{\min\left\{d_j^{\star}(s),\frac{1}{S}\right\}}} \leq \sum_{j=h}^{H}\sum_s d_j^{\star}(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})}{d_j^{\star}(s)}} + \sum_{j=h}^{H}\sum_s d_j^{\star}(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})}{1/S}}\\ &= \sum_{j=h}^{H}\sum_s \sqrt{d_j^{\star}(s)\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})} + \sqrt{S}\sum_{j=h}^{H}\sum_s d_j^{\star}(s)\sqrt{\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})}\\ &\leq \sqrt{HS}\cdot\sqrt{\sum_{j=h}^{H}\sum_s d_j^{\star}(s)\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})} + \sqrt{S}\sqrt{\sum_{j=h}^{H}\sum_s d_j^{\star}(s)}\sqrt{\sum_{j=h}^{H}\sum_s d_j^{\star}(s)\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})}\\ &= 2\sqrt{HS}\cdot\sqrt{\sum_{j=h}^{H}\sum_s d_j^{\star}(s)\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1})}\\ &\leq 4\sqrt{HS\left(H^2+H\sum_{j=h}^{H}\left\langle d_j^{\star},b_j^{\star}\right\rangle\right)} \leq 4\sqrt{H^3S}+4\sqrt{H^2S\sum_{j=h}^{H}\left\langle d_j^{\star},b_j^{\star}\right\rangle},\end{aligned}$$

where the third line makes use of the Cauchy-Schwarz inequality, and the last line would hold as long as we could establish the following inequality

$$\sum_{j=h}^{H}\sum_s d_j^{\star}(s)\mathsf{Var}_{P_{j,s,\pi_j^{\star}(s)}}(\widehat{V}_{j+1}) \leq 4H^2 + 4H\sum_{j=h}^{H}\left\langle d_j^{\star},b_j^{\star}\right\rangle \tag{131}$$

for all $h\in[H]$ with probability exceeding $1-4\delta$. Substitution into (130) yields

$$\begin{aligned}\sum_{j=h}^{H}\sum_s d_j^{\star}(s)b_j^{\star}(s) &\leq \sqrt{\frac{32c_{\mathsf{b}}C^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}}\left\{4\sqrt{H^3S}+4\sqrt{H^2S\sum_{j=h}^{H}\left\langle d_j^{\star},b_j^{\star}\right\rangle}\right\} + \sum_{j=h}^{H}\frac{96c_{\mathsf{b}}HSC^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}\\ &\leq 16\sqrt{\frac{2c_{\mathsf{b}}H^2SC^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}}\sqrt{\sum_{j=h}^{H}\left\langle d_j^{\star},b_j^{\star}\right\rangle} + 16\sqrt{\frac{2c_{\mathsf{b}}H^3SC^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}} + \frac{96c_{\mathsf{b}}H^2SC^{\star}_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}\end{aligned}$$

$$\leq \frac{1}{2}\sum_{j=h}^{H}\langle d_j^\star, b_j^\star\rangle + \frac{256c_{\mathsf{b}}H^2SC^\star_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K} + 16\sqrt{\frac{2c_{\mathsf{b}}H^3SC^\star_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}} + \frac{96c_{\mathsf{b}}H^2SC^\star_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K},$$

where the last inequality follows from the elementary inequality $2xy \leq x^2+y^2$. Rearranging terms, we are left with

$$\begin{aligned}\sum_{j=h}^{H}\sum_{s} d_j^\star(s)b_j^\star(s) &\leq 32\sqrt{\frac{2c_{\mathsf{b}}H^3SC^\star_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}} + \frac{704c_{\mathsf{b}}H^2SC^\star_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}\\ &\leq 40\sqrt{\frac{2c_{\mathsf{b}}H^3SC^\star_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}},\end{aligned}$$

provided that $K \geq 3872HSC^\star_{\mathsf{clipped}}\log\frac{NH}{\delta}$. This taken collectively with (127) completes the proof of Theorem 6, as long as the inequality (131) can be validated.

Proof of inequality (131). First of all, we observe that

$$\begin{aligned}\widehat{V}_j(s) + 2b_j^\star\big(s,\pi_j^\star(s)\big) - P_{j,s,\pi_j^\star(s)}\widehat{V}_{j+1} &= \widehat{V}_j(s) - \widehat{P}_{j,s,\pi_j^\star(s)}\widehat{V}_{j+1} + 2b_j^\star\big(s,\pi_j^\star(s)\big) + \big(\widehat{P}_{j,s,\pi_j^\star(s)} - P_{j,s,\pi_j^\star(s)}\big)\widehat{V}_{j+1}\\ &\overset{\text{(i)}}{\geq} \widehat{V}_j(s) - \widehat{P}_{j,s,\pi_j^\star(s)}\widehat{V}_{j+1} + b_j^\star\big(s,\pi_j^\star(s)\big)\\ &\geq \widehat{V}_j(s) - \Big\{r\big(s,\pi_j^\star(s)\big) + \widehat{P}_{j,s,\pi_j^\star(s)}\widehat{V}_{j+1} - b_j^\star\big(s,\pi_j^\star(s)\big)\Big\}\\ &\geq \max_a \widehat{Q}_j(s,a) - \widehat{Q}_j\big(s,\pi_j^\star(s)\big) \geq 0\end{aligned}$$

for any $s\in\mathcal{S}$, where (i) is a consequence of (120), and the last line arises from (52) and (53). This implies the non-negativity of the vector $\widehat{V}_j + 2b_j^\star - P_j^\star\widehat{V}_{j+1}$, which in turn allows one to deduce that

$$\begin{aligned}\widehat{V}_j\circ\widehat{V}_j - \big(P_j^\star\widehat{V}_{j+1}\big)\circ\big(P_j^\star\widehat{V}_{j+1}\big) &= \big(\widehat{V}_j + P_j^\star\widehat{V}_{j+1}\big)\circ\big(\widehat{V}_j - P_j^\star\widehat{V}_{j+1}\big)\\ &\leq \big(\widehat{V}_j + P_j^\star\widehat{V}_{j+1}\big)\circ\big(\widehat{V}_j + 2b_j^\star - P_j^\star\widehat{V}_{j+1}\big)\\ &\leq 2H\big(\widehat{V}_j + 2b_j^\star - P_j^\star\widehat{V}_{j+1}\big),\end{aligned} \tag{132}$$

where the last line relies on Lemma 6. Consequently, we can demonstrate that

$$\begin{aligned}&\sum_{j=h}^{H}\sum_{s} d_j^\star(s)\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big) = \sum_{j=h}^{H}\big\langle d_j^\star, P_j^\star\big(\widehat{V}_{j+1}\circ\widehat{V}_{j+1}\big) - \big(P_j^\star\widehat{V}_{j+1}\big)\circ\big(P_j^\star\widehat{V}_{j+1}\big)\big\rangle\\
&= \sum_{j=h}^{H}\big(d_j^\star\big)^\top P_j^\star\widehat{V}_{j+1}\circ\widehat{V}_{j+1} - \big\langle d_j^\star, \big(P_j^\star\widehat{V}_{j+1}\big)\circ\big(P_j^\star\widehat{V}_{j+1}\big)\big\rangle\\
&\overset{\text{(i)}}{\leq} \sum_{j=h}^{H}\Big(\big\langle d_{j+1}^\star, \widehat{V}_{j+1}\circ\widehat{V}_{j+1}\big\rangle - \big\langle d_j^\star, \widehat{V}_j\circ\widehat{V}_j\big\rangle + 2H\big\langle d_j^\star, \widehat{V}_j + 2b_j^\star - P_j^\star\widehat{V}_{j+1}\big\rangle\Big)\\
&\overset{\text{(ii)}}{=} \sum_{j=h}^{H}\Big(\big\langle d_{j+1}^\star, \widehat{V}_{j+1}\circ\widehat{V}_{j+1}\big\rangle - \big\langle d_j^\star, \widehat{V}_j\circ\widehat{V}_j\big\rangle\Big) + 2H\sum_{j=h}^{H}\Big(\big\langle d_j^\star, \widehat{V}_j\big\rangle - \big\langle d_{j+1}^\star, \widehat{V}_{j+1}\big\rangle\Big) + 4H\sum_{j=h}^{H}\big\langle d_j^\star, b_j^\star\big\rangle\\
&= \big\langle d_{H+1}^\star, \widehat{V}_{H+1}\circ\widehat{V}_{H+1}\big\rangle - \big\langle d_h^\star, \widehat{V}_h\circ\widehat{V}_h\big\rangle + 2H\Big(\big\langle d_h^\star, \widehat{V}_h\big\rangle - \big\langle d_{H+1}^\star, \widehat{V}_{H+1}\big\rangle\Big) + 4H\sum_{j=h}^{H}\big\langle d_j^\star, b_j^\star\big\rangle\\
&\leq \big\|d_{H+1}^\star\big\|_1\big\|\widehat{V}_{H+1}\circ\widehat{V}_{H+1}\big\|_\infty + 2H\big\|d_h^\star\big\|_1\big\|\widehat{V}_h\big\|_\infty + 4H\sum_{j=h}^{H}\big\langle d_j^\star, b_j^\star\big\rangle\end{aligned}$$

$$\overset{\text{(iii)}}{\leq} 3H^2 + 4H\sum_{j=h}^{H}\langle d_j^\star, b_j^\star\rangle,$$

where (i) arises from (132) as well as the basic property $\left(d_j^\star\right)^\top P_j^\star = \left(d_{j+1}^\star\right)^\top$, (ii) follows by rearranging terms and using the property $\left(d_j^\star\right)^\top P_j^\star = \left(d_{j+1}^\star\right)^\top$ once again, and (iii) holds due to the fact that $\|\widehat{V}_h\|_\infty \leq H$ and $\|d_h^\star\|_1 = 1$. This concludes the proof of (131).

8 Discussion

Our primary contribution has been to pin down the sample complexity of model-based offline RL for the tabular settings, by establishing its (near) minimax optimality for both infinite- and finite-horizon MDPs. While reliable estimation of the transition kernel is often infeasible in the sample-starved regime, it does not preclude the success of this "plug-in" approach in learning the optimal policy. Encouragingly, the sample complexity characterization we have derived holds for the entire range of target accuracy level ε, thus revealing that sample optimality comes into effect without incurring any burn-in cost. This is in stark contrast to all prior results, which either suffered from sample sub-optimality or required a large burn-in sample size in order to yield optimal efficiency. We have demonstrated that sophisticated techniques like variance reduction are not necessary, as long as Bernstein-style lower confidence bounds are carefully employed to capture the variance of the estimates in each iteration.

Turning to future directions, we first note that the two-fold subsampling adopted in Algorithm 3 is likely unnecessary; it would be of interest to develop sharp analysis for the VI-LCB algorithm without sample splitting, which would call for more refined analysis in order to handle the complicated statistical dependency between different time steps. Notably, while avoiding sample splitting cannot improve the sample complexity in an order-wise sense, the potential gain in terms of the pre-constants as well as the algoritmic simplicity might be of practical interest. Moreover, given the appealing memory efficiency of model-free algorithms, understanding whether one can design sample-optimal model-free offline algorithms with minimal burn-in periods is another open direction. Moving beyond tabular settings, it would be of great interest to extend our analysis to accommodate model-based offline RL in more general scenarios; examples include MDPs with low-complexity linear representations, and offline RL involving multiple agents.

Acknowledgements

Y. Wei is supported in part by the NSF grants CCF-2106778, DMS-2147546/2015447, the NSF CAREER award DMS-2143215, and the Google Research Scholar Award. Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grants FA9550-19-1-0030 and FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009, CCF-1907661, DMS-2014279, IIS-2218713 and IIS-2218773. L. Shi and Y. Chi are supported in part by the grants ONR N00014-19-1-2404, NSF CCF-2106778 and DMS-2134080, and CAREER award ECCS-1818571. L. Shi is also gratefully supported by the Leo Finzi Memorial Fellowship, Wei Shen and Xuehong Zhang Presidential Fellowship, and Liang Ji-Dian Graduate Fellowship at Carnegie Mellon University. Part of this work was done while G. Li, Y. Chen and Y. Wei were visiting the Simons Institute for the Theory of Computing.

A Proof of auxiliary lemmas: infinite-horizon MDPs

A.1 Proof of Lemma 1

Before embarking on the proof, we introduce several notation. To make explicit the dependency on V, we shall express the penalty term using the following notation throughout this subsection:

$$b(s,a;V) = \min\left\{\max\left\{\sqrt{\frac{c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(V)},\ \frac{2c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\},\ \frac{1}{1-\gamma}\right\} + \frac{5}{N} \tag{133}$$

For any $Q, Q_1, Q_2 \in \mathbb{R}^{SA}$, we write

$$V(s) \coloneqq \max_a Q(s,a), \qquad V_1(s) \coloneqq \max_a Q_1(s,a) \qquad \text{and} \qquad V_2(s) \coloneqq \max_a Q_2(s,a) \tag{134}$$

for all $s \in \mathcal{S}$. Unless otherwise noted, we assume that

$$Q(s,a), Q_1(s,a), Q_2(s,a) \in \left[0, \frac{1}{1-\gamma}\right] \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}$$

throughout this subsection. In addition, let us define another operator $\widetilde{\mathcal{T}}_{\mathsf{pe}}$ obeying

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a) = r(s,a) - b(s,a;V) + \gamma \widehat{P}_{s,a} V \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A} \tag{135}$$

for any $Q \in \mathbb{R}^{SA}$. It is self-evident that

$$\widehat{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a) = \max\left\{\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a), 0\right\} \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{136}$$

γ-contraction. The main step of the proof lies in showing the monotonicity of the operator $\widetilde{\mathcal{T}}_{\mathsf{pe}}$ in the sense that

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q) \le \widetilde{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q}) \qquad \text{for any } Q \le \widetilde{Q}. \tag{137}$$

Suppose that this claim is valid for the moment, then one can demonstrate that: for any $Q_1, Q_2 \in \mathbb{R}^{SA}$,

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2) \le \widetilde{\mathcal{T}}_{\mathsf{pe}}\big(Q_2 + \|Q_1 - Q_2\|_\infty 1\big) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2), \tag{138a}$$

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2) \ge \widetilde{\mathcal{T}}_{\mathsf{pe}}\big(Q_2 - \|Q_1 - Q_2\|_\infty 1\big) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2), \tag{138b}$$

with 1 denoting the all-one vector. Additionally, observe that

$$\mathsf{Var}_{\widehat{P}_{s,a}}(V) = \mathsf{Var}_{\widehat{P}_{s,a}}(V + c \cdot 1) \qquad \text{and hence} \qquad b(s,a;V) = b(s,a;V + c \cdot 1)$$

for any constant c, which together with the identity $\widehat{P}1 = 1$ immediately leads to

$$\begin{aligned}
\left\|\widetilde{\mathcal{T}}_{\mathsf{pe}}\big(Q_2 - \|Q_1 - Q_2\|_\infty 1\big) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty &\le \gamma \left\|\widehat{P}\big(\|Q_1 - Q_2\|_\infty 1\big)\right\|_\infty = \gamma \|Q_1 - Q_2\|_\infty, \\
\left\|\widetilde{\mathcal{T}}_{\mathsf{pe}}\big(Q_2 + \|Q_1 - Q_2\|_\infty 1\big) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty &\le \gamma \left\|\widehat{P}\big(\|Q_1 - Q_2\|_\infty 1\big)\right\|_\infty = \gamma \|Q_1 - Q_2\|_\infty.
\end{aligned}$$

Taking this together with (138) yields

$$\left\|\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty,$$

which combined with the basic property $\left\|\widehat{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty \le \left\|\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty$ (as a result of (136)) justifies that

$$\left\|\widehat{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty. \tag{139}$$

The remainder of the proof is thus devoted to establishing the monotonicity property (137).

Proof of the monotonicity property (137). Consider any point $Q \in \mathbb{R}^{SA}$, and we would like to examine the derivative of $\widetilde{\mathcal{T}}_{\mathsf{pe}}$ at point Q. Towards this end, we consider any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and divide into several cases.

- *Case 1:* $\max\left\{\sqrt{\frac{c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)} \mathsf{Var}_{\widehat{P}_{s,a}}(V)}, \frac{2c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\} > \frac{1}{1-\gamma}$. In this case, the penalty term (133) simplifies to

$$b(s,a;V) = \frac{1}{1-\gamma} + \frac{5}{N}.$$

Taking the derivative of $\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a)$ w.r.t. the s'-th component of V leads to

$$\frac{\partial\big(\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a)\big)}{\partial V(s')} = \frac{\partial\big(r(s,a) - \frac{1}{1-\gamma} + \gamma \widehat{P}_{s,a} V\big)}{\partial V(s')} = \gamma \widehat{P}(s' \,|\, s,a) \geq 0 \tag{140}$$

for any $s' \in \mathcal{S}$.

- *Case 2:* $\sqrt{\frac{c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)} \mathsf{Var}_{\widehat{P}_{s,a}}(V)} < \frac{2c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} < \frac{1}{1-\gamma}$. The penalty (133) in this case reduces to

$$b(s,a;V) = \frac{2c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} + \frac{5}{N},$$

an expression that is independent of V. As a result, repeating the argument for Case 1 indicates that (140) continues to hold for this case.

- *Case 3:* $\frac{2c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} < \sqrt{\frac{c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)} \mathsf{Var}_{\widehat{P}_{s,a}}(V)} < \frac{1}{1-\gamma}$. In this case, the penalty term is given by

$$b(s,a;V) = \sqrt{\frac{c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)} \mathsf{Var}_{\widehat{P}_{s,a}}(V)} + \frac{5}{N}.$$

Note that in this case, we necessarily have

$$\mathsf{Var}_{\widehat{P}_{s,a}}(V) \geq \frac{4c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a)},$$

which together with the definition in (8) indicates that

$$\widehat{P}_{s,a}(V \circ V) - \big(\widehat{P}_{s,a} V\big)^2 \geq \frac{4c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a)} > 0. \tag{141}$$

As a result, for any $s' \in \mathcal{S}$, taking the derivative of $b(s,a;V)$ w.r.t. the s'-th component of V gives

$$\begin{aligned}
\frac{\partial b(s,a;V)}{\partial V(s')} &= \sqrt{\frac{c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)}} \frac{\partial \sqrt{\widehat{P}_{s,a}(V \circ V) - \big(\widehat{P}_{s,a} V\big)^2}}{\partial V(s')} \\
&= \sqrt{\frac{c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)}} \frac{\widehat{P}(s' \,|\, s,a) V(s') - \big(\widehat{P}_{s,a} V\big) \widehat{P}(s' \,|\, s,a)}{\sqrt{\widehat{P}_{s,a}(V \circ V) - \big(\widehat{P}_{s,a} V\big)^2}} \\
&\leq \sqrt{\frac{c_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)}} \frac{\widehat{P}(s' \,|\, s,a) V(s')}{\sqrt{\widehat{P}_{s,a}(V \circ V) - \big(\widehat{P}_{s,a} V\big)^2}} \\
&\leq \frac{1}{2}(1-\gamma) \widehat{P}(s' \,|\, s,a) V(s') \leq \gamma \widehat{P}(s' \,|\, s,a),
\end{aligned}$$

where the penultimate inequality relies on (141), and the last inequality is valid since $V(s') = \max_a Q(s',a) \leq \frac{1}{1-\gamma}$ and $\gamma \geq 1/2$. In turn, the preceding relation allows one to derive

$$\frac{\partial\big(\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a)\big)}{\partial V(s')} = \gamma \widehat{P}(s' \,|\, s,a) - \frac{\partial b(s,a;V)}{\partial V(s')} \geq 0$$

for any $s' \in \mathcal{S}$.

Putting the above cases together reveals that

$$\frac{\partial\big(\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a)\big)}{\partial V(s')} \geq 0 \qquad \text{for all } (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$$

holds almost everywhere (except for the boundary points of these cases). Recognizing that $\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)$ is continuous in Q and that V is non-decreasing in Q, one can immediately conclude that

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q) \leq \widetilde{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q}) \qquad \text{for any } Q \leq \widetilde{Q}. \tag{142}$$

Existence and uniqueness of fixed points. To begin with, note that for any $0 \leq Q \leq \frac{1}{1-\gamma}\cdot 1$, one has $0 \leq \widehat{\mathcal{T}}_{\mathsf{pe}}(Q) \leq \frac{1}{1-\gamma}\cdot 1$. If we produce the following sequence recursively:

$$Q^{(0)} = 0 \qquad \text{and} \qquad Q^{(t+1)} = \widehat{\mathcal{T}}_{\mathsf{pe}}(Q^{(t)}) \quad \text{for all } t \geq 0,$$

then the standard proof for the Banach fixed-point theorem (e.g., Agarwal et al. (2001, Theorem 1)) tells us that $Q^{(t)}$ converges to some point $Q^{(\infty)}$ as $t \to \infty$. Clearly, $Q^{(\infty)}$ is a fixed point of $\widehat{\mathcal{T}}_{\mathsf{pe}}$ obeying $0 \leq Q^{(\infty)} \leq \frac{1}{1-\gamma}\cdot 1$.

We then turn to justifying the uniqueness of fixed points of $\widehat{\mathcal{T}}_{\mathsf{pe}}$. Suppose that there exists another point $\widetilde{Q}$ obeying $\widetilde{Q} = \widehat{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q})$, which clearly satisfies $\widetilde{Q} \geq 0$. If $\|\widetilde{Q}\|_\infty > \frac{1}{1-\gamma}$, then

$$\big\|\widetilde{Q}\big\|_\infty = \big\|\widehat{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q})\big\|_\infty \leq \|r\|_\infty + \gamma\|\widehat{P}\|_1\big\|\widetilde{Q}\big\|_\infty \leq 1 + \gamma\big\|\widetilde{Q}\big\|_\infty < (1-\gamma)\big\|\widetilde{Q}\big\|_\infty + \gamma\big\|\widetilde{Q}\big\|_\infty = \big\|\widetilde{Q}\big\|_\infty,$$

resulting in contradiction. Consequently, one necessarily has $0 \leq \widetilde{Q} \leq \frac{1}{1-\gamma}\cdot 1$. Further, the γ-contraction property (139) implies that

$$\big\|\widetilde{Q} - Q^{(\infty)}\big\|_\infty = \big\|\widehat{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q}) - \widehat{\mathcal{T}}_{\mathsf{pe}}(Q^{(\infty)})\big\|_\infty \leq \gamma\big\|\widetilde{Q} - Q^{(\infty)}\big\|_\infty.$$

Given that $\gamma < 1$, this inequality cannot happen unless $\widetilde{Q} = Q^\infty$, thus confirming the uniqueness of Q^∞.

A.2 Proof of Lemma 2

Let us first recall the monotone non-decreasing property (137) of the operator $\widetilde{\mathcal{T}}_{\mathsf{pe}}$ defined in (135), which taken together with the property (136) readily yields

$$\widehat{\mathcal{T}}_{\mathsf{pe}}(Q) \leq \widehat{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q}) \tag{143}$$

for any Q and $\widetilde{Q}$ obeying $Q \leq \widetilde{Q}$, $0 \leq Q \leq \frac{1}{1-\gamma}\cdot 1$ and $0 \leq \widetilde{Q} \leq \frac{1}{1-\gamma}\cdot 1$ (with 1 the all-one vector). Given that $\widehat{Q}_0 = 0 \leq \widehat{Q}^\star_{\mathsf{pe}}$, we can apply (143) to obtain

$$\widehat{Q}_1 = \widehat{\mathcal{T}}_{\mathsf{pe}}(Q_0) \leq \widehat{\mathcal{T}}_{\mathsf{pe}}(\widehat{Q}^\star_{\mathsf{pe}}) = \widehat{Q}^\star_{\mathsf{pe}}.$$

Repeat this argument recursively to arrive at

$$\widehat{Q}_\tau \leq \widehat{Q}^\star_{\mathsf{pe}} \qquad \text{for all } \tau \geq 0.$$

In addition, it comes directly from Lemma 1 that

$$\begin{aligned}\big\|\widehat{Q}_\tau - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty &= \big\|\widehat{\mathcal{T}}_{\mathsf{pe}}(\widehat{Q}_{\tau-1}) - \widehat{\mathcal{T}}_{\mathsf{pe}}(\widehat{Q}^\star_{\mathsf{pe}})\big\|_\infty \leq \gamma\big\|\widehat{Q}_{\tau-1} - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \\ &\leq \cdots \leq \gamma^\tau\big\|\widehat{Q}_0 - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \\ &\leq \frac{\gamma^\tau}{1-\gamma}\end{aligned} \tag{144}$$

for any $\tau \geq 0$, where the last inequality is valid since $\widehat{Q}_0 = 0$ and $\|\widehat{Q}^\star_{\mathsf{pe}}\|_\infty \leq \frac{1}{1-\gamma}$ (see Lemma 1). The other claim (33) also follows immediately by taking the right-hand side of (144) to be no larger than $1/N$.

A.3 Proof of Lemma 4

For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, if $\frac{Nd^{\mathsf{b}}(s,a)}{12} < \frac{2}{3}\log\frac{SN}{\delta}$, then it is self-evident that this pair satisfies (69). As a consequence, it suffices to focus attention on the following set of state-action pairs:

$$\mathcal{N}_{\mathsf{large}} := \left\{(s,a) \,\middle|\, d^{\mathsf{b}}(s,a) \geq \frac{8\log\frac{SN}{\delta}}{N}\right\}. \tag{145}$$

To bound the cardinality of $\mathcal{N}_{\mathsf{large}}$, we make the observation that

$$\left|\mathcal{N}_{\mathsf{large}}\right| \cdot \frac{8\log\frac{SN}{\delta}}{N} \leq \sum_{(s,a)\in\mathcal{N}_{\mathsf{large}}} d^{\mathsf{b}}(s,a) \leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\mathsf{b}}(s,a) \leq 1,$$

thus leading to the crude bound

$$\left|\mathcal{N}_{\mathsf{large}}\right| \leq \frac{N}{8\log\frac{SN}{\delta}} \leq \frac{N}{8}. \tag{146}$$

Let us now look at any $(s,a) \in \mathcal{N}_{\mathsf{large}}$. Given that $N(s,a)$ can be viewed as the sum of N independent Bernoulli random variables each with mean $d^{\mathsf{b}}(s,a)$, we can apply the Bernstein inequality to yield

$$\mathbb{P}\Big\{\big|N(s,a) - Nd^{\mathsf{b}}(s,a)\big| \geq \tau\Big\} \leq 2\exp\left(-\frac{\tau^2/2}{v_{s,a}+\tau/3}\right)$$

for any $\tau \geq 0$, where we define

$$v_{s,a} := N\mathsf{Var}\Big(\mathbb{1}\big\{(s_i,a_i) = (s,a)\big\}\Big) \leq Nd^{\mathsf{b}}(s,a).$$

A little algebra then yields that with probability at least $1-\delta$,

$$\big|N(s,a) - Nd^{\mathsf{b}}(s,a)\big| \leq \sqrt{4v_{s,a}\log\frac{2}{\delta}} + \frac{2}{3}\log\frac{2}{\delta} \leq \sqrt{4Nd^{\mathsf{b}}(s,a)\log\frac{2}{\delta}} + \log\frac{2}{\delta}. \tag{147}$$

Combining this result with the union bound over $(s,a) \in \mathcal{N}_{\mathsf{large}}$ and making use of (146) give: with probability at least $1-\delta$,

$$\big|N(s,a) - Nd^{\mathsf{b}}(s,a)\big| \leq \sqrt{4Nd^{\mathsf{b}}(s,a)\log\frac{N}{\delta}} + \log\frac{N}{\delta} \tag{148}$$

holds simultaneously for all $(s,a) \in \mathcal{N}_{\mathsf{large}}$. Recalling that $Nd^{\mathsf{b}}(s,a) \geq 8\log\frac{NS}{\delta}$ holds for any $(s,a) \in \mathcal{N}_{\mathsf{large}}$, we can easily verify that

$$N(s,a) \geq Nd^{\mathsf{b}}(s,a) - \left(\sqrt{4Nd^{\mathsf{b}}(s,a)\log\frac{N}{\delta}} + \log\frac{N}{\delta}\right) \geq \frac{Nd^{\mathsf{b}}(s,a)}{12}, \tag{149}$$

thereby establishing (69) for any $(s,a) \in \mathcal{N}_{\mathsf{large}}$. This concludes the proof.

A.4 Proof of Lemma 5

If $N(s,a) = 0$, then the inequalities hold trivially. Hence, it is sufficient to focus on the case where $N(s,a) > 0$. Before proceeding, we make note of a key Bernstein-style result; the proof is deferred to Appendix A.4.1.

Lemma 9. *Consider any given pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ with $N(s,a) > 0$. Let $V \in \mathbb{R}^S$ be any vector independent of $\widehat{P}_{s,a}$ obeying $\|V\|_\infty \leq \frac{1}{1-\gamma}$. With probability at least $1-4\delta$, one has*

$$\big|(\widehat{P}_{s,a} - P_{s,a})V\big| \leq \sqrt{\frac{48\mathsf{Var}_{\widehat{P}_{s,a}}(V)\log\frac{N}{\delta}}{N(s,a)}} + \frac{48\log\frac{N}{\delta}}{(1-\gamma)N(s,a)} \tag{150a}$$

$$\mathsf{Var}_{\widehat{P}_{s,a}}(V) \leq 2\mathsf{Var}_{P_{s,a}}(V) + \frac{5\log\frac{N}{\delta}}{3(1-\gamma)^2N(s,a)} \tag{150b}$$

Remark 6. In words, Lemma 9 develops a Bernstein bound (150a) on $\big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\big|$ that makes clear the importance of the variance parameter. Lemma 9 (cf. (150b)) also ascertains that the variance w.r.t. the empirical distribution $\widehat{P}_{s,a}$ does not deviate much from the variance w.r.t. the true distribution $P_{s,a}$.

Equipped with this result, we are now ready to present the proof of Lemma 5, which is built upon a leave-one-out decoupling argument and consists of the following steps.

Step 1: construction of auxiliary state-absorbing MDPs. Recall that $\widehat{\mathcal{M}}$ is the empirical MDP. For each state $s \in \mathcal{S}$ and each scalar $u \geq 0$, we construct an auxiliary state-absorbing MDP $\widehat{\mathcal{M}}^{s,u}$ in a way that makes it identical to the empirical MDP $\widehat{\mathcal{M}}$ except for state s. More specifically, the transition kernel of the auxiliary MDP $\widehat{\mathcal{M}}^{s,u}$ — denoted by $P^{s,u}$ — is chosen such that

$$\begin{aligned} P^{s,u}(\,\widetilde{s}\,|\,s,a) &= \mathbb{1}(\widetilde{s}=s) && \text{for all } (\,\widetilde{s},a) \in \mathcal{S}\times\mathcal{A}, \\ P^{s,u}(\cdot\,|\,s',a) &= \widehat{P}(\cdot\,|\,s',a) && \text{for all } (s',a)\in\mathcal{S}\times\mathcal{A} \text{ and } s'\neq s; \end{aligned}$$

and the reward function of $\widehat{\mathcal{M}}^{s,u}$ — denoted by $r^{s,u}$ — is set to be

$$\begin{aligned} r^{s,u}(s,a) &= u && \text{for all } a\in\mathcal{A}, \\ r^{s,u}(s',a) &= r(s',a) && \text{for all } (s',a)\in\mathcal{S}\times\mathcal{A} \text{ and } s'\neq s. \end{aligned}$$

In words, the probability transition kernel of $\widehat{\mathcal{M}}^{s,u}$ is obtained by dropping all randomness of $\widehat{P}_{s,a}$ ($a\in\mathcal{A}$) that concerns state s and making s an absorbing state. In addition, let us define the pessimistic Bellman operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{s,u}$ based on the auxiliary MDP $\widehat{\mathcal{M}}^{s,u}$ such that

$$\widehat{\mathcal{T}}_{\mathsf{pe}}^{s,u}(Q)(s,a) \coloneqq \max\Big\{r^{s,u}(s,a) + \gamma P_{s,a}^{s,u}V - b^{s,u}(s,a;V),\,0\Big\} \tag{151}$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, where the penalty term is taken to be

$$b^{s,u}(s,a;V) = \min\left\{\max\left\{\sqrt{\frac{c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{P^{s,u}(\cdot\,|\,s,a)}(V)},\ \frac{2c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\},\ \frac{1}{1-\gamma}\right\} + \frac{5}{N}. \tag{152}$$

Step 2: the correspondence between the empirical MDP and auxiliary MDP. Taking

$$u^\star = (1-\gamma)\widehat{V}_{\mathsf{pe}}^\star(s) + \min\left\{\frac{2c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)\max_a N(s,a)},\ \frac{1}{1-\gamma}\right\} + \frac{5}{N}, \tag{153}$$

we claim that there exists a fixed point $\widehat{Q}_{s,u^\star}^\star$ of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{s,u^\star}$ whose corresponding value function $\widehat{V}_{s,u^\star}^\star$ coincides with $\widehat{V}_{\mathsf{pe}}^\star$. To justify this, it suffices to verify the following properties:

- Consider any $a\in\mathcal{A}$. Given that $P^{s,u}(\cdot\,|\,s,a)$ only has a single non-zero entry (equal to 1), it is easily seen that $\mathsf{Var}_{P^{s,u}(\cdot\,|\,s,a)}(V)=0$ holds for any V and any u, thus indicating that

$$b^{s,u}(s,a;V) = \min\left\{\frac{2c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)},\ \frac{1}{1-\gamma}\right\} + \frac{5}{N}. \tag{154}$$

  Consequently, for state s, one has

$$\begin{aligned} \max_a\Big\{r^{s,u^\star}(s,a) - b^{s,u^\star}\big(s,a;\widehat{V}_{\mathsf{pe}}^\star\big) + \gamma\big\langle P^{s,u^\star}(\cdot\,|\,s,a),\ \widehat{V}_{\mathsf{pe}}^\star\big\rangle\Big\} &= \max_a\Big\{u^\star - b^{s,u^\star}\big(s,a;\widehat{V}_{\mathsf{pe}}^\star\big) + \gamma\widehat{V}_{\mathsf{pe}}^\star(s)\Big\} \\ &= u^\star - \min_a b^{s,u^\star}\big(s,a;\widehat{V}_{\mathsf{pe}}^\star\big) + \gamma\widehat{V}_{\mathsf{pe}}^\star(s) \\ &= (1-\gamma)\widehat{V}_{\mathsf{pe}}^\star(s) + \gamma\widehat{V}_{\mathsf{pe}}^\star(s) \\ &= \widehat{V}_{\mathsf{pe}}^\star(s), \end{aligned} \tag{155}$$

  where the third identity makes use of our choice (153) of $u^\star$ and (154).

- Next, consider any $s' \neq s$ and any $a \in \mathcal{A}$. We make the observation that

$$\begin{aligned}&\max\Big\{r^{s,u^\star}(s',a) - b^{s,u^\star}\big(s',a;\widehat{V}^\star_{\mathsf{pe}}\big) + \gamma\big\langle P^{s,u^\star}(\cdot\,|\,s',a),\ \widehat{V}^\star_{\mathsf{pe}}\big\rangle,\ 0\Big\}\\ &\quad = \max\Big\{r(s',a) - b\big(s',a;\widehat{V}^\star_{\mathsf{pe}}\big) + \gamma\big\langle \widehat{P}(\cdot\,|\,s',a),\ \widehat{V}^\star_{\mathsf{pe}}\big\rangle,\ 0\Big\} = \widehat{Q}^\star_{\mathsf{pe}}(s',a),\end{aligned} \tag{156}$$

  where the last relation holds since $\widehat{Q}^\star_{\mathsf{pe}}$ is a fixed point of $\widehat{\mathcal{T}}_{\mathsf{pe}}$.

Armed with (155) and (156), we see that $\widehat{Q}^\star_{s,u^\star} = \widehat{\mathcal{T}}^{s,u^\star}_{\mathsf{pe}}\big(\widehat{Q}^\star_{s,u^\star}\big)$ by taking

$$\begin{aligned}&\max_{a\in\mathcal{A}} \widehat{Q}^\star_{s,u^\star}(s,a) = \widehat{V}^\star_{\mathsf{pe}}(s),\\ &\widehat{Q}^\star_{s,u^\star}(s',a) = \widehat{Q}^\star_{\mathsf{pe}}(s',a) \qquad\qquad \text{for all } s' \neq s \text{ and } a \in \mathcal{A}.\end{aligned}$$

This readily confirms the existence of a fixed point of $\widehat{\mathcal{T}}^{s,u^\star}_{\mathsf{pe}}$ whose corresponding value coincides with $\widehat{V}^\star_{\mathsf{pe}}$.

Step 3: building an ϵ-net. Consider any $(s,a) \in \mathcal{S}\times\mathcal{A}$ with $N(s,a) > 0$. Construct a set $\mathcal{U}_{\mathrm{cover}}$ as follows

$$\mathcal{U}_{\mathrm{cover}} \coloneqq \left\{\frac{i}{N}\ |\ 1 \le i \le N u_{\max}\right\}, \tag{157}$$

with $u_{\max} = \min\left\{\frac{2c_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}, \frac{1}{1-\gamma}\right\} + \frac{5}{N} + 1$. This can be viewed as the ϵ-net (Vershynin, 2018) of the range $[0, u_{\max}] \subseteq \big[0, \frac{2}{1-\gamma}\big]$ with $\epsilon = 1/N$. Let us construct an auxiliary MDP $\widehat{\mathcal{M}}^{s,u}$ as in Step 1 for each $u \in \mathcal{U}_{\mathrm{cover}}$. Repeating the argument in the proof of Lemma 1 (see Section A.1), we can easily show that there exists a unique fixed point $\widehat{Q}^\star_{s,u}$ of $\widehat{\mathcal{M}}^{s,u}$, which also obeys $0 \le \widehat{Q}^\star_{s,u} \le \frac{1}{1-\gamma}\cdot 1$. In what follows, we denote by $\widehat{V}^\star_{s,u}$ the corresponding value function of $\widehat{Q}^\star_{s,u}$.

Recognizing that $\widehat{\mathcal{M}}^{s,u}$ is statistically independent from $\widehat{P}_{s,a}$ for any $u \in \mathcal{U}_{\mathrm{cover}}$ (by construction), we can apply Lemma 9 in conjunction with the union bound (over all $u \in \mathcal{U}_{\mathsf{cover}}$) to show that, with probability exceeding $1-\delta$,

$$\left|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widehat{V}^\star_{s,u}\right| \le \sqrt{\frac{48\log\frac{8N^2}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u}\big)} + \frac{48\log\frac{8N^2}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}, \tag{158a}$$

$$\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u}\big) \le 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u}\big) + \frac{5\log\frac{8N^2}{(1-\gamma)\delta}}{3(1-\gamma)^2N(s,a)} \tag{158b}$$

hold simultaneously for all $u \in \mathcal{U}_{\mathrm{cover}}$. Clearly, the total number of (s,a) pairs with $N(s,a) > 0$ cannot exceed N. Thus, taking the union bound over all these pairs yield that, with probability at least $1-\delta$,

$$\left|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widehat{V}^\star_{s,u}\right| \le \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u}\big)} + \frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}, \tag{159a}$$

$$\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u}\big) \le 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u}\big) + \frac{5\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2N(s,a)} \tag{159b}$$

hold simultaneously for all $(s,a,u) \in \mathcal{S}\times\mathcal{A}\times\mathcal{U}_{\mathrm{cover}}$ obeying $N(s,a) > 0$.

Step 4: a covering argument. In this step, we shall work on the high-probability event (158) that holds simultaneously for all $u \in \mathcal{U}_{\mathrm{cover}}$. Given that $\widehat{V}^\star_{\mathsf{pe}}$ satisfies the trivial bound $0 \le \widehat{V}^\star_{\mathsf{pe}}(s) \le \frac{1}{1-\gamma}$ for all $s \in \mathcal{S}$, one can find some $u_0 \in \mathcal{U}_{\mathrm{cover}}$ such that $|u_0 - u^\star| \le 1/N$, where we recall the choice of $u^\star$ in (153). From the definition of the MDP $\widehat{\mathcal{M}}^{s,u}$ and the operator (151), it is readily seen that

$$\big\|\widehat{\mathcal{T}}^{s,u_0}_{\mathsf{pe}}(Q) - \widehat{\mathcal{T}}^{s,u^\star}_{\mathsf{pe}}(Q)\big\|_\infty \le |u_0 - u^\star| \le \frac{1}{N}$$

holds for any $Q \in \mathbb{R}^{SA}$. Consequently, we can use γ-contraction of the operator to obtain

$$\begin{aligned}
\big\|\widehat{Q}^\star_{s,u_0} - \widehat{Q}^\star_{s,u^\star}\big\|_\infty &= \Big\|\widehat{\mathcal{T}}^{s,u_0}_{\mathsf{pe}}\big(\widehat{Q}^\star_{s,u_0}\big) - \widehat{\mathcal{T}}^{s,u^\star}_{\mathsf{pe}}\big(\widehat{Q}^\star_{s,u^\star}\big)\Big\|_\infty \\
&\le \Big\|\widehat{\mathcal{T}}^{s,u^\star}_{\mathsf{pe}}\big(\widehat{Q}^\star_{s,u_0}\big) - \widehat{\mathcal{T}}^{s,u^\star}_{\mathsf{pe}}\big(\widehat{Q}^\star_{s,u^\star}\big)\Big\|_\infty + \Big\|\widehat{\mathcal{T}}^{s,u_0}_{\mathsf{pe}}\big(\widehat{Q}^\star_{s,u_0}\big) - \widehat{\mathcal{T}}^{s,u^\star}_{\mathsf{pe}}\big(\widehat{Q}^\star_{s,u_0}\big)\Big\|_\infty \\
&\le \gamma\big\|\widehat{Q}^\star_{s,u_0} - \widehat{Q}^\star_{s,u^\star}\big\|_\infty + \frac{1}{N},
\end{aligned}$$

which implies that

$$\big\|\widehat{Q}^\star_{s,u_0} - \widehat{Q}^\star_{s,u^\star}\big\|_\infty \le \frac{1}{(1-\gamma)N}$$

and therefore

$$\big\|\widehat{V}^\star_{s,u_0} - \widehat{V}^\star_{s,u^\star}\big\|_\infty \le \big\|\widehat{Q}^\star_{s,u_0} - \widehat{Q}^\star_{s,u^\star}\big\|_\infty \le \frac{1}{(1-\gamma)N}.$$

This in turn allows us to demonstrate that

$$\begin{aligned}
&\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u_0}\big) - \mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u^\star}\big) \\
&\quad= P_{s,a}\Big(\big(\widehat{V}^\star_{s,u_0} - P_{s,a}\widehat{V}^\star_{s,u_0}\big)\circ\big(\widehat{V}^\star_{s,u_0} - P_{s,a}\widehat{V}^\star_{s,u_0}\big) - \big(\widehat{V}^\star_{s,u^\star} - P_{s,a}\widehat{V}^\star_{s,u^\star}\big)\circ\big(\widehat{V}^\star_{s,u^\star} - P_{s,a}\widehat{V}^\star_{s,u^\star}\big)\Big) \\
&\quad\le P_{s,a}\Big(\big(\widehat{V}^\star_{s,u_0} - P_{s,a}\widehat{V}^\star_{s,u^\star}\big)\circ\big(\widehat{V}^\star_{s,u_0} - P_{s,a}\widehat{V}^\star_{s,u^\star}\big) - \big(\widehat{V}^\star_{s,u^\star} - P_{s,a}\widehat{V}^\star_{s,u^\star}\big)\circ\big(\widehat{V}^\star_{s,u^\star} - P_{s,a}\widehat{V}^\star_{s,u^\star}\big)\Big) \\
&\quad\le P_{s,a}\Big(\big(\widehat{V}^\star_{s,u_0} - P_{s,a}\widehat{V}^\star_{s,u^\star} + \widehat{V}^\star_{s,u^\star} - P_{s,a}\widehat{V}^\star_{s,u^\star}\big)\circ\big(\widehat{V}^\star_{s,u_0} - \widehat{V}^\star_{s,u^\star}\big)\Big) \\
&\quad\le \frac{2}{1-\gamma}\Big|P_{s,a}\big(\widehat{V}^\star_{s,u_0} - \widehat{V}^\star_{s,u^\star}\big)\Big| \le \frac{2}{1-\gamma}\big\|\widehat{V}^\star_{s,u_0} - \widehat{V}^\star_{s,u^\star}\big\|_\infty \le \frac{2}{(1-\gamma)^2N},
\end{aligned}$$

where the third line comes from the fact that $\mathbb{E}[X] = \arg\min_c \mathbb{E}[(X-c)^2]$, and the last line relies on the property $0 \le \widehat{V}^\star_{s,u_0}, \widehat{V}^\star_{s,u^\star} \le \frac{1}{1-\gamma}$. In addition, by swapping $\widehat{V}^\star_{s,u_0}$ and $\widehat{V}^\star_{s,u^\star}$, we can derive

$$\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u^\star}\big) - \mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u_0}\big) \le \frac{2}{(1-\gamma)^2N},$$

and then

$$\Big|\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u_0}\big) - \mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u^\star}\big)\Big| \le \frac{2}{(1-\gamma)^2N}. \tag{160}$$

Clearly, this bound (160) continues to be valid if we replace $P_{s,a}$ with $\widehat{P}_{s,a}$.

With the above perturbation bounds in mind, we can invoke the triangle inequality and (159a) to reach

$$\begin{aligned}
\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widehat{V}^\star_{\mathsf{pe}}\Big| &= \Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widehat{V}^\star_{s,u^\star}\Big| \le \Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widehat{V}^\star_{s,u_0}\Big| + \Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\big(\widehat{V}^\star_{s,u^\star} - \widehat{V}^\star_{s,u_0}\big)\Big| \\
&\le \Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widehat{V}^\star_{s,u_0}\Big| + \frac{2}{N(1-\gamma)} \\
&\le \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u_0}\big)} + \frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} + \frac{2}{N(1-\gamma)} \\
&\le \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u^\star}\big)} + \sqrt{\frac{96\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)^2N(s,a)}} + \frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} + \frac{2}{N(1-\gamma)} \\
&\le \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u^\star}\big)} + \frac{60\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)},
\end{aligned} \tag{161}$$

where the second line holds since

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\big(\widehat{V}^\star_{s,u^\star} - \widehat{V}^\star_{s,u_0}\big)\Big| \le \big(\big\|\widehat{P}_{s,a}\big\|_1 + \big\|P_{s,a}\big\|_1\big)\big\|\widehat{V}^\star_{s,u^\star} - \widehat{V}^\star_{s,u_0}\big\|_\infty \le \frac{2}{N(1-\gamma)},$$

the penultimate line is valid due to (160), and the last line holds true under the conditions that $T \geq N(s,a)$ and that T is sufficiently large. Moreover, apply (159b) and the triangle inequality to arrive at

$$\begin{aligned}
\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big) &= \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big) \leq \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big) + \Big|\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big) - \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big)\Big| \\
&\overset{\text{(i)}}{\leq} 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big) + \frac{5\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N(s,a)} + \frac{2}{(1-\gamma)^2 N} \\
&\leq 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big) + 2\Big|\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big) - \mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big)\Big| + \frac{5\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N(s,a)} + \frac{2}{(1-\gamma)^2 N} \\
&\overset{\text{(ii)}}{\leq} 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big) + \frac{6}{(1-\gamma)^2 N} + \frac{5\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N(s,a)} \\
&\leq 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big) + \frac{23\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N(s,a)},
\end{aligned} \tag{162}$$

where (i) arise from (159b) and (160), (ii) follows from (160), and the last line holds true since $N \geq N(s,a)$.

Step 5: extending the bounds to $\widetilde{V}$. Consider any $\widetilde{V}$ obeying $\|\widetilde{V} - \widehat{V}^{\star}_{\mathsf{pe}}\|_\infty \leq \frac{1}{N}$ and $\|\widetilde{V}\|_\infty \leq \frac{1}{1-\gamma}$. Invoke (161) and the triangle inequality to arrive at

$$\begin{aligned}
\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widetilde{V}\Big| &\leq \Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widehat{V}^{\star}_{\mathsf{pe}}\Big| + \Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\big(\widehat{V}^{\star}_{\mathsf{pe}} - \widetilde{V}\big)\Big| \\
&\leq \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big)} + \frac{60\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} + \frac{2}{N}, \\
&\leq 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big)} + \frac{62\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} \\
&= 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big)} + \frac{62\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)},
\end{aligned} \tag{163}$$

where the penultimate inequality relies on $N \geq N(s,a)$, and the second line holds since

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\big(\widehat{V}^{\star}_{\mathsf{pe}} - \widetilde{V}\big)\Big| \leq \big(\big\|\widehat{P}_{s,a}\big\|_1 + \big\|P_{s,a}\big\|_1\big)\big\|\widehat{V}^{\star}_{\mathsf{pe}} - \widetilde{V}\big\|_\infty \leq \frac{2}{N}.$$

Given that $\big\|\widetilde{V} - \widehat{V}^{\star}_{\mathsf{pe}}\big\|_\infty \leq 1/N$, we can repeat the argument for (160) allows one to demonstrate that

$$\Big|\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big) - \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widetilde{V}\big)\Big| \leq \frac{2}{(1-\gamma)^2 N}$$

which taken together with (163) and the basic inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ gives

$$\begin{aligned}
\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widetilde{V}\Big| &\leq 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widetilde{V}\big)} + 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\cdot\frac{2}{(1-\gamma)^2 N}} + \frac{62\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} \\
&\leq 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widetilde{V}\big)} + \frac{74\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}.
\end{aligned}$$

Additionally, repeating the argument for (162) leads to another desired inequality:

$$\begin{aligned}
\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widetilde{V}\big) &\leq 2\mathsf{Var}_{P_{s,a}}\big(\widetilde{V}\big) + \frac{6}{(1-\gamma)N} + \frac{23\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N(s,a)} \\
&\leq 2\mathsf{Var}_{P_{s,a}}\big(\widetilde{V}\big) + \frac{41\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a)}.
\end{aligned}$$

A.4.1 Proof of Lemma 9

In this proof, we shalle often use $\mathsf{Var}_{s,a}$ to abbreviate $\mathsf{Var}_{P_{s,a}}$ for notational simplicity. Before proceeding, let us define the following vector

$$\overline{V} = V - (P_{s,a}V)1, \tag{164}$$

with 1 denoting the all-one vector. It is clearly seen that

$$P_{s,a}\big(\overline{V} \circ \overline{V}\big) = P_{s,a}\big(V \circ V\big) - \big(P_{s,a}V\big)^2 = \mathsf{Var}_{s,a}(V). \tag{165}$$

In addition, we make note of the following basic facts that will prove useful:

$$\|V\|_\infty \le \frac{1}{1-\gamma}, \qquad \|\overline{V}\|_\infty \le \frac{1}{1-\gamma}, \qquad \|\overline{V} \circ \overline{V}\|_\infty \le \|\overline{V}\|_\infty^2 \le H^2, \tag{166a}$$

$$\mathsf{Var}_{s,a}\big(\overline{V} \circ \overline{V}\big) \le P_{s,a}\big(\overline{V} \circ \overline{V} \circ \overline{V} \circ \overline{V}\big) \le \frac{1}{(1-\gamma)^2} P_{s,a}\big(\overline{V} \circ \overline{V}\big) = \frac{1}{(1-\gamma)^2}\mathsf{Var}_{s,a}\big(V\big). \tag{166b}$$

Proof of inequality (150a). If $0 < N(s,a) < 48\log\frac{N}{\delta}$, then we can immediately see that

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\Big| \le \|V\|_\infty \le \frac{1}{1-\gamma} \le \frac{48\log\frac{N}{\delta}}{(1-\gamma)N(s,a)}, \tag{167}$$

and hence the claim (150a) is valid. As a result, it suffices to focus on the case where

$$N(s,a) \ge 48\log\frac{N}{\delta}. \tag{168}$$

Note that the total number of pairs (s,a) with nonzero $N(s,a)$ cannot exceed N. Akin to (148), taking the Bernstein inequality together with (166) and invoking the union bound, we can demonstrate that with probability at least $1-4\delta$,

$$\begin{aligned}
\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\Big| &\le \sqrt{\frac{4\mathsf{Var}_{s,a}(V)\log\frac{N}{\delta}}{N(s,a)}} + \frac{2\|V\|_\infty \log\frac{N}{\delta}}{3N(s,a)} \\
&\le \sqrt{\frac{4\mathsf{Var}_{s,a}(V)\log\frac{N}{\delta}}{N(s,a)}} + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)N(s,a)} \qquad (169a) \\
\Big|\big(P_{s,a} - \widehat{P}_{s,a}\big)\big(\overline{V} \circ \overline{V}\big)\Big| &\le \sqrt{\frac{4\mathsf{Var}_{s,a}\big(\overline{V} \circ \overline{V}\big)\log\frac{N}{\delta}}{N(s,a)}} + \frac{2\|\overline{V} \circ \overline{V}\|_\infty \log\frac{N}{\delta}}{3N(s,a)} \\
&\le \sqrt{\frac{4\mathsf{Var}_{s,a}(V)\log\frac{N}{\delta}}{(1-\gamma)^2N(s,a)}} + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)^2N(s,a)} \qquad (169b)
\end{aligned}$$

hold simultaneously over all (s,a) with $N(s,a) > 0$. Note, however, that the Bernstein bounds in (169) involve the variance $\mathsf{Var}_{s,a}(V)$; we still need to connect $\mathsf{Var}_{s,a}(V)$ with its empirical estimate $\mathsf{Var}_{\widehat{P}_{s,a}}(V)$.

In the sequel, let us look at two cases separately.

- *Case 1:* $\mathsf{Var}_{s,a}(V) \le \frac{9\log\frac{N}{\delta}}{(1-\gamma)^2N(s,a)}$. In this case, our bound (169a) immediately leads to

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\Big| \le \frac{7\log\frac{N}{\delta}}{(1-\gamma)N(s,a)}. \tag{170}$$

- *Case 2:* $\mathsf{Var}_{s,a}(V) > \frac{9\log\frac{N}{\delta}}{(1-\gamma)^2N(s,a)}$. We first single out the following useful identity:

$$\widehat{P}_{s,a}\big(\overline{V} \circ \overline{V}\big) - \mathsf{Var}_{\widehat{P}_{s,a}}(V) = \widehat{P}_{s,a}\big(\overline{V} \circ \overline{V}\big) - \Big[\widehat{P}_{s,a}\big(V \circ V\big) - \big(\widehat{P}_{s,a}V\big)^2\Big]$$

$$= \widehat{P}_{s,a}(V \circ V) - 2\big(\widehat{P}_{s,a}V\big)\big(P_{s,a}V\big) + \big(P_{s,a}V\big)^2 - \Big[\widehat{P}_{s,a}(V \circ V) - \big(\widehat{P}_{s,a}V\big)^2\Big]$$
$$= \big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\big|^2. \tag{171}$$

Combining (171) with (169b) then implies that, with probability exceeding $1 - 4\delta$,

$$\mathsf{Var}_{s,a}(V) = P_{s,a}\big(\overline{V} \circ \overline{V}\big) = \big(P_{s,a} - \widehat{P}_{s,a}\big)\big(\overline{V} \circ \overline{V}\big) + \widehat{P}_{s,a}\big(\overline{V} \circ \overline{V}\big)$$
$$= \big(P_{s,a} - \widehat{P}_{s,a}\big)\big(\overline{V} \circ \overline{V}\big) + \Big\{\big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\big|^2 + \mathsf{Var}_{\widehat{P}_{s,a}}(V)\Big\} \tag{172}$$
$$\le \sqrt{\frac{4\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}}\sqrt{\mathsf{Var}_{s,a}(V)} + \big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\big|^2 + \mathsf{Var}_{\widehat{P}_{s,a}}(V) + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)}$$
$$\le \frac{2}{3}\mathsf{Var}_{s,a}(V) + \big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\big|^2 + \mathsf{Var}_{\widehat{P}_{s,a}}(V) + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)}, \tag{173}$$

where the second line arises from the identity (171), the penultimate inequality results from (169b), and the last inequality holds true due to the assumption $\mathsf{Var}_{s,a}(V) > \frac{9\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}$ in this case. Rearranging terms of the above inequality, we are left with

$$\mathsf{Var}_{s,a}(V) \le 3\big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\big|^2 + 3\mathsf{Var}_{\widehat{P}_{s,a}}(V) + \frac{2\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}$$

Taking this upper bound on $\mathsf{Var}_{s,a}(V)$ collectively with (169a) and using a little algebra lead to

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\Big| \le \sqrt{\frac{12\log\frac{N}{\delta}}{N(s,a)}}\big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\big| + \sqrt{\frac{12\mathsf{Var}_{\widehat{P}_{s,a}}(V)\log\frac{N}{\delta}}{N(s,a)}} + \frac{5\log\frac{N}{\delta}}{(1-\gamma)N(s,a)} \tag{174}$$

with probability at least $1 - 4\delta$. When $N(s,a) \ge 48\log\frac{N}{\delta}$ (cf. (168)), one has $\sqrt{\frac{12\log\frac{N}{\delta}}{N(s,a)}} \le 1/2$. Substituting this into (174) and rearranging terms, we arrive at

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\Big| \le \sqrt{\frac{48\mathsf{Var}_{\widehat{P}_{s,a}}(V)\log\frac{N}{\delta}}{N(s,a)}} + \frac{10\log\frac{N}{\delta}}{(1-\gamma)N(s,a)}$$

with probability at least $1 - 4\delta$.

Putting the above two cases together establishes the advertised bound (150a).

Proof of inequality (150b). It follows from (172) and (169a) that with probability at least $1 - 4\delta$,

$$\mathsf{Var}_{s,a}(V) \ge -\Big|\big(P_{s,a} - \widehat{P}_{s,a}\big)\big(\overline{V} \circ \overline{V}\big)\Big| + \mathsf{Var}_{\widehat{P}_{s,a}}(V)$$
$$\ge -\sqrt{\frac{4\mathsf{Var}_{s,a}(V)\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}} - \frac{2\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)} + \mathsf{Var}_{\widehat{P}_{s,a}}(V),$$

or equivalently,

$$\mathsf{Var}_{\widehat{P}_{s,a}}(V) \le \mathsf{Var}_{s,a}(V) + 2\sqrt{\frac{\mathsf{Var}_{s,a}(V)\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}} + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)}.$$

Invoke the elementary inequality $2xy \le x^2 + y^2$ to establish the claimed bound:

$$\mathsf{Var}_{\widehat{P}_{s,a}}(V) \le \mathsf{Var}_{s,a}(V) + \left(\mathsf{Var}_{s,a}(V) + \frac{\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}\right) + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)}$$
$$= 2\mathsf{Var}_{s,a}(V) + \frac{5\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)}.$$

B Proof of auxiliary lemmas: episodic finite-horizon MDPs

B.1 Proof of Lemma 3

(a) Let us begin by proving the claim (57a). Recall from our construction that $\mathcal{D}^{\mathsf{aux}}$ is composed of the second half of the sample trajectories, and hence for each $s \in \mathcal{S}$ and $1 \le h \le H$,

$$N_h^{\mathsf{aux}}(s) = \sum_{k=K/2+1}^{K} \mathbb{1}\{s_h^k = s\}$$

can be viewed as the sum of $K/2$ independent Bernoulli random variables, each with mean $d_h^{\mathsf{b}}(s)$. According to the union bound and the Bernstein inequality, we obtain

$$\begin{aligned}\mathbb{P}\left\{\exists(s,h) \in \mathcal{S}\times[H] : \left|N_h^{\mathsf{aux}}(s) - \frac{K}{2}d_h^{\mathsf{b}}(s)\right| \ge \tau\right\} &\le \sum_{s\in\mathcal{S},h\in[H]} \mathbb{P}\left\{\left|N_h^{\mathsf{aux}}(s) - \frac{K}{2}d_h^{\mathsf{b}}(s)\right| \ge \tau\right\} \\ &\le 2SH\exp\left(-\frac{\tau^2/2}{v_{s,h}+\tau/3}\right)\end{aligned}$$

for any $\tau \ge 0$, where

$$v_{s,h} \coloneqq \frac{K}{2}\mathsf{Var}\big(\mathbb{1}\{s_h^t = s\}\big) = \frac{Kd_h^{\mathsf{b}}(s)\big(1-d_h^{\mathsf{b}}(s)\big)}{2} \le \frac{Kd_h^{\mathsf{b}}(s)}{2}.$$

A little algebra then yields that with probability at least $1-2\delta$, one has

$$\left|N_h^{\mathsf{aux}}(s) - \frac{K}{2}d_h^{\mathsf{b}}(s)\right| \le \sqrt{4v_{s,h}\log\frac{HS}{\delta}} + \frac{2}{3}\log\frac{HS}{\delta} \le \sqrt{2Kd_h^{\mathsf{b}}(s)\log\frac{HS}{\delta}} + \log\frac{HS}{\delta} \tag{175}$$

simultaneously for all $s \in \mathcal{S}$ and all $1 \le h \le H$. The same argument also reveals that with probability exceeding $1-2\delta$,

$$\left|N_h^{\mathsf{main}}(s) - \frac{K}{2}d_h^{\mathsf{b}}(s)\right| \le \sqrt{2Kd_h^{\mathsf{b}}(s)\log\frac{HS}{\delta}} + \log\frac{HS}{\delta} \tag{176}$$

holds simultaneously for all $s \in \mathcal{S}$ and all $1 \le h \le H$. Combine (175) and (176) to show that

$$\left|N_h^{\mathsf{main}}(s) - N_h^{\mathsf{aux}}(s)\right| \le 2\sqrt{2Kd_h^{\mathsf{b}}(s)\log\frac{HS}{\delta}} + 2\log\frac{HS}{\delta} \tag{177}$$

for all $s \in \mathcal{S}$ and all $1 \le h \le H$.

To establish the claimed result (57a), we divide into two cases.

- *Case 1:* $N_h^{\mathsf{aux}}(s) \le 100\log\frac{HS}{\delta}$. By construction, it is easily seen that

$$N_h^{\mathsf{trim}}(s) = \max\left\{N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s)\log\frac{HS}{\delta}},\, 0\right\} = 0 \le N_h^{\mathsf{main}}(s). \tag{178}$$

- *Case 2:* $N_h^{\mathsf{aux}}(s) > 100\log\frac{HS}{\delta}$. In this case, invoking (175) reveals that

$$\frac{K}{2}d_h^{\mathsf{b}}(s) + \sqrt{2Kd_h^{\mathsf{b}}(s)\log\frac{HS}{\delta}} + \log\frac{HS}{\delta} \ge N_h^{\mathsf{aux}}(s) > 100\log\frac{HS}{\delta},$$

and hence one necessarily has

$$Kd_h^{\mathsf{b}}(s) \ge (9\sqrt{2})^2\log\frac{HS}{\delta} \ge 100\log\frac{HS}{\delta}. \tag{179}$$

In turn, this property (179) taken collectively with (148) ensures that

$$N_h^{\mathsf{aux}}(s) \geq \frac{K}{2} d_h^{\mathsf{b}}(s) - \sqrt{2K d_h^{\mathsf{b}}(s) \log \frac{HS}{\delta}} - \log \frac{HS}{\delta} \geq \frac{K}{4} d_h^{\mathsf{b}}(s). \tag{180}$$

Therefore, in the case with $N_h^{\mathsf{aux}}(s) > 100 \log \frac{HS}{\delta}$, we can demonstrate that

$$\begin{aligned} N_h^{\mathsf{trim}}(s) &= \max\left\{ N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s) \log \frac{HS}{\delta}},\, 0 \right\} = N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s) \log \frac{HS}{\delta}} \\ &\overset{\text{(i)}}{\leq} N_h^{\mathsf{aux}}(s) - 5\sqrt{K d_h^{\mathsf{b}}(s) \log \frac{HS}{\delta}} \overset{\text{(ii)}}{\leq} N_h^{\mathsf{aux}}(s) - \left\{ 2\sqrt{2K d_h^{\mathsf{b}}(s) \log \frac{HS}{\delta}} + 2 \log \frac{HS}{\delta} \right\} \\ &\overset{\text{(iii)}}{\leq} N_h^{\mathsf{main}}(s), \end{aligned} \tag{181}$$

where (i) comes from Condition (180), (ii) is valid under the condition (179), and (iii) holds true with probability at least $1 - 2\delta$ due to the inequality (177).

Putting the above two cases together establishes the claim (57a).

(b) We now turn to the second claim (57b). Towards this, we first claim that the following bound holds simultaneously for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ with probability exceeding $1 - 2\delta$:

$$N_h^{\mathsf{trim}}(s, a) \geq N_h^{\mathsf{trim}}(s) \pi_h^{\mathsf{b}}(a \,|\, s) - \sqrt{4 N_h^{\mathsf{trim}}(s) \pi_h^{\mathsf{b}}(a \,|\, s) \log \frac{KH}{\delta}} - \log \frac{KH}{\delta}. \tag{182}$$

Let us take this claim as given for the moment, and return to establish it towards the end of this section. We shall discuss the following two cases separately.

- If $K d_h^{\mathsf{b}}(s, a) = K d_h^{\mathsf{b}}(s) \pi_h^{\mathsf{b}}(a \,|\, s) > 1600 \log \frac{KH}{\delta}$, then it follows from (180) (with slight modification) that

$$N_h^{\mathsf{aux}}(s) \geq \frac{K}{4} d_h^{\mathsf{b}}(s) \geq 400 \log \frac{KH}{\delta}. \tag{183}$$

  This property together with the definition of $N_h^{\mathsf{trim}}(s)$ in turn allows us to derive

$$\begin{aligned} N_h^{\mathsf{trim}}(s) &\geq N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s) \log \frac{KH}{\delta}} \geq \frac{K}{4} d_h^{\mathsf{b}}(s) - 10\sqrt{\frac{K}{4} d_h^{\mathsf{b}}(s) \log \frac{KH}{\delta}} \\ &\geq \frac{K}{8} d_h^{\mathsf{b}}(s), \end{aligned}$$

  and as a result,

$$N_h^{\mathsf{trim}}(s) \pi_h^{\mathsf{b}}(a \,|\, s) \geq \frac{K}{8} d_h^{\mathsf{b}}(s) \pi_h^{\mathsf{b}}(a \,|\, s) = \frac{K}{8} d_h^{\mathsf{b}}(s, a) \geq 200 \log \frac{KH}{\delta},$$

  where the last inequality arises from the assumption of this case. Taking this lower bound with (182) implies that

$$\begin{aligned} N_h^{\mathsf{trim}}(s, a) &\geq \frac{K}{8} d_h^{\mathsf{b}}(s, a) - \sqrt{\frac{K}{2} d_h^{\mathsf{b}}(s, a) \log \frac{KH}{\delta}} - \log \frac{KH}{\delta} \\ &\geq \frac{K}{8} d_h^{\mathsf{b}}(s, a) - 2\sqrt{K d_h^{\mathsf{b}}(s, a) \log \frac{KH}{\delta}}. \end{aligned}$$

- If $K d_h^{\mathsf{b}}(s, a) \leq 1600 \log \frac{KH}{\delta}$, then one can easily verify that

$$\frac{K}{8} d_h^{\mathsf{b}}(s, a) - 5\sqrt{K d_h^{\mathsf{b}}(s, a) \log \frac{KH}{\delta}} \leq 0 \leq N_h^{\mathsf{trim}}(s, a).$$

Putting these two cases together concludes the proof, provided that the claim (182) is valid.

Proof of inequality (182). Let us look at two cases separately.

- If $N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) \le 4\log\frac{KH}{\delta}$, then the right-hand side of (182) is negative, and hence the claim (182) holds trivially.
- We then turn attention to the following set:

$$\mathcal{A}_{\mathsf{large}} \coloneqq \Big\{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H] \;\Big|\; N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) > 4\log\frac{KH}{\delta}\Big\}. \tag{184}$$

Recognizing that

$$\begin{aligned}\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) &= \sum_{(s,h)\in\mathcal{S}\times[H]} N_h^{\mathsf{trim}}(s)\sum_{a\in\mathcal{A}}\pi_h^{\mathsf{b}}(a\,|\,s) = \sum_{(s,h)\in\mathcal{S}\times[H]} N_h^{\mathsf{trim}}(s)\\ &\le \sum_{(s,h)\in\mathcal{S}\times[H]} N_h^{\mathsf{aux}}(s) = \frac{KH}{2},\end{aligned}$$

we can immediately bound the cardinality of $\mathcal{A}_{\mathsf{large}}$ as follows:

$$\big|\mathcal{A}_{\mathsf{large}}\big| < \frac{\sum_{(s,a,h)} N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s)}{4\log\frac{KH}{\delta}} \le KH/2. \tag{185}$$

Additionally, it follows from our construction that: conditional on $N_h^{\mathsf{trim}}(s)$, $N_h^{\mathsf{main}}(s)$ and the high-probability event (57a), $N_h^{\mathsf{trim}}(s,a)$ can be viewed as the sum of $\min\big\{N_h^{\mathsf{trim}}(s), N_h^{\mathsf{main}}(s)\big\} = N_h^{\mathsf{trim}}(s)$ independent Bernoulli random variables each with mean $\pi_h(a\,|\,s)$. As a result, repeating the Bernstein-type argument in (148) on the event (57a) reveals that, with probability at least $1-2\delta/(KH)$,

$$N_h^{\mathsf{trim}}(s,a) \ge N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) - \sqrt{4N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s)\log\frac{KH}{\delta}} - \log\frac{KH}{\delta} \tag{186}$$

for any fixed triple (s,a,h). Taking the union bound over all $(s,a,h)\in\mathcal{A}_{\mathsf{large}}$ and using the bound (185) imply that with probability exceeding $1-\delta$, (186) holds simultaneously for all $(s,a,h)\in\mathcal{A}_{\mathsf{large}}$.

Combining the above two cases allows one to conclude that with probability at least $1-\delta$, the advertised property (182) holds simultaneously for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$.

B.2 Proof of the instance-dependent statistical bound (65)

To establish relation (65), we make use of relation (171) as follows: for any $1\le h\le H$,

$$\begin{aligned}&\big\langle d_h^\star, V_h^\star - V_h^{\widehat{\pi}}\big\rangle \le 2\sum_{j=h}^{H}\big\langle d_j^\star, b_j^\star\big\rangle\\ &\le 2\sum_{j=h}^{H}\sum_s d_j^\star(s)\sqrt{\frac{32c_{\mathsf{b}}\log\frac{NH}{\delta}}{Kd_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(\widehat{V}_{j+1})} + \frac{192c_{\mathsf{b}}H^2SC^\star_{\mathsf{clipped}}\log\frac{NH}{\delta}}{K}\\ &\le 12\sum_{j=h}^{H}\sum_s d_j^\star(s)\sqrt{\frac{c_{\mathsf{b}}\log\frac{NH}{\delta}}{Kd_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(V^\star_{j+1})} + 20\Big(\frac{c_{\mathsf{b}}H^3SC^\star\log\frac{NH}{\delta}}{K}\Big)^{3/4},\end{aligned} \tag{187}$$

provided that $K\ge 100c_{\mathsf{b}}HSC^\star\log\frac{NH}{\delta}$. To see why the last inequality in (187) holds, it suffices to observe that

$$\sum_s d_j^\star(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(\widehat{V}_{j+1})}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \sum_s d_j^\star(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(V_{j+1}^\star)}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}} + \sum_s d_j^\star(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(\widehat{V}_{j+1}-V_{j+1}^\star)}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}} \\
&\overset{\text{(ii)}}{\leq} \sum_s d_j^\star(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(V_{j+1}^\star)}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}} + \sum_s d_j^\star(s)\sqrt{\frac{H\cdot\big\langle P_{j,s,\pi_j^\star(s)},\widehat{V}_{j+1}-V_{j+1}^\star\big\rangle}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}} \\
&\overset{\text{(iii)}}{\leq} \sum_s d_j^\star(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(V_{j+1}^\star)}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}} + \sqrt{S}\sqrt{H\sum_s \frac{\big(d_j^\star(s)\big)^2}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}\big\langle P_{j,s,\pi_j^\star(s)},\widehat{V}_{j+1}-V_{j+1}^\star\big\rangle} \\
&\overset{\text{(iv)}}{\leq} \sum_s d_j^\star(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(V_{j+1}^\star)}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}} + \sqrt{HSC^\star\sum_s d_j^\star(s)\big\langle P_{j,s,\pi_j^\star(s)},\widehat{V}_{j+1}-V_{j+1}^\star\big\rangle} \\
&= \sum_s d_j^\star(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(V_{j+1}^\star)}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}} + \sqrt{HSC^\star\sum_s d_{j+1}^\star(s)\big[V_{j+1}^\star(s)-\widehat{V}_{j+1}(s)\big]} \\
&\overset{\text{(v)}}{\leq} \sum_s d_j^\star(s)\sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(V_{j+1}^\star)}{d_j^{\mathsf{b}}\big(s,\pi_j^\star(s)\big)}} + \sqrt{HSC^\star\cdot 80\sqrt{\frac{2c_{\mathsf{b}}H^3SC^\star\log\frac{NH}{\delta}}{K}}},
\end{aligned}
\tag{188}
$$

where (i) holds due to the elementary inequality $\sqrt{\mathsf{Var}(X+Y)}\leq\sqrt{\mathsf{Var}(X)}+\sqrt{\mathsf{Var}(Y)}$; (ii) follows since

$$
\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}(\widehat{V}_{j+1}-V_{j+1}^\star)\leq\big\langle P_{j,s,\pi_j^\star(s)},(\widehat{V}_{j+1}-V_{j+1}^\star)^2\big\rangle\leq H\big\langle P_{j,s,\pi_j^\star(s)},\widehat{V}_{j+1}-V_{j+1}^\star\big\rangle,
$$

which comes from the fact that $V_{j+1}^\star\geq\widehat{V}_{j+1}\geq 0$ and $\|V_{j+1}^\star\|_\infty\leq H$; (iii) invokes the Cauchy-Schwarz inequality; (iv) makes use of the definition of $C^\star_{\mathsf{clipped}}$; and (v) is obtained by applying (116) of Theorem 6.

C Proof of minimax lower bounds

C.1 Preliminary facts

For any two distributions P and Q, we denote by $\mathsf{KL}(P\,\|\,Q)$ the Kullback-Leibler (KL) divergence of P and Q. Letting $\mathsf{Ber}(p)$ be the Bernoulli distribution with mean p, we also introduce

$$
\mathsf{KL}(p\,\|\,q)\coloneqq p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}\qquad\text{and}\qquad\chi^2(p\,\|\,q)\coloneqq\frac{(p-q)^2}{q}+\frac{(p-q)^2}{1-q},
\tag{189}
$$

which represent respectively the KL divergence and the chi-square divergence of $\mathsf{Ber}(p)$ from $\mathsf{Ber}(q)$ (Tsybakov, 2009). We make note of the following useful properties about the KL divergence.

Lemma 10. *For any $p,q\in\big[\frac{1}{2},1\big)$ and $p>q$, it holds that*

$$
\mathsf{KL}(p\,\|\,q)\leq\mathsf{KL}(q\,\|\,p)\leq\chi^2(q\,\|\,p)=\frac{(p-q)^2}{p(1-p)}.
\tag{190}
$$

Proof. The second inequality in (190) is a well-known relation between KL divergence and chi-square divergence; see Tsybakov (2009, Lemma 2.7). As a result, it suffices to justify the first inequality. Towards this end, let us introduce $a=\frac{p+q}{2}\in\big[\frac{1}{2},1\big]$ and $b=\frac{p-q}{2}\in\big[0,\frac{1}{4}\big]$, which allow us to re-parameterize (p,q) as $p=a+b$ and $q=a-b$. The definition (189) together with a little algebra gives

$$
\begin{aligned}
\mathsf{KL}(p\,\|\,q)-\mathsf{KL}(q\,\|\,p)&=(p+q)\log\frac{p}{q}+(2-p-q)\log\frac{1-p}{1-q}\\
&=2a\log\left(\frac{a+b}{a-b}\right)+2(1-a)\log\frac{1-a-b}{1-a+b}\eqqcolon g(a,b).
\end{aligned}
$$

Taking the derivative w.r.t. b yields

$$\frac{\partial g(a,b)}{\partial b} = 2a\left\{\frac{1}{a+b}+\frac{1}{a-b}\right\} - 2\,(1-a)\left\{\frac{1}{1-a+b}+\frac{1}{1-a-b}\right\} = f(a) - f(1-a) \le 0,$$

with $f(x) \coloneqq \frac{2x}{x+b} + \frac{2x}{x-b}$ (for $x > b$). Here, the last inequality follows since $f(\cdot)$ is a decreasing function and that $a \ge 1-a$. This implies that $g(a,b)$ is non-increasing in $b \ge 0$ for any given a, which in turn leads to

$$\mathsf{KL}(p \,\|\, q) - \mathsf{KL}(q \,\|\, p) = g(a,b) \le g(a,0) = 0$$

as claimed. □

C.2 Proof of Theorem 2

We now construct some hard problem instances and use them to establish the minimax lower bounds claimed in Theorem 2. It is assumed throughout this subsection that

$$\frac{2}{3} \le \gamma < 1 \qquad \text{and} \qquad \frac{14(1-\gamma)\varepsilon}{\gamma} \le \frac{1}{2}. \tag{191}$$

C.2.1 Construction of hard problem instances

Construction of the hard MDPs. Let us introduce two MDPs $\{\mathcal{M}_\theta = (\mathcal{S}, \mathcal{A}, P_\theta, r, \gamma) \mid \theta \in \{0,1\}\}$ parameterized by θ, which involve S states and 2 actions as follows:

$$\mathcal{S} = \{0, 1, \ldots, S-1\} \qquad \text{and} \qquad \mathcal{A} = \{0,1\}.$$

We single out a crucial state distribution (supported on the state subset $\{0,1\}$) as follows:

$$\mu(s) = \frac{1}{CS}\mathbb{1}\{s=0\} + \Big(1 - \frac{1}{CS}\Big)\mathbb{1}\{s=1\} \tag{192}$$

for some quantity $C > 0$ obeying

$$\frac{1}{CS} \le \frac{1}{4\gamma}. \tag{193}$$

We shall make clear the relation between C and the concentrability coefficient $C^\star_{\mathsf{clipped}}$ shortly (see (203)). Armed with this distribution, we are ready to define the transition kernel P_θ of the MDP $\mathcal{M}_\theta$ as follows:

$$P_\theta(s' \,|\, s,a) = \begin{cases} p\mathbb{1}\{s'=0\} + (1-p)\mu(s') & \text{for} \quad (s,a) = (0,\theta), \\ q\mathbb{1}\{s'=0\} + (1-q)\mu(s') & \text{for} \quad (s,a) = (0,1-\theta), \\ \mathbb{1}\{s'=1\} & \text{for} \quad (s,a) = (1,0), \\ (2\gamma-1)\mathbb{1}\{s'=1\} + 2(1-\gamma)\mu(s') & \text{for} \quad (s,a) = (1,1), \\ \gamma\mathbb{1}(s'=s\} + (1-\gamma)\mu(s') & \text{for} \quad s > 1, \end{cases} \tag{194}$$

where the parameters p and q are chosen to be

$$p = \gamma + \frac{14(1-\gamma)^2\varepsilon}{\gamma}, \qquad q = \gamma - \frac{14(1-\gamma)^2\varepsilon}{\gamma}. \tag{195}$$

In view of the assumptions (191), one has

$$p > q \ge \gamma - \frac{1-\gamma}{2} \ge \frac{1}{2}. \tag{196}$$

As can be clearly seen from the construction, if the MDP is initialized to either state 0 or state 1, then it will never leave the state subset $\{0,1\}$. In addition, the reward function for any MDP $\mathcal{M}_\theta$ is chosen to be

$$r(s,a) = \begin{cases} 1 & \text{for} \quad s = 0, \\ \frac{1}{2} & \text{for} \quad (s,a) = (1,0), \\ 0 & \text{for} \quad (s,a) = (1,1), \\ 0 & \text{for} \quad s > 1, \end{cases} \tag{197}$$

where the reward gained in state 0 is clearly higher than that in other states.

Value functions and optimal policies. Next, let us take a moment to compute the value functions of the constructed MDPs and identify the optimal policies. For notational clarity, for the MDP $\mathcal{M}_\theta$ with $\theta \in \{0,1\}$, we denote by $\pi_\theta^\star$ the optimal policy, and let V_θ^π (resp. $V_\theta^\star$) represent the value function of policy π (resp. $\pi_\theta^\star$). The lemma below collects several useful properties about the value functions and the optimal policies; the proof is deferred to Appendix C.2.3.

Lemma 11. *Consider any $\theta \in \{0,1\}$ and any policy π. One has*

$$V_\theta^\pi(0) = \frac{1+\gamma(1-x_{\pi,\theta})\mu(1)V_\theta^\pi(1)}{1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)} = V_\theta^\pi(1) + \frac{1-(1-\gamma)V_\theta^\pi(1)}{1-\gamma\mu(0)-\gamma\mu(1)x_{\pi,\theta}}, \tag{198}$$

where we define

$$x_{\pi,\theta} \coloneqq p\pi(\theta\,|\,0) + q\pi(1-\theta\,|\,0). \tag{199}$$

In addition, the optimal policy $\pi_\theta^\star$ and the optimal value function obey

$$\pi_\theta^\star(\theta\,|\,0) = 1, \qquad \pi_\theta^\star(0\,|\,1) = 1, \qquad \text{and} \qquad V_\theta^\star(1) = \frac{1}{2(1-\gamma)}. \tag{200}$$

Construction of the batch dataset. Given any constructed MDP $\mathcal{M}_\theta$, we generate a dataset containing N i.i.d. samples $\{(s_i, a_i, s_i')\}_{1\le i\le N}$ according to (17), where the initial state distribution ρ^{b} and behavior policy π^{b} are chosen to be:

$$\rho^{\mathsf{b}}(s) = \mu(s) \qquad \text{and} \qquad \pi^{\mathsf{b}}(a\,|\,s) = 1/2, \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A},$$

with μ denoting the distribution defined in (192). Interestingly, the occupancy state distribution of this dataset coincides with μ, in the sense that

$$d^{\mathsf{b}}(s) = \mu(s) \qquad \text{and} \qquad d^{\mathsf{b}}(s,a) = \mu(s)/2, \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{201}$$

Moreover, letting us choose the test distribution ρ in a way that

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{if } s > 0. \end{cases} \tag{202}$$

we can also characterize the single-policy clipped concentrability coefficient $C^\star_{\mathsf{clipped}}$ of the dataset w.r.t. the constructed MDP $\mathcal{M}_\theta$ as follows

$$C^\star_{\mathsf{clipped}} = 2C. \tag{203}$$

The proof of the claims (201) and (203) can be found in Appendix C.2.4.

C.2.2 Establishing the minimax lower bound

Equipped with the above construction, we are ready to develop our lower bounds. We remind the reader of the test distribution ρ chosen in (202), and hence we need to control $\langle \rho, V_\theta^\star - V_\theta^{\widehat{\pi}}\rangle = V_\theta^\star(0) - V_\theta^{\widehat{\pi}}(0)$ with $\widehat{\pi}$ representing a policy estimate (computed based on the batch dataset).

Step 1: converting $\widehat{\pi}$ into an estimate $\widehat{\theta}$ of θ. Consider first an arbitrary policy π. By combining the definition (199) with the properties (200), we see that $x_{\pi_\theta^\star,\theta} = p$, which together with (198) gives

$$\begin{aligned}\langle \rho, V_\theta^\star - V_\theta^\pi\rangle = V_\theta^\star(0) - V_\theta^\pi(0) &= \frac{1+\gamma(1-p)\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)p+\mu(0)\big)} - \frac{1+\gamma(1-x_{\pi,\theta})\mu(1)V_\theta^\pi(1)}{1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)} \\ &\ge \frac{1+\gamma(1-p)\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)p+\mu(0)\big)} - \frac{1+\gamma(1-x_{\pi,\theta})\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)}\end{aligned}$$

$$\geq \frac{21\varepsilon}{8}\big(1-\pi(\theta\,|\,0)\big). \tag{204}$$

Here, the second line holds since $V_\theta^\pi \leq V_\theta^\star$, and the last inequality will be established in Appendix C.2.4.

Denoting by $\mathbb{P}_\theta$ the probability distribution when the MDP is $\mathcal{M}_\theta$, suppose for the moment that the policy estimate $\widehat{\pi}$ achieves

$$\mathbb{P}_\theta\big\{\big\langle \rho, V_\theta^\star - V_\theta^{\widehat{\pi}}\big\rangle \leq \varepsilon\big\} \geq \frac{7}{8},$$

then in view of (204), one necessarily has $\widehat{\pi}(\theta\,|\,0) \geq \frac{13}{21}$ with probability at least 7/8. If this were true, then we could then construct the following estimate $\widehat{\theta}$ for θ:

$$\widehat{\theta} = \arg\max_a \widehat{\pi}(a\,|\,0), \tag{205}$$

which would necessarily satisfy

$$\mathbb{P}_\theta\big(\widehat{\theta}=\theta\big) \geq \mathbb{P}_\theta\big\{\widehat{\pi}(\theta\,|\,0) > 1/2\big\} \geq \mathbb{P}_\theta\bigg\{\widehat{\pi}(\theta\,|\,0) \geq \frac{13}{21}\bigg\} \geq \frac{7}{8}. \tag{206}$$

In what follows, we would like to show that (206) cannot happen — i.e., one cannot possibly find such a good estimator for θ — without a sufficient number of samples.

Step 2: probability of error in testing two hypotheses. The next step lies in studying the feasibility of differentiating two hypotheses $\theta = 0$ and $\theta = 1$. Define the minimax probability of error as follows

$$p_{\mathrm{e}} \coloneqq \inf_{\psi} \max\big\{\mathbb{P}_0(\psi \neq 0),\, \mathbb{P}_1(\psi \neq 1)\big\}, \tag{207}$$

where the infimum is taken over all possible tests ψ (based on the batch dataset in hand). Letting μ_θ^{b} denote the distribution of a sample (s_i, a_i, s_i') under the MDP $\mathcal{M}_\theta$ and recalling that the samples are independently generated, one can demonstrate that

$$\begin{aligned} p_{\mathrm{e}} &\geq \frac{1}{4}\exp\Big(-N\mathsf{KL}\big(\mu_0^{\mathsf{b}} \,\|\, \mu_1^{\mathsf{b}}\big)\Big) \\ &= \frac{1}{4}\exp\bigg\{-\frac{1}{2}N\mu(0)\Big(\mathsf{KL}\big(P_0(\cdot\,|\,0,0) \,\|\, P_1(\cdot\,|\,0,0)\big) + \mathsf{KL}\big(P_0(\cdot\,|\,0,1) \,\|\, P_1(\cdot\,|\,0,1)\big)\Big)\bigg\}. \end{aligned} \tag{208}$$

Here, the first inequality results from Tsybakov (2009, Theorem 2.2) and the additivity property of the KL divergence (cf. Tsybakov (2009, Page 85)), and the second line holds true since

$$\begin{aligned} \mathsf{KL}(\mu_0^{\mathsf{b}} \,\|\, \mu_1^{\mathsf{b}}) &= \sum_{s,a,s'} \mu(s)\pi^{\mathsf{b}}(a\,|\,s)P_0(s'\,|\,s,a)\log\frac{\mu(s)\pi^{\mathsf{b}}(a\,|\,s)P_0(s'\,|\,s,a)}{\mu(s)\pi^{\mathsf{b}}(a\,|\,s)P_1(s'\,|\,s,a)} \\ &= \frac{1}{2}\mu(0)\sum_a\sum_{s'} P_0(s'\,|\,0,a)\log\frac{P_0(s'\,|\,0,a)}{P_1(s'\,|\,0,a)} \\ &= \frac{1}{2}\mu(0)\sum_a \mathsf{KL}\big(P_0(\cdot\,|\,0,a) \,\|\, P_1(\cdot\,|\,0,a)\big), \end{aligned}$$

where the second line is valid since $P_0(\cdot\,|\,s,a)$ and $P_1(\cdot\,|\,s,a)$ differ only when $s=0$.

Next, we turn attention to the KL divergence of interest. Recall that

$$P_0(0\,|\,0,0) = \bigg(1-\frac{1}{CS}\bigg)p + \frac{1}{CS}, \qquad P_1(0\,|\,0,0) = \bigg(1-\frac{1}{CS}\bigg)q + \frac{1}{CS}.$$

Given that $p \geq q \geq 1/2$ (see (196)), we can apply Lemma 10 to arrive at

$$\mathsf{KL}\big(P_0(\cdot\,|\,0,0) \,\|\, P_1(\cdot\,|\,0,0)\big) = \mathsf{KL}\left(\bigg(1-\frac{1}{CS}\bigg)p + \frac{1}{CS} \,\|\, \bigg(1-\frac{1}{CS}\bigg)q + \frac{1}{CS}\right)$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \frac{\big(1-\frac{1}{CS}\big)^2 (p-q)^2}{\big((1-\frac{1}{CS})\,p+\frac{1}{CS}\big)\big(1-p-(1-p)\frac{1}{CS}\big)} \\
&\leq \frac{\big(1-\frac{1}{CS}\big)^2 (p-q)^2}{p\big((1-p)\big(1-\frac{1}{CS}\big)\big)} \\
&\overset{\text{(ii)}}{=} \frac{784(1-\gamma)^4\varepsilon^2}{\gamma^2\left(\gamma+\frac{14(1-\gamma)^2\varepsilon}{\gamma}\right)\left(1-\gamma-\frac{14(1-\gamma)^2\varepsilon}{\gamma}\right)} \\
&\overset{\text{(iii)}}{\leq} \frac{1568(1-\gamma)^4\varepsilon^2}{\gamma^3(1-\gamma)} \overset{\text{(iv)}}{\leq} 12544(1-\gamma)^3\varepsilon^2,
\end{aligned}
$$

where (i) arises from Lemma 10, (ii) follows from the definitions of p and q (195), (iii) holds true as long as $\frac{14(1-\gamma)^2\varepsilon}{\gamma} \leq \frac{1-\gamma}{2}$, and (iv) results from the assumption $\gamma \in [\frac{1}{2}, 1)$. Evidently, the same upper bound holds for $\mathsf{KL}\big(P_0(\cdot\,|\,0,1) \,\|\, P_1(\cdot\,|\,0,1)\big)$ as well. Substitution back into (208) reveals that: if the sample size does not exceed

$$
N \leq \frac{CS\log 2}{12544(1-\gamma)^3\varepsilon^2} = \frac{C^\star_{\mathsf{clipped}} S\log 2}{25088(1-\gamma)^3\varepsilon^2}, \tag{209}
$$

then one necessarily has

$$
p_{\mathrm{e}} \geq \frac{1}{4}\exp\Big(-12544N\mu(0)(1-\gamma)^3\varepsilon^2\Big) = \frac{1}{4}\exp\Big(-\frac{12544N(1-\gamma)^3\varepsilon^2}{CS}\Big) \geq \frac{1}{8}. \tag{210}
$$

Step 3: putting all this together. To finish up, suppose that there exists an estimator $\widehat{\pi}$ such that

$$
\mathbb{P}_0\big\{\big\langle \rho, V_0^\star - V_0^{\widehat{\pi}}\big\rangle > \varepsilon\big\} < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1\big\{\big\langle \rho, V_0^\star - V_0^{\widehat{\pi}}\big\rangle > \varepsilon\big\} < \frac{1}{8}.
$$

Then in view of our arguments in Step 1, the estimator $\widehat{\theta}$ defined in (205) must satisfy

$$
\mathbb{P}_0\big(\widehat{\theta} \neq \theta\big) < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1\big(\widehat{\theta} \neq \theta\big) < \frac{1}{8}.
$$

This, however, cannot possibly happen under the sample size condition (209); otherwise it contradicts the lower bound (210).

C.2.3 Proof of Lemma 11

To begin with, for any policy π, the value function of state 0 obeys

$$
\begin{aligned}
V_\theta^\pi(0) &= \mathop{\mathbb{E}}_{a\sim\pi(\cdot\,|\,0)}\Big[r(0,a) + \gamma\sum_{s'} P_\theta(s'\,|\,0,a)V_\theta^\pi(s')\Big] \\
&= 1 + \gamma\pi(\theta\,|\,0)\Big[\big(p+(1-p)\mu(0)\big)V_\theta^\pi(0) + (1-p)\mu(1)V_\theta^\pi(1)\Big] \\
&\quad + \gamma\pi(1-\theta\,|\,0)\Big[\big(q+(1-q)\mu(0)\big)V_\theta^\pi(0) + (1-q)\mu(1)V_\theta^\pi(1)\Big] \\
&= 1 + \gamma\Big[p\pi(\theta\,|\,0) + q\pi(1-\theta\,|\,0) + \mu(0) - p\pi(\theta\,|\,0)\mu(0) - q\pi(1-\theta\,|\,0)\mu(0)\Big]V_\theta^\pi(0) \\
&\quad + \gamma\mu(1)\Big[1 - p\pi(\theta\,|\,0) - q\pi(1-\theta\,|\,0)\Big]V_\theta^\pi(1) \\
&\overset{\text{(i)}}{=} 1 + \gamma\Big[x_{\pi,\theta} + (1-x_{\pi,\theta})\mu(0)V_\theta^\pi(0) + (1-x_{\pi,\theta})\mu(1)V_\theta^\pi(1)\Big] \\
&\overset{\text{(ii)}}{=} 1 + \gamma\Big[\big(\mu(1)x_{\pi,\theta} + \mu(0)\big)V_\theta^\pi(0) + (1-x_{\pi,\theta})\mu(1)V_\theta^\pi(1)\Big],
\end{aligned} \tag{211}
$$

where in (i) we have defined the following quantity

$$
x_{\pi,\theta} = p\pi(\theta\,|\,0) + q\pi(1-\theta\,|\,0) = q + (p-q)\pi(\theta\,|\,0), \tag{212}
$$

and (ii) relies on the fact that $\mu(0)+\mu(1)=1$. Rearranging terms in (211), we are left with

$$V_\theta^\pi(0)=\frac{1+\gamma(1-x_{\pi,\theta})\mu(1)V_\theta^\pi(1)}{1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)}=V_\theta^\pi(1)+\frac{1-(1-\gamma)V_\theta^\pi(1)}{1-\gamma\mu(0)-\gamma\mu(1)x_{\pi,\theta}}. \tag{213}$$

Additionally, the value function of state 1 can be calculated as

$$\begin{aligned}
V_\theta^\pi(1)&=\pi(0\,|\,1)\left(\frac{1}{2}+\gamma V_\theta^\pi(1)\right)+\pi(1\,|\,1)\gamma\Big[\big((2\gamma-1)+2(1-\gamma)\mu(1)\big)V_\theta^\pi(1)+2(1-\gamma)\mu(0)V_\theta^\pi(0)\Big]\\
&=\pi(0\,|\,1)\left(\frac{1}{2}+\gamma V_\theta^\pi(1)\right)+\pi(1\,|\,1)\gamma\left[\left(1-\frac{2(1-\gamma)}{CS}\right)V_\theta^\pi(1)+\frac{2(1-\gamma)}{CS}V_\theta^\pi(0)\right] \qquad (214)\\
&\overset{\text{(i)}}{\le}\pi(0\,|\,1)\left(\frac{1}{2}+\gamma V_\theta^\pi(1)\right)+\pi(1\,|\,1)\gamma\left[\left(1-\frac{2(1-\gamma)}{CS}\right)V_\theta^\pi(1)+\frac{2(1-\gamma)}{CS}\frac{1}{1-\gamma}\right]\\
&\overset{\text{(ii)}}{\le}\pi(0\,|\,1)\left(\frac{1}{2}+\gamma V_\theta^\pi(1)\right)+\pi(1\,|\,1)\left[\frac{1}{2}+\gamma\left(1-\frac{2(1-\gamma)}{CS}\right)V_\theta^\pi(1)\right]\\
&=\frac{1}{2}+\gamma V_\theta^\pi(1)-\frac{2\gamma(1-\gamma)}{CS}V_\theta^\pi(1)\pi(1\,|\,1), \qquad (215)
\end{aligned}$$

where (i) arises from the elementary property $0\le V_\theta^\pi(s)\le\frac{1}{1-\gamma}$ for any π and $s\in\mathcal{S}$, and (ii) comes from the assumption (193). The above observation reveals several facts:

- If we take $\pi(0\,|\,1)=1$, then (214) tells us that

$$V_\theta^\pi(1)=\frac{1}{2}+\gamma V_\theta^\pi(1)\qquad\Longrightarrow\qquad V_\theta^\pi(1)=\frac{1}{2(1-\gamma)}. \tag{216}$$

- It also follows from (215) that for any policy π, one has

$$V_\theta^\pi(1)\le\frac{1}{2}+\gamma V_\theta^\pi(1)\qquad\Longrightarrow\qquad V_\theta^\pi(1)\le\frac{1}{2(1-\gamma)}. \tag{217}$$

These two facts taken collectively imply that the optimal policy and the optimal value function obey

$$\pi_\theta^\star(0\,|\,1)=1\qquad\text{and}\qquad V_\theta^\star(1)=\frac{1}{2(1-\gamma)}. \tag{218}$$

Next, we have learned from (213) that

$$V_\theta^\star(0)=V_\theta^\star(1)+\frac{1-(1-\gamma)V_\theta^\star(1)}{1-\gamma\mu(0)-\gamma\mu(1)x_{\pi_\theta^\star,\theta}}.$$

Note that $1-(1-\gamma)V_\theta^\star(1)\ge 1-(1-\gamma)\frac{1}{1-\gamma}=0$. Since the function

$$g(x)=V_\theta^\star(1)+\frac{1-(1-\gamma)V_\theta^\star(1)}{1-\gamma\mu(0)-\gamma\mu(1)x}$$

is increasing in x and that $x_{\pi,\theta}$ (cf. (212)) is increasing in $\pi(\theta\,|\,0)$ (given that $p\ge q$), one can easily see that the optimal policy obeys

$$\pi_\theta^\star(\theta\,|\,0)=1. \tag{219}$$

C.2.4 Proof of auxiliary properties

Proof of claim (201). We begin by proving the property (201). Towards this, let us abuse the notation by considering a MDP trajectory denoted by $\{(s_t,a_t)\}_{t\ge0}$, and suppose that it starts from $s_0\sim\rho^{\mathsf{b}}=\mu$. It can be straightforwardly calculated that

$$\mathbb{P}\{s_1=0\}=\sum\nolimits_s\mu(s)\Big\{\pi^{\mathsf{b}}(0\,|\,s)\mathbb{P}\{s_1=0\mid s_0=s,a_0=0\}+\pi^{\mathsf{b}}(1\,|\,s)\mathbb{P}\{s_1=0\mid s_0=s,a_0=1\}\Big\}$$

$$= \mu(0)\left\{\frac{1}{2}P_\theta(0\,|\,0,0)+\frac{1}{2}P_\theta(0\,|\,0,1)\right\}+\mu(1)\left\{\frac{1}{2}P_\theta(0\,|\,1,0)+\frac{1}{2}P_\theta(0\,|\,1,1)\right\}$$
$$= \mu(0)\left\{\gamma+(1-\gamma)\mu(0)\right\}+\mu(1)\left\{(1-\gamma)\mu(0)\right\}=\mu(0),$$

where the last identity holds since $\mu(0)+\mu(1)=1$. Similarly, one can derive $\mathbb{P}\{s_1=1\}=\mu(1)$, thus indicating that $s_1\sim\mu$. Repeating this analysis reveals that $s_t\sim\mu$ for any $t\geq 0$. Consequently, one has

$$d^{\mathsf{b}}(s)=(1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\big(s_t=s\mid s_0\sim\rho^{\mathsf{b}};\pi^{\mathsf{b}}\big)\right]=\mu(s),\qquad\forall s\in\mathcal{S}.$$

Additionally, it it observed that

$$d^{\mathsf{b}}(s,a)=d^{\mathsf{b}}(s)\pi^{\mathsf{b}}(a\,|\,s)=\mu(s)/2. \tag{220}$$

Proof of claim (203)**.** Consider the MDP $\mathcal{M}_\theta$, whose optimal policy $\pi^\star_\theta$ satisfies $\pi^\star_\theta(\theta\,|\,0)=1$ (see Lemma 11). Let us generate a MDP trajectory denoted by $\{(s_t,a_t)\}_{t\geq 0}$ with $a_t\sim\pi^\star_\theta(\cdot\,|\,s_t)$, where we have again abused notation as long as it is clear from the context. In this case, we can deduce that

$$d^\star(0,\theta)=(1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\big(s_t=0\mid s_0\sim\rho;\pi^\star_\theta\big)\pi^\star_\theta(\theta\,|\,0)\right]=(1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\big(s_t=0\mid s_0\sim\rho;\pi^\star_\theta\big)\right]$$
$$\overset{\text{(i)}}{\geq}(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\rho(0)\big[\mathbb{P}_\theta(0\,|\,0,\theta)\big]^t\overset{\text{(ii)}}{\geq}(1-\gamma)\sum_{t=0}^{\infty}\rho(0)\gamma^{2t}=\frac{1-\gamma}{1-\gamma^2}=\frac{1}{1+\gamma}\geq\frac{1}{2},$$

where in (i) we compute, for each t, the probability of a special trajectory with $s_1=\cdots=s_t=0$ and $a_0=\cdots=a_{t-1}=\theta$, and (ii) holds true since $P_\theta(0\,|\,0,\theta)\geq p\geq\gamma$. Taking this together with (220) yields

$$\frac{\min\left\{d^\star(0,\theta),\frac{1}{S}\right\}}{d^{\mathsf{b}}(0,\theta)}=\frac{2}{S\mu(0)}=2C,$$
$$\frac{\min\left\{d^\star(0,1-\theta),\frac{1}{S}\right\}}{d^{\mathsf{b}}(0,1-\theta)}=\frac{\min\left\{d^\star(0,1-\theta),\frac{1}{S}\right\}}{d^{\mathsf{b}}(0,\theta)}\leq\frac{\min\left\{d^\star(0,\theta),\frac{1}{S}\right\}}{d^{\mathsf{b}}(0,\theta)}=2C. \tag{221}$$

In addition, it is easily seen that $d^\star(s,a)=0$ for any $s>1$, and that

$$\frac{\min\left\{d^\star(1,a),\frac{1}{S}\right\}}{d^{\mathsf{b}}(1,a)}\leq\frac{1/S}{\mu(1)/2}=\frac{2}{S(1-1/CS)}\leq\frac{4}{S}\leq 2C,$$

where the first inequality comes from (220), the first identity uses the definition (192), and the last two inequalities result from an immediate consequence of (193) and $\gamma\geq 1/2$, i.e.,

$$\frac{1}{CS}\leq\frac{1}{4\gamma}\leq\frac{1}{2}. \tag{222}$$

As a result, putting the above relations together leads to

$$C^\star_{\mathsf{clipped}}=\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\min\left\{d^\star(s,a),\frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)}=\frac{\min\left\{d^\star(0,\theta),\frac{1}{S}\right\}}{d^{\mathsf{b}}(0,\theta)}=2C. \tag{223}$$

Proof of inequality (204)**.** Observing the basic identity (using $\mu(0)+\mu(1)=1$)

$$\frac{1+\gamma(1-x)\mu(1)V^\star_\theta(1)}{1-\gamma\big(\mu(1)x+\mu(0)\big)}=V^\star_\theta(1)+\frac{1-(1-\gamma)V^\star_\theta(1)}{1-\gamma\mu(0)-\gamma\mu(1)x},$$

we can obtain

$$\frac{1+\gamma(1-p)\mu(1)V^\star_\theta(1)}{1-\gamma\big(\mu(1)p+\mu(0)\big)}-\frac{1+\gamma(1-x_{\pi,\theta})\mu(1)V^\star_\theta(1)}{1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)}=\frac{1-(1-\gamma)V^\star_\theta(1)}{1-\gamma\mu(0)-\gamma\mu(1)p}-\frac{1-(1-\gamma)V^\star_\theta(1)}{1-\gamma\mu(0)-\gamma\mu(1)x_{\pi,\theta}}$$

$$= \big(1-(1-\gamma)V_\theta^\star(1)\big)\underbrace{\frac{\gamma\mu(1)(p-x_{\pi,\theta})}{\big[1-\gamma\big(\mu(1)p+\mu(0)\big)\big]\big[1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)\big]}}_{=:\,\alpha}$$
$$= \frac{\gamma\mu(1)(p-x_{\pi,\theta})}{2\alpha}, \tag{224}$$

where the last relation arises from the fact (200).

The remainder of the proof boils down to controlling α. Making use of the definition of p (cf. (195)), $\mu(s)$ (cf. (192)) and x_π (cf. (199)), we can demonstrate that

$$\begin{aligned}\alpha &= \left[1-\gamma\left(\left(1-\frac{1}{CS}\right)p+\frac{1}{CS}\right)\right]\left[1-\gamma\left(\left(1-\frac{1}{CS}\right)x_{\pi,\theta}+\frac{1}{CS}\right)\right] \le (1-\gamma p)(1-\gamma x_{\pi,\theta})\\ &\overset{\text{(i)}}{\le} (1-\gamma p)(1-\gamma q) \overset{\text{(ii)}}{\le} \left(1-\gamma\frac{p+q}{2}\right)^2\\ &= (1-\gamma^2)^2 = (1-\gamma)^2(1+\gamma)^2 \le 4(1-\gamma)^2,\end{aligned} \tag{225}$$

where (i) holds true owing to the trivial fact that $x_{\pi,\theta}\ge q$ for any policy π (as long as $p\ge q$), and (ii) is a consequence of the AM-GM inequality. Substituting it into (224) and using the definition (199) give

$$\begin{aligned}\frac{1+\gamma(1-p)\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)p+\mu(0)\big)} - \frac{1+\gamma(1-x_{\pi,\theta})\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)} &= \frac{\gamma\mu(1)(p-x_{\pi,\theta})}{2\alpha} \ge \frac{\gamma\mu(1)(p-x_{\pi,\theta})}{8(1-\gamma)^2}\\ &= \frac{\gamma\mu(1)}{8(1-\gamma)^2}(p-q)\pi(1-\theta\,|\,0)\\ &\ge \frac{3\gamma}{32(1-\gamma)^2}\frac{28(1-\gamma)^2\varepsilon}{\gamma}\pi(1-\theta\,|\,0) = \frac{21\varepsilon}{8}\big(1-\pi(\theta\,|\,0)\big).\end{aligned}$$

C.3 Proof of Theorem 4

To establish Theorem 4, we shall first generate a collection of hard problem instances (including MDPs and the associated batch datasets), and then conduct sample complexity analyses over these hard instances.

C.3.1 Construction of hard problem instances

Construction of the hard MDPs. To begin with, for any integer $H\ge 32$, let us consider a set $\Theta\subseteq\{0,1\}^H$ of H-dimensional vectors, which we shall construct shortly. We then generate a collection of MDPs

$$\mathsf{MDP}(\Theta) = \Big\{\mathcal{M}_\theta = \big(\mathcal{S},\mathcal{A},P^\theta=\{P_h^{\theta_h}\}_{h=1}^H,\{r_h\}_{h=1}^H,H\big)\mid \theta=[\theta_h]_{1\le h\le H}\in\Theta\Big\}, \tag{226}$$

where

$$\mathcal{S}=\{0,1,\ldots,S-1\},\qquad\text{and}\qquad \mathcal{A}=\{0,1\}.$$

To define the transition kernel of these MDPs, we find it convenient to introduce the following state distribution supported on the state subset $\{0,1\}$:

$$\mu(s) = \frac{1}{CS}\mathbb{1}\{s=0\}+\Big(1-\frac{1}{CS}\Big)\mathbb{1}\{s=1\}, \tag{227}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and $C>0$ is some constant that will determine the concentrability coefficient $C^\star_{\mathsf{clipped}}$ (as we shall detail momentarily). It is assumed that

$$\frac{1}{CS}\le\frac{1}{4}. \tag{228}$$

With this distribution in mind, we can specify the transition kernel $P^\theta = \{P_h^{\theta_h}\}_{h=1}^H$ of the MDP $\mathcal{M}_\theta$ as follows:

$$P_h^{\theta_h}(s' \,|\, s,a) = \begin{cases} p\mathbb{1}\{s'=0\} + (1-p)\mu(s') & \text{if} \quad (s,a) = (0,\theta_h) \\ q\mathbb{1}\{s'=0\} + (1-q)\mu(s') & \text{if} \quad (s,a) = (0,1-\theta_h) \\ \mathbb{1}\{s'=1\} & \text{if} \quad (s,a) = (1,0) \\ \big(1-\frac{2c_1}{H}\big)\mathbb{1}\{s'=1\} + \frac{2c_1}{H}\mu(s') & \text{if} \quad (s,a) = (1,1) \\ \big(1-\frac{1}{H}\big)\mathbb{1}\{s'=s\} + \frac{1}{H}\mu(s') & \text{if} \quad s>1 \end{cases} \tag{229}$$

for any $(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$, where p and q are set to be

$$p = 1 - \frac{c_1}{H} + \frac{c_2\varepsilon}{H^2} \qquad \text{and} \qquad q = 1 - \frac{c_1}{H} - \frac{c_2\varepsilon}{H^2} \tag{230}$$

for $c_1 = 1/4$ and $c_2 = 4096$ such that

$$\frac{c_2\varepsilon}{H^2} \le \frac{c_1}{2H} \le \frac{1}{8}. \tag{231}$$

It is readily seen from the above assumption that

$$p > q \ge \frac{1}{2}. \tag{232}$$

In view of the transition kernel (229), the MDP will never leave the state subset $\{0,1\}$ if its initial state belongs to $\{0,1\}$. The reward function of all these MDPs is chosen to be

$$r_h(s,a) = \begin{cases} 1 & \text{if} \quad s=0 \\ \frac{1}{2} & \text{if} \quad (s,a) = (1,0) \\ 0 & \text{if} \quad (s,a) = (1,1) \\ 0 & \text{if} \quad s>1 \end{cases} \tag{233}$$

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

Finally, let us choose the set $\Theta \subseteq \{0,1\}^H$. By virtue of the Gilbert-Varshamov lemma (Gilbert, 1952), one can construct $\Theta \subseteq \{0,1\}^H$ in a way that

$$|\Theta| \ge e^{H/8} \qquad \text{and} \qquad \|\theta - \widetilde{\theta}\|_1 \ge \frac{H}{8} \quad \text{for any } \theta, \widetilde{\theta} \in \Theta \text{ obeying } \theta \ne \widetilde{\theta}. \tag{234}$$

In other words, the set Θ we construct contains an exponentially large number of vectors that are sufficiently separated. This property plays an important role in the ensuing analysis.

Value functions and optimal policies. Next, we look at the value functions of the constructed MDPs and identify the optimal policies. For the sake of notational clarity, for the MDP $\mathcal{M}_\theta$, we denote by $\pi^{\star,\theta} = \{\pi_h^{\star,\theta}\}_{h=1}^H$ the optimal policy, and let $V_h^{\pi,\theta}$ (resp. $V_h^{\star,\theta}$) indicate the value function of policy π (resp. $\pi^{\star,\theta}$) at time step h. The following lemma collects a couple of useful properties concerning the value functions and optimal policies; the proof can be found in Appendix C.3.3.

Lemma 12. *Consider any $\theta \in \Theta$ and any policy π. Then it holds that*

$$V_h^{\pi,\theta}(0) = 1 + \big(\mu(1)x_h^{\pi,\theta} + \mu(0)\big)V_{h+1}^{\pi,\theta}(0) + (1-x_h^{\pi,\theta})\mu(1)V_{h+1}^{\pi,\theta}(1) \tag{235}$$

for any $h \in [H]$, where

$$x_h^{\pi,\theta} = p\pi_h(\theta_h \,|\, 0) + q\pi_h(1-\theta_h \,|\, 0). \tag{236}$$

In addition, for any $h \in [H]$, the optimal policies and the optimal value functions obey

$$\pi_h^{\star,\theta}(\theta_h \,|\, 0) = 1, \qquad V_h^{\star,\theta}(0) \ge \tfrac{2}{3}(H+1-h), \tag{237a}$$

$$\pi_h^{\star,\theta}(0 \,|\, 1) = 1, \qquad V_h^{\star,\theta}(1) = \tfrac{1}{2}(H+1-h), \tag{237b}$$

provided that $0 < c_1 \le 1/2$.

Construction of the batch dataset. A batch dataset is then generated, which consists of K *independent* sample trajectories each of length H. The initial state distribution ρ^{b} and the behavior policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{h=1}^H$ (according to (45)) are chosen as follows:

$$\rho^{\mathsf{b}}(s) = \mu(s) \quad \text{and} \quad \pi_h^{\mathsf{b}}(a \,|\, s) = \frac{1}{2}, \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H],$$

where μ has been defined in (227). As it turns out, for any MDP $\mathcal{M}_\theta$, the occupancy distributions of the above batch dataset admit the following simple characterization:

$$d_h^{\mathsf{b}}(s) = \mu(s), \qquad d_h^{\mathsf{b}}(s,a) = \frac{1}{2}\mu(s), \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{238}$$

Additionally, we shall choose the initial state distribution ρ as follows

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0, \\ 0, & \text{if } s > 0. \end{cases} \tag{239}$$

With this choice of ρ, the single-policy clipped concentrability coefficient $C^\star_{\mathsf{clipped}}$ and the quantity C are intimately connected as follows:

$$C^\star_{\mathsf{clipped}} = 2C. \tag{240}$$

The proof of the claims (238) and (240) can be found in Appendix C.3.4.

C.3.2 Establishing the minimax lower bound

We are now positioned to establish our sample complexity lower bounds. Recalling our choice of ρ in (239), our proof seeks to control the quantity

$$\big\langle \rho, V_1^{\star,\theta} - V_1^{\widehat{\pi},\theta} \big\rangle = V_1^{\star,\theta}(0) - V_1^{\widehat{\pi},\theta}(0),$$

where $\widehat{\pi}$ is any policy estimator computed based on the batch dataset.

Step 1: converting $\widehat{\pi}$ into an estimate $\widehat{\theta}$ of θ. Towards this, we first make the following claim: for an arbitrary policy π obeying

$$\sum_{h=1}^H \big\| \pi_h(\cdot \,|\, 0) - \pi_h^{\star,\theta}(\cdot \,|\, 0) \big\|_1 \geq \frac{H}{8}, \tag{241}$$

one has

$$\big\langle \rho, V_1^{\star,\theta} - V_1^{\pi,\theta} \big\rangle > \varepsilon. \tag{242}$$

We shall postpone the proof of this claim to Appendix C.3.4. Suppose for the moment that there exists a policy estimate $\widehat{\pi}$ that achieves

$$\mathbb{P}\Big\{ \big\langle \rho, V_1^{\star,\theta} - V_1^{\widehat{\pi},\theta} \big\rangle \leq \varepsilon \Big\} \geq \frac{3}{4}, \tag{243}$$

then in view of (242), we necessarily have

$$\mathbb{P}\left\{ \sum_{h=1}^H \big\| \widehat{\pi}_h(\cdot \,|\, 0) - \pi_h^{\star,\theta}(\cdot \,|\, 0) \big\|_1 < H/8 \right\} \geq \frac{3}{4}. \tag{244}$$

With the above observation in mind, we are motivated to construct the following estimate $\widehat{\theta}$ for $\theta \in \Theta$:

$$\widehat{\theta} = \arg\min_{\overline{\theta} \in \Theta} \sum_{h=1}^H \big\| \widehat{\pi}_h(\cdot \,|\, 0) - \pi_h^{\star,\overline{\theta}}(\cdot \,|\, 0) \big\|_1. \tag{245}$$

If $\sum_h \big\|\widehat{\pi}(\cdot\,|\,0) - \pi^{\star,\theta}(\cdot\,|\,0)\big\|_1 < H/8$ holds for some $\theta \in \Theta$, then for any $\widetilde{\theta} \in \Theta$ with $\widetilde{\theta} \neq \theta$ one has

$$\begin{aligned}\sum_{h=1}^{H} \big\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\widetilde{\theta}}(\cdot\,|\,0)\big\|_1 &\geq \sum_{h=1}^{H} \big\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h^{\star,\widetilde{\theta}}(\cdot\,|\,0)\big\|_1 - \sum_{h=1}^{H} \big\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\big\|_1 \\ &= 2\big\|\theta - \widetilde{\theta}\big\|_1 - \sum_{h=1}^{H} \big\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\big\|_1 \\ &> \frac{H}{4} - \frac{H}{8} = \frac{H}{8}, \end{aligned} \tag{246}$$

where the first inequality holds by the triangle inequality, the second line arises from the fact $\pi_h^{\star,\theta}(\theta_h\,|\,0) = 1$ for all $1 \leq h \leq H$ (see (237)), and the last line comes from the properties (234) about Θ. Putting (245) and (246) together implies that $\widehat{\theta} = \theta$ if

$$\sum_{h=1}^{H} \big\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\big\|_1 < \frac{H}{8} < \sum_{h=1}^{H} \big\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\widetilde{\theta}}(\cdot\,|\,0)\big\|_1$$

is valid for all $\widetilde{\theta} \in \Theta$ with $\widetilde{\theta} \neq \theta$. As a consequence,

$$\mathbb{P}\big(\widehat{\theta} = \theta\big) \geq \mathbb{P}\left(\sum_{h=1}^{H} \big\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\big\|_1 < \frac{H}{8}\right) \geq \frac{3}{4}. \tag{247}$$

In the sequel, we aim to demonstrate that (247) cannot possibly happen without enough samples, which would in turn contradict (243).

Step 2: probability of error in testing multiple hypotheses. Next, we turn attention to a $|\Theta|$-ary hypothesis testing problem. For any $\theta \in \Theta$, denote by $\mathbb{P}_\theta$ the probability distribution when the MDP is $\mathcal{M}_\theta$. We will then study the minimax probability of error defined as follows:

$$p_{\mathrm{e}} \coloneqq \inf_{\psi} \max_{\theta \in \Theta} \mathbb{P}_\theta(\psi \neq \theta), \tag{248}$$

where the infimum is taken over all possible tests ψ (constructed based on the batch dataset available).

Let $\mu^{\mathsf{b},\theta}$ (resp. $\mu_h^{\mathsf{b},\theta_h}(s_h)$) represent the distribution of a sample trajectory $\{s_1, a_1, s_2, a_2, \cdots, s_H, a_H\}$ (resp. a sample (a_h, s_{h+1}) conditional on s_h) for the MDP $\mathcal{M}_\theta$. Recalling that the K trajectories in the batch dataset are independently generated, one obtains

$$\begin{aligned} p_{\mathrm{e}} &\overset{\text{(i)}}{\geq} 1 - \frac{K \max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}} \mathsf{KL}\big(\mu^{\mathsf{b},\theta} \,\|\, \mu^{\mathsf{b},\widetilde{\theta}}\big) + \log 2}{\log|\Theta|} \\ &\overset{\text{(ii)}}{\geq} 1 - \frac{8K}{H} \max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}} \mathsf{KL}\big(\mu^{\mathsf{b},\theta} \,\|\, \mu^{\mathsf{b},\widetilde{\theta}}\big) - \frac{8\log 2}{H} \\ &\overset{\text{(iii)}}{\geq} \frac{1}{2} - \frac{8K}{H} \max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}} \mathsf{KL}\big(\mu^{\mathsf{b},\theta} \,\|\, \mu^{\mathsf{b},\widetilde{\theta}}\big), \end{aligned} \tag{249}$$

where (i) arises from Fano's inequality (cf. (Tsybakov, 2009, Corollary 2.6)) and the additivity property of the KL divergence (cf. Tsybakov (2009, Page 85)), (ii) holds since $|\Theta| \geq e^{H/8}$ (according to our construction (234)), and (iii) is valid when $H \geq 16\log 2$. Recalling that the occupancy state distribution d_h^{b} is the same for any MDP $\mathcal{M}_\theta$ with $\theta \in \Theta$ (see (238)), one can invoke the chain rule of the KL divergence (Duchi, 2018, Lemma 5.2.8) and the Markovian nature of the sample trajectories to obtain

$$\mathsf{KL}\big(\mu^{\mathsf{b},\theta} \,\|\, \mu^{\mathsf{b},\widetilde{\theta}}\big) = \sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h \sim d_h^{\mathsf{b}}} \Big[\mathsf{KL}\big(\mu_h^{\mathsf{b},\theta_h}(s_h) \,\|\, \mu_h^{\mathsf{b},\widetilde{\theta}_h}(s_h)\big)\Big] = \frac{1}{2}\mu(0) \sum_{h=1}^{H} \sum_{a\in\{0,1\}} \mathsf{KL}\left(P_h^{\theta_h}(\cdot\,|\,0,a) \,\|\, P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,a)\right),$$

where the last identity holds true since (by construction and (238))

$$\begin{aligned}\mathop{\mathbb{E}}_{s_h\sim d_h^{\mathsf{b}}}\Big[\mathsf{KL}\big(\mu_h^{\mathsf{b},\theta_h}(s_h)\,\|\,\mu_h^{\mathsf{b},\widetilde{\theta}_h}(s_h)\big)\Big] &= \sum_s d_h^{\mathsf{b}}(s)\left\{\sum_{a,s'}\pi_h^{\mathsf{b}}(a\,|\,s)P_h^{\theta_h}(s'\,|\,s,a)\log\frac{\pi_h^{\mathsf{b}}(a\,|\,s)P_h^{\theta_h}(s'\,|\,s,a)}{\pi_h^{\mathsf{b}}(a\,|\,s)P_h^{\widetilde{\theta}_h}(s'\,|\,s,a)}\right\}\\ &= \frac{1}{2}\mu(0)\sum_a\sum_{s'}P_h^{\theta_h}(s'\,|\,0,a)\log\frac{P_h^{\theta_h}(s'\,|\,0,a)}{P_h^{\widetilde{\theta}_h}(s'\,|\,0,a)}\\ &= \frac{1}{2}\mu(0)\sum_a\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,a)\,\|\,P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,a)\big).\end{aligned}$$

Substitution into (249) yields

$$p_{\mathrm{e}}\geq\frac{1}{2}-\frac{4K\mu(0)}{H}\max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}}\sum_{h=1}^{H}\Big[\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,0)\,\|\,P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,0)\big)+\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,1)\,\|\,P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,1)\big)\Big]. \tag{250}$$

It then boils down to bounding the KL divergence terms in (250). If $\theta_h=\widetilde{\theta}_h$, then it is self-evident that

$$\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,0)\,\|\,P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,0)\big)+\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,1)\,\|\,P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,1)\big)=0. \tag{251}$$

Consider now the case that $\theta_h\neq\widetilde{\theta}_h$, and suppose without loss of generality that $\theta_h=0$ and $\widetilde{\theta}_h=1$. It is seen that

$$\begin{aligned}P_h^{\theta_h}(0\,|\,0,0)&=P_h^{\theta_h}(\theta_h\,|\,0,0)=\left(1-\frac{1}{CS}\right)p+\frac{1}{CS},\\ P_h^{\widetilde{\theta}_h}(0\,|\,0,0)&=P_h^{\widetilde{\theta}_h}\big(1-\widetilde{\theta}_h\,|\,0,0\big)=\left(1-\frac{1}{CS}\right)q+\frac{1}{CS}.\end{aligned}$$

Given that $p\geq q\geq 1/2$ (see (232)), we can apply Lemma 10 to arrive at

$$\begin{aligned}\mathsf{KL}\left(P_h^{\theta_h}(0\,|\,0,0)\,\|\,P_h^{\widetilde{\theta}_h}(0\,|\,0,0)\right)&=\mathsf{KL}\left(\left(1-\frac{1}{CS}\right)p+\frac{1}{CS}\,\|\,\left(1-\frac{1}{CS}\right)q+\frac{1}{CS}\right)\\ &\leq\frac{\left(1-\frac{1}{CS}\right)^2(p-q)^2}{\left(\left(1-\frac{1}{CS}\right)p+\frac{1}{CS}\right)\left(1-p-(1-p)\frac{1}{CS}\right)}\\ &\overset{\text{(i)}}{\leq}\frac{\left(1-\frac{1}{CS}\right)^2(p-q)^2}{\left(\left(1-\frac{1}{CS}\right)p\right)\left((1-p)(1-\frac{1}{CS})\right)}=\frac{4(c_2)^2\varepsilon^2}{H^4p(1-p)}\\ &\overset{\text{(ii)}}{=}\frac{4(c_2)^2\varepsilon^2}{H^4\left(1-\frac{c_1}{H}+\frac{c_2\varepsilon}{H^2}\right)\left(\frac{c_1}{H}-\frac{c_2\varepsilon}{H^2}\right)}\\ &\leq\frac{4(c_2)^2\varepsilon^2}{H^4\frac{1}{2}\frac{c_1}{2H}}=\frac{16(c_2)^2\varepsilon^2}{c_1H^3},\end{aligned} \tag{252}$$

where (i) and (ii) make use of the definition (230) of (p,q), and the last line follows as long as $\frac{c_2\varepsilon}{H^2}\leq\frac{c_1}{2H}\leq\frac{1}{4}$. Similarly, it can be easily verified that $\mathsf{KL}\big(P_h^{\theta_h}(0\,|\,0,1)\,\|\,P_h^{\widetilde{\theta}_h}(0\,|\,0,1)\big)$ can be upper bounded in the same way. Substituting (252) and (251) back into (250) indicates that: if the sample size obeys

$$N=KH\leq\frac{c_1CSH^4}{512(c_2)^2\varepsilon^2}=\frac{c_1C^\star_{\mathsf{clipped}}SH^4}{1024(c_2)^2\varepsilon^2}, \tag{253}$$

then one necessarily has

$$\begin{aligned}p_{\mathrm{e}}&\geq\frac{1}{2}-\frac{4K\mu(0)}{H}\max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}}\sum_{h=1}^{H}\Big[\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,0)\,\|\,P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,0)\big)+\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,1)\,\|\,P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,1)\big)\Big]\\ &\geq\frac{1}{2}-\frac{4K\mu(0)}{H}\sum_{h=1}^{H}\frac{32(c_2)^2\varepsilon^2}{c_1H^3}\geq\frac{1}{4}.\end{aligned} \tag{254}$$

Step 3: combining the above results. Suppose that there exists an estimator $\widehat{\pi}$ satisfying

$$\max_{\theta\in\Theta}\mathbb{P}_\theta\left\{\left\langle\rho, V_1^{\star,\theta}-V_1^{\widehat{\pi},\theta}\right\rangle\geq\varepsilon\right\}<\frac{1}{4}, \tag{255}$$

where $\mathbb{P}_\theta$ denotes the probability when the MDP is $\mathcal{M}_\theta$. Then in view of the analysis in Step 1, we must have

$$\mathbb{P}_\theta\left(\sum_{h=1}^{H}\left\|\widehat{\pi}(\cdot\,|\,0)-\pi^{\star,\theta}(\cdot\,|\,0)\right\|_1<\frac{H}{8}\right)\geq\frac{3}{4},\qquad\text{for all }\theta\in\Theta,$$

and as a consequence of (247), the estimator $\widehat{\theta}$ defined in (245) must satisfy

$$\mathbb{P}_\theta\big(\widehat{\theta}\neq\theta\big)<\frac{1}{4},\qquad\text{for all }\theta\in\Theta. \tag{256}$$

Nevertheless, this cannot possibly happen under the sample size condition (253); otherwise it is contradictory to the result in (254). This concludes the proof by inserting $c_1=1/4$ and $c_2=4096$.

C.3.3 Proof of Lemma 12

To start with, for any policy π, it is observed that the value function of state $s=0$ at step h is

$$\begin{aligned}
V_h^{\pi,\theta}(0)&=\mathop{\mathbb{E}}_{a\sim\pi_h(\cdot\,|\,0)}\left[1+\sum_{s'}P_h^{\theta_h}(s'\,|\,0,a)V_{h+1}^{\pi,\theta}(s')\right]\\
&=1+\pi_h(\theta_h\,|\,0)\Big[\big(p+(1-p)\mu(0)\big)V_{h+1}^{\pi,\theta}(0)+(1-p)\mu(1)V_{h+1}^{\pi,\theta}(1)\Big]\\
&\quad+\pi(1-\theta_h\,|\,0)\Big[\big(q+(1-q)\mu(0)\big)V_{h+1}^{\pi,\theta}(0)+(1-q)\mu(1)V_{h+1}^{\pi,\theta}(1)\Big]\\
&=1+\Big[p\pi_h(\theta_h\,|\,0)+q\pi(1-\theta_h\,|\,0)+\mu(0)-p\pi_h(\theta_h\,|\,0)\mu(0)-q\pi(1-\theta_h\,|\,0)\mu(0)\Big]V_{h+1}^{\pi,\theta}(0)\\
&\quad+\mu(1)\Big[1-p\pi_h(\theta_h\,|\,0)-q\pi(1-\theta_h\,|\,0)\Big]V_{h+1}^{\pi,\theta}(1)\\
&\overset{\text{(i)}}{=}1+\Big[x_h^{\pi,\theta}+(1-x_h^{\pi,\theta})\mu(0)V_{h+1}^{\pi,\theta}(0)+(1-x_h^{\pi,\theta})\mu(1)V_{h+1}^{\pi,\theta}(1)\Big]\\
&\overset{\text{(ii)}}{=}1+\big(\mu(1)x_h^{\pi}+\mu(0)\big)V_{h+1}^{\pi,\theta}(0)+(1-x_h^{\pi})\mu(1)V_{h+1}^{\pi,\theta}(1),
\end{aligned}\tag{257}$$

where (i) is valid due to the choice

$$x_h^{\pi,\theta}=p\pi_h(\theta_h\,|\,0)+q\pi_h(1-\theta_h\,|\,0), \tag{258}$$

and (ii) holds since $\mu(0)+\mu(1)=1$.

Additionally, the value function of state 1 at any step h obeys

$$\begin{aligned}
V_h^{\pi,\theta}(1)&=\pi_h(0\,|\,1)\left(\frac{1}{2}+V_{h+1}^{\pi,\theta}(1)\right)+\pi_h(1\,|\,1)\left[\left(1-\frac{2c_1}{HCS}\right)V_{h+1}^{\pi,\theta}(1)+\frac{2c_1}{HCS}V_{h+1}^{\pi,\theta}(0)\right] &(259)\\
&\overset{\text{(i)}}{\leq}\pi_h(0\,|\,1)\left(\frac{1}{2}+V_{h+1}^{\pi,\theta}(1)\right)+\pi_h(1\,|\,1)\left[\left(1-\frac{2c_1}{HCS}\right)V_{h+1}^{\pi,\theta}(1)+\frac{2c_1}{HCS}(H-h)\right]\\
&\overset{\text{(ii)}}{\leq}\pi_h(0\,|\,1)\left(\frac{1}{2}+V_{h+1}^{\pi,\theta}(1)\right)+\pi_h(1\,|\,1)\left[\frac{1}{2}+\left(1-\frac{2c_1}{HCS}\right)V_{h+1}^{\pi,\theta}(1)\right]\\
&=\frac{1}{2}+V_{h+1}^{\pi,\theta}(1)-\frac{2c_1}{HCS}\pi_h(1\,|\,1)V_{h+1}^{\pi,\theta}(1), &(260)
\end{aligned}$$

where (i) arises from the basic fact $0\leq V_h^{\pi,\theta}(s)\leq H-h+1$ for any policy π and all $(s,h)\in\mathcal{S}\times[H]$, and (ii) holds since $\frac{2c_1}{HCS}(H-h)\leq\frac{1}{2}$ for c_1 small enough. The above results lead to several immediate facts.

- If we choose π such that $\pi_h(0\,|\,1) = 1$ for all $h \in [H]$, then (259) tells us that

$$V_h^{\pi,\theta}(1) = \frac{1}{2} + V_{h+1}^{\pi,\theta}(1). \tag{261}$$

  A recursive application of this relation reveals that

$$V_h^{\pi,\theta}(1) = \frac{1}{2} + V_{h+1}^{\pi,\theta}(1) = \cdots = \sum_{j=h}^{H} \frac{1}{2} = \frac{1}{2}(H+1-h). \tag{262}$$

- For any policy π, applying (260) recursively tells us that

$$V_h^{\pi,\theta}(1) \le \frac{1}{2} + V_{h+1}^{\pi,\theta}(1) \le \cdots \le \sum_{j=h}^{H} \frac{1}{2} = \frac{1}{2}(H+1-h). \tag{263}$$

The above two facts taken collectively imply that the optimal policy and optimal value function obey

$$\pi_h^{\star,\theta}(0\,|\,1) = 1, \qquad V_h^{\star,\theta}(1) = \frac{1}{2}(H+1-h), \qquad \forall h \in [H]. \tag{264}$$

We then return to state 0. By taking π such that $\pi_h(\theta_h\,|\,0) = 1$ (and hence $x_h^{\pi,\theta} = p$) for all $h \in [H]$, one can invoke (257) to derive

$$\begin{aligned} V_h^{\pi,\theta}(0) &= 1 + \big(\mu(1)p + \mu(0)\big)V_{h+1}^{\pi,\theta}(0) + (1-p)\mu(1)V_{h+1}^{\pi,\theta}(1) \\ &\ge 1 + pV_{h+1}^{\pi,\theta}(0) \ge \sum_{j=0}^{H-h} p^j \ge \sum_{j=0}^{H-h} \Big(1 - \frac{c_1}{H}\Big)^j = \frac{1 - \big(1-\frac{c_1}{H}\big)^{H-h+1}}{c_1/H} \\ &\ge \frac{2}{3}(H+1-h). \end{aligned} \tag{265}$$

To see that why the last inequality holds, it suffices to observe that

$$\Big(1 - \frac{c_1}{H}\Big)^{H-h+1} \le \exp\Big(-\frac{c_1}{H}(H-h+1)\Big) \le 1 - \frac{2c_1(H-h+1)}{3H},$$

as long as $c_1 \le 0.5$, which follows due to the elementary inequalities $1 - x \le \exp(-x)$ for any $x \ge 0$ and $\exp(-x) \le 1 - 2x/3$ for any $0 \le x \le 1/2$. Combine (265) with (264) to reach

$$V_h^{\star,\theta}(0) \ge V_h^{\pi,\theta}(0) \ge \frac{2}{3}(H+1-h) > V_h^{\star,\theta}(1). \tag{266}$$

Moreover, it follows from (257) that

$$\begin{aligned} V_h^{\star,\theta}(0) &= 1 + \big(\mu(1)x_h^{\pi^\star,\theta,\theta} + \mu(0)\big)V_{h+1}^{\star,\theta}(0) + (1 - x_h^{\pi^\star,\theta,\theta})\mu(1)V_{h+1}^{\star,\theta}(1) \\ &= 1 + \mu(0)V_{h+1}^{\star,\theta}(0) + \mu(1)V_{h+1}^{\star,\theta}(1) + \mu(1)\big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\big)x_h^{\pi^\star,\theta,\theta}. \end{aligned} \tag{267}$$

Observing that the function

$$\mu(1)\big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\big)x \tag{268}$$

is increasing in x (as a result of (266)) and that $x_h^{\pi,\theta}$ is increasing in $\pi_h(\theta_h\,|\,0)$ (since $p \ge q$), we can readily conclude that the optimal policy in state 0 obeys

$$\pi_h^{\star,\theta}(\theta_h\,|\,0) = 1, \qquad \text{for all } h \in [H]. \tag{269}$$

C.3.4 Proof of auxiliary properties

Throughout this section, we shall suppress the dependency on θ in the notation $d_h^\star$ whenever it is clear from the context.

Proof of claim (238). For any MDP $\mathcal{M}_\theta$, from the definition of $d_h^{\mathsf{b}}(s,a)$ in (46) and the Markov property, it is clearly seen that

$$d_{h+1}^{\mathsf{b}}(s) = d_{h+1}^{\pi^{\mathsf{b}}}(s;\rho^{\mathsf{b}}) = \mathbb{P}(s_{h+1}=s \,|\, s_h \sim d_h^{\mathsf{b}};\pi^{\mathsf{b}}), \qquad \forall (s,h)\in\mathcal{S}\times[H]. \tag{270}$$

Recalling that $d_1^{\mathsf{b}}(s)=\rho^{\mathsf{b}}(s)=\mu(s)$ for all $s\in\mathcal{S}$, one can then show that

$$\begin{aligned}
d_2^{\mathsf{b}}(0) &= \mathbb{P}\{s_2=0 \,|\, s_1\sim d_1^{\mathsf{b}};\pi^{\mathsf{b}}\} \\
&= \mu(0)\left[\pi_1^{\mathsf{b}}(\theta_1\,|\,0)P_1^{\theta_1}(0\,|\,0,\theta_1)+\pi_1^{\mathsf{b}}(1-\theta_1\,|\,0)P_1^{\theta_1}(0\,|\,0,1-\theta_1)\right] \\
&\quad + \mu(1)\left[\pi_1^{\mathsf{b}}(0\,|\,1)P_1^{\theta_1}(0\,|\,1,0)+\pi_1^{\mathsf{b}}(1\,|\,1)P_1^{\theta_1}(0\,|\,1,1)\right] \\
&= \frac{\mu(0)}{2}\left[P_1^{\theta_1}(0\,|\,0,\theta_1)+P_1^{\theta_1}(0\,|\,0,1-\theta_1)\right]+\frac{\mu(1)}{2}\left[P_1^{\theta_1}(0\,|\,1,0)+P_1^{\theta_1}(0\,|\,1,1)\right] \\
&= \frac{\mu(0)}{2}\left[(p+q)+(2-p-q)\mu(0)\right]+\frac{\mu(1)}{2}\mu(0)\frac{2c_1}{H} \\
&= \frac{\mu(0)}{2}\left[2-\frac{2c_1}{H}+\frac{2c_1}{H}\mu(0)\right]+\frac{\mu(1)}{2}\mu(0)\frac{2c_1}{H}=\mu(0),
\end{aligned}$$

where the last inequality holds since $\mu(1)+\mu(0)=1$. Similarly, it can be verified that $d_1^{\mathsf{b}}(1)=\mu(1)$, thereby implying that $d_2^{\mathsf{b}}=\mu$. Repeating this argument recursively for steps $h=2,\ldots,H$ confirms that

$$d_h^{\mathsf{b}}(s)=\mu(s), \qquad \forall (s,h)\in\mathcal{S}\times[H]. \tag{271}$$

This further allows one to demonstrate that

$$d_h^{\mathsf{b}}(s,a)=d_h^{\mathsf{b}}(s)\pi_h^{\mathsf{b}}(a\,|\,s)=\mu(s)\pi_h^{\mathsf{b}}(a\,|\,s)=\mu(s)/2, \qquad \forall (s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]. \tag{272}$$

Proof of claim (240). Consider any MDP $\mathcal{M}_\theta$, for which we have shown in Lemma 12 that $\pi_h^{\star,\theta}(\theta_h\,|\,0)=1$ for all $h\in[H]$. It is observed that

$$\begin{aligned}
d_h^\star(0,\theta_h) &= d_h^\star(0)\pi_h^{\star,\theta}(\theta_h\,|\,0)=d_h^\star(0)=\mathbb{P}\big\{s_h=0\,|\,s_{h-1}\sim d_{h-1}^\star;\pi^{\star,\theta}\big\} \\
&\geq d_{h-1}^\star(0)\pi_{h-1}^{\star,\theta}(\theta_{h-1}\,|\,0)P_{h-1}^{\theta_{h-1}}(0\,|\,0,\theta_{h-1})=d_{h-1}^\star(0)P_{h-1}^{\theta_{h-1}}(0\,|\,0,\theta_{h-1}) \\
&\geq \cdots \geq d_1^\star(0)\prod_{j=0}^{h-1}P_j^{\theta_j}(0\,|\,0,\theta_j)=\rho(0)\prod_{j=0}^{h-1}P_j^{\theta_j}(0\,|\,0,\theta_j) \\
&\geq \rho(0)\prod_{j=0}^{h-1}p\geq\left(1-\frac{c_1}{H}\right)^H>\frac{1}{2},
\end{aligned} \tag{273}$$

where the last line makes use of the properties $p\geq 1-c_1/H$, $\rho(0)=1$, and

$$\left(1-\frac{c_1}{H}\right)^H\geq\left(1-\frac{1}{2H}\right)^H>\frac{1}{2}$$

provided that $0<c_1<1/2$. Combining this with (238), we arrive at

$$\begin{aligned}
\max_{h\in[H]}\frac{\min\left\{d_h^\star(0,\theta_h),\frac{1}{S}\right\}}{d_h^{\mathsf{b}}(0,\theta_h)} &= \frac{2}{S\mu(0)}=2C, \\
\max_{h\in[H]}\frac{\min\left\{d_h^\star(0,1-\theta_h),\frac{1}{S}\right\}}{d_h^{\mathsf{b}}(0,1-\theta_h)} &= \max_{h\in[H]}\frac{\min\left\{d_h^\star(0)\pi_h^\star(1-\theta_h\,|\,0),\frac{1}{S}\right\}}{d_h^{\mathsf{b}}(0,1-\theta_h)}=0, \\
\max_{a\in\{0,1\},h\in[H]}\frac{\min\left\{d_h^\star(1,a),\frac{1}{S}\right\}}{d_h^{\mathsf{b}}(1,a)} &\overset{\text{(i)}}{\leq}\frac{1/S}{\mu(1)/2}\overset{\text{(ii)}}{=}\frac{2}{S\left(1-\frac{1}{SC}\right)}\leq\frac{4}{S}\leq 2C,
\end{aligned}$$

where (i) arises from (238), (ii) relies on the definition in (227), and the final two inequalities come from the assumption in (228). Taking this together with the straightforward condition $d_h^\star(s)=0$ $(s>1)$ yields

$$C^\star_{\mathsf{clipped}}=\max_{h\in[H]}\frac{\min\left\{d_h^\star(0,\theta_h),\frac{1}{S}\right\}}{d_h^{\mathsf{b}}(0,\theta_h)}=2C. \tag{274}$$

Proof of inequality (242). By virtue of (236) and (237), we see that $x_h^{\pi^{\star,\theta},\theta} = p$ for all $h \in [H]$, which combined with (235) gives

$$\begin{aligned}
\big\langle \rho, V_h^{\star,\theta} - V_h^{\pi,\theta} \big\rangle &= V_h^{\star,\theta}(0) - V_h^{\pi,\theta}(0) \\
&= \big(\mu(1)p + \mu(0)\big) V_{h+1}^{\star,\theta}(0) + (1-p)\mu(1) V_{h+1}^{\star,\theta}(1) \\
&\quad - \big(\mu(1)x_h^{\pi,\theta} + \mu(0)\big) V_{h+1}^{\pi,\theta}(0) - (1 - x_h^{\pi,\theta})\mu(1) V_{h+1}^{\pi,\theta}(1) \\
&\overset{\text{(i)}}{\geq} \big(\mu(1)x_h^{\pi,\theta} + \mu(0)\big)\Big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\Big) + \mu(1)\big(p - x_h^{\pi,\theta}\big) V_{h+1}^{\star,\theta}(0) \\
&\quad + (1-p)\mu(1) V_{h+1}^{\star,\theta}(1) - \big(1 - x_h^{\pi,\theta}\big)\mu(1) V_{h+1}^{\star,\theta}(1) \\
&= \big(\mu(1)x_h^{\pi,\theta} + \mu(0)\big)\Big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\Big) + \big(p - x_h^{\pi,\theta}\big)\mu(1)\Big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\Big) \\
&\overset{\text{(ii)}}{\geq} q\Big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\Big) + \big(p - x_h^{\pi,\theta}\big)\mu(1)\Big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\Big) \\
&\overset{\text{(iii)}}{\geq} q\Big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\Big) + \frac{3}{8}(p-q)\big\|\pi_h^{\star,\theta}(0) - \pi_h(0)\big\|_1 \Big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\Big) \\
&\overset{\text{(iv)}}{\geq} q\Big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\Big) + \frac{c_2\varepsilon}{8H^2}(H+1-h)\big\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\big\|_1,
\end{aligned} \tag{275}$$

where (i) holds since $V_{h+1}^{\pi,\theta}(1) \leq V_{h+1}^{\star,\theta}(1)$, (ii) follows from the fact that $x_h^{\pi} \geq q$ for any π and $h \in [H]$, and (iv) arises from the facts (237) and the choice (230) of (p,q). To see why (iii) is valid, it suffices to note that $\mu(1) = 1 - \frac{1}{CS} \geq \frac{3}{4}$ (as a consequence of (227) and (228)) and

$$p - x_h^{\pi,\theta} = (p-q)\big(1 - \pi_h(\theta_h\,|\,0)\big) = \frac{1}{2}(p-q)\big(1 - \pi_h(\theta_h\,|\,0) + \pi_h(1-\theta_h\,|\,0)\big) = \frac{1}{2}(p-q)\big\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\big\|_1.$$

To continue, under the condition

$$\sum_{h=1}^{H} \big\|\pi_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\big\|_1 \geq \frac{H}{8}, \tag{276}$$

applying the relation in (275) recursively yields

$$\begin{aligned}
V_1^{\star,\theta}(0) - V_1^{\pi,\theta}(0) &\geq \sum_{h=1}^{H} q^{h-1} \frac{c_2\varepsilon}{8H^2}(H+1-h)\big\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\big\|_1 \\
&= \sum_{h=1}^{H} \left(1 - \frac{c_1}{H} - \frac{c_2\varepsilon}{H^2}\right)^{h-1} \frac{c_2\varepsilon}{8H^2}(H+1-h)\big\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\big\|_1 \\
&\overset{\text{(i)}}{>} \frac{c_2\varepsilon}{16H^2} \sum_{h=1}^{H} (H+1-h)\big\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\big\|_1 \\
&= \frac{c_2\varepsilon}{16H^2} \sum_{h=1}^{H} h\big\|\pi_{H+1-h}^{\star,\theta}(\cdot\,|\,0) - \pi_{H+1-h}(\cdot\,|\,0)\big\|_1 \\
&\overset{\text{(ii)}}{\geq} \frac{c_2\varepsilon}{16H^2} \sum_{h=1}^{\lfloor H/16 \rfloor} 2h = \frac{c_2\varepsilon}{8H^2}\left\lfloor \frac{H}{16} \right\rfloor \left(\left\lfloor \frac{H}{16} \right\rfloor + 1\right).
\end{aligned} \tag{277}$$

Here, (i) follows since

$$\left(1 - \frac{c_1}{H} - \frac{c_2\varepsilon}{H^2}\right)^{h-1} \geq \left(1 - \frac{2c_1}{H}\right)^{H} > \frac{1}{2}, \qquad \text{for all } h \in [H]$$

holds as long as $0 < c_1 \leq 1/4$ and $c_2\varepsilon/H \leq c_1$. To see why (ii) is valid, we note that for any $0 \leq x_1, \cdots, x_H \leq x_{\max}$ obeying $\sum_{i=1}^{H} x_i \geq x_{\text{sum}}$, the following elementary inequality holds:

$$\sum_{i=1}^{H} x_i a_i \geq \sum_{i=1}^{\lfloor x_{\text{sum}}/x_{\max} \rfloor} x_{\max} a_i;$$

this together with $\big\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\big\|_1 \leq 2$ and (276) reveals that (by taking $a_h = h$ and $x_h = \big\|\pi_{H+1-h}^{\star,\theta}(\cdot\,|\,0) - \pi_{H+1-h}(\cdot\,|\,0)\big\|_1$)

$$\sum_{h=1}^{H} h\big\|\pi_{H+1-h}^{\star,\theta}(\cdot\,|\,0) - \pi_{H+1-h}(\cdot\,|\,0)\big\|_1 \geq \sum_{h=1}^{\lfloor H/16\rfloor} 2h,$$

thus validating inequality (ii). As a result, we can continue the derivation to obtain

$$(277) \geq \frac{c_2\varepsilon}{8H^2}\frac{\frac{H}{16}\big(\frac{H}{16}+1\big)}{2} > \varepsilon, \tag{278}$$

provided that $c_2 \geq 4096$.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2021). Reinforcement learning: Theory and algorithms. *Technical report*.

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

Agarwal, R. P., Meehan, M., and O'regan, D. (2001). *Fixed point theory and applications*, volume 141. Cambridge university press.

Auer, P. and Ortner, R. (2006). Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19.

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. In *Advances in Neural Information Processing Systems*, pages 8002–8011.

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. (2017). *Dynamic programming and optimal control (4th edition)*. Athena Scientific.

Bourel, H., Maillard, O., and Talebi, M. S. (2020). Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning*, pages 1056–1066. PMLR.

Buckman, J., Gelada, C., and Bellemare, M. G. (2020). The importance of pessimism in fixed-dataset policy optimization. In *International Conference on Learning Representations*.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR.

Chen, M., Li, Y., Wang, E., Yang, Z., Wang, Z., and Zhao, T. (2021a). Pessimism meets invariance: Provably efficient offline mean-field multi-agent RL. *Advances in Neural Information Processing Systems*, 34.

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. (2021b). A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567*.

Cui, Q. and Du, S. S. (2022). When is offline two-player zero-sum Markov game solvable? *arXiv preprint arXiv:2201.03522*.

Cui, Q. and Yang, L. F. (2021). Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence*, pages 1496–1504. PMLR.

Diehl, C., Sievernich, T., Krüger, M., Hoffmann, F., and Bertran, T. (2021). Umbrella: Uncertainty-aware model-based offline reinforcement learning leveraging planning. *arXiv preprint arXiv:2111.11097*.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. (2017). Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058. PMLR.

Duan, Y., Jia, Z., and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR.

Duan, Y., Wang, M., and Wainwright, M. J. (2021). Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*.

Duchi, J. C. (2018). Introductory lectures on stochastic optimization. *The mathematics of data*, 25:99–186.

Ebert, F., Finn, C., Dasari, S., Xie, A., Lee, A., and Levine, S. (2018). Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*.

Even-Dar, E. and Mansour, Y. (2003). Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2019). A theoretical analysis of deep Q-learning. *arXiv e-prints*, pages arXiv–1901.

Farahmand, A.-m., Szepesvári, C., and Munos, R. (2010). Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23.

Filippi, S., Cappé, O., and Garivier, A. (2010). Optimism in reinforcement learning and kullback-leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122. IEEE.

Fruit, R., Pirotta, M., and Lazaric, A. (2020). Improved analysis of UCRL2 with empirical Bernstein inequality. *arXiv preprint arXiv:2007.05456*.

Gilbert, E. N. (1952). A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522.

He, J., Zhou, D., and Gu, Q. (2021). Nearly minimax optimal reinforcement learning for discounted MDPs. *Advances in Neural Information Processing Systems*, 34:22288–22300.

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

Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR.

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, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

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.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63.

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.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lange, S., Gabel, T., and Riedmiller, M. (2012). Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer.

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

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2024a). Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):203–221.

Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. (2021). 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., and Chen, Y. (2024b). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):222–236.

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. (2022). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473.

Li, G., Yan, Y., Chen, Y., and Fan, J. (2023). Minimax-optimal reward-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2304.07278*.

Li, G., Zhan, W., Lee, J. D., Chi, Y., and Chen, Y. (2024c). Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Li, L., Munos, R., and Szepesvári, C. (2014). On minimax optimal offline policy evaluation. *arXiv preprint arXiv:1409.3653*.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2020). Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*.

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

Murphy, S. (2005). A generalization error for Q-learning. *Journal of Machine Learning Research*, 6:1073–1097.

Nguyen-Tang, T., Gupta, S., and Venkatesh, S. (2021). Sample complexity of offline reinforcement learning with deep ReLU networks. *arXiv preprint arXiv:2103.06671*.

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR.

Pananjady, A. and Wainwright, M. J. (2020). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585.

Prudencio, R. F., Maximo, M. R., and Colombini, E. L. (2022). A survey on offline reinforcement learning: Taxonomy, review, and open problems. *arXiv preprint arXiv:2203.01387*.

Qian, J., Fruit, R., Pirotta, M., and Lazaric, A. (2019). Exploration bonus for regret minimization in discrete and continuous average reward MDPs. *Advances in Neural Information Processing Systems*, 32.

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. (2022). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 68(12):8156–8196.

Ren, T., Li, J., Dai, B., Du, S. S., and Sanghavi, S. (2021). Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Shi, C., Luo, S., Zhu, H., and Song, R. (2022a). Statistically efficient advantage learning for offline reinforcement learning in infinite horizons. *arXiv preprint arXiv:2202.13163*.

Shi, L. and Chi, Y. (2022). Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*.

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

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018). 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.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359.

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

Szepesvári, C. (1998). The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, pages 1064–1070.

Talebi, M. S. and Maillard, O.-A. (2018). Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pages 770–805. PMLR.

Tang, S. and Wiens, J. (2021). Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pages 2–35. PMLR.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR.

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

Uehara, M., Huang, J., and Jiang, N. (2020). Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR.

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

Uehara, M., Zhang, X., and Sun, W. (2022). Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. (2019a). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

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

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

Wang, Y., Dong, K., Chen, X., and Wang, L. (2019). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *International Conference on Learning Representations*.

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

Xie, T. and Jiang, N. (2021). Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *arXiv preprint arXiv:2106.04895*.

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., Yang, Z., Wang, Z., and Liang, Y. (2021). A unified off-policy evaluation approach for general value function. *arXiv preprint arXiv:2107.02711*.

Yan, Y., Li, G., Chen, Y., and Fan, J. (2022). Model-based reinforcement learning is minimax-optimal for offline zero-sum Markov games. *arXiv preprint arXiv:2206.04044*.

Yan, Y., Li, G., Chen, Y., and Fan, J. (2023). The efficacy of pessimism in asynchronous Q-learning. *IEEE Transactions on Information Theory*, 69(11):7185–7219.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. (2020). Off-policy evaluation via the regularized Lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. (2022). 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., 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.

Zanette, A., Wainwright, M. J., and Brunskill, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. D. (2022). Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*.

Zhang, Z., Chen, Y., Lee, J. D., and Du, S. S. (2023). Settling the sample complexity of online reinforcement learning. *arXiv preprint arXiv:2307.13586*.

Zhang, Z., Ji, X., and Du, S. (2021a). Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. 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. (2021b). Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, pages 12653–12662. PMLR.

Zhong, H., Xiong, W., Tan, J., Wang, L., Zhang, T., Wang, Z., and Yang, Z. (2022). Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. *arXiv preprint arXiv:2202.07511*.

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR.