The Annals of Statistics
2024, Vol. 52, No. 1, 233–260
https://doi.org/10.1214/23-AOS2342

SETTLING THE SAMPLE COMPLEXITY OF MODEL-BASED OFFLINE REINFORCEMENT LEARNING

BY GEN LI[1,a], LAIXI SHI[2,b], YUXIN CHEN[3,c], YUEJIE CHI[4,e] AND YUTING WEI[3,d]

[1]*Department of Statistics, Chinese University of Hong Kong,* [a]*genli@cuhk.edu.hk*
[2]*Department of Computing Mathematical Sciences, California Institute of Technology,* [b]*laixis@caltech.edu*
[3]*Department of Statistics and Data Science, The Wharton School, University of Pennsylvania,* [c]*yuxinc@wharton.upenn.edu,* [d]*ytwei@wharton.upenn.edu*
[4]*Department of Electrical and Computer Engineering, Carnegie Mellon University,* [e]*yuejiechi@cmu.edu*

This paper is concerned with offline reinforcement learning (RL), which learns using precollected data without further exploration. Effective offline RL would be able to accommodate distribution shift and limited data coverage. However, prior results 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 any 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

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

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.

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—for example, 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 happens 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))

Received February 2023; revised November 2023.
MSC2020 subject classifications. 62C20.
Key words and phrases. Markov decision process, offline reinforcement learning, distribution shift, sample complexity, minimax optimality.

and healthcare (Tang and Wiens (2021)); see Levine et al. (2020), Prudencio, Maximo and Colombini (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, Gabel and Riedmiller (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 data set 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, Gelada and Bellemare (2020), Chen et al. (2021), Jin, Yang and Wang (2021), Kumar et al. (2020), Rashidinejad et al. (2022), Xie et al. (2021), Zanette, Wainwright and Brunskill (2021)). This is reminiscent of the optimism principle in the face of uncertainty for online exploration (Azar, Osband and Munos (2017), Jaksch, Ortner and Auer (2010), Lai and Robbins (1985)), but works for drastically different reasons (to be detailed 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 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 offline RL algorithm to achieve the smallest possible sample complexity.

Despite extensive recent activities, however, existing statistical guarantees for the above paradigm remained highly 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 2.1 and 3.1 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.

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

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

  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 (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

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

  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

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

  in other words, an enormous burn-in sample size is needed 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. (2022), 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

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

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

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

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?

This 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 $\varepsilon \in (0, \frac{1}{1-\gamma}]$ for discounted infinite-horizon MDPs and $\varepsilon \in (0, H]$ for finite-horizon MDPs.

More precisely, 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, VI-LCB with Bernstein-style penalty finds an ε-optimal policy with a sample complexity of

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

for any given accuracy level $\varepsilon \in (0, \frac{1}{1-\gamma}]$ (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. The above sample complexity (6) remains valid if $C^\star_{\mathsf{clipped}}$ is replaced by $C^\star$.

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 $C^\star_{\mathsf{clipped}} \le C^\star$; and the parameter $d^{\mathsf{b}}_{\min} := \frac{1}{\min_{s,a,h}\{d^{\mathsf{b}}_h(s,a):d^{\mathsf{b}}_h(s,a)>0\}}$ employed in Yin and Wang (2021) could be exceedingly small, with d^{b}_h the occupancy distribution of the data set. 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

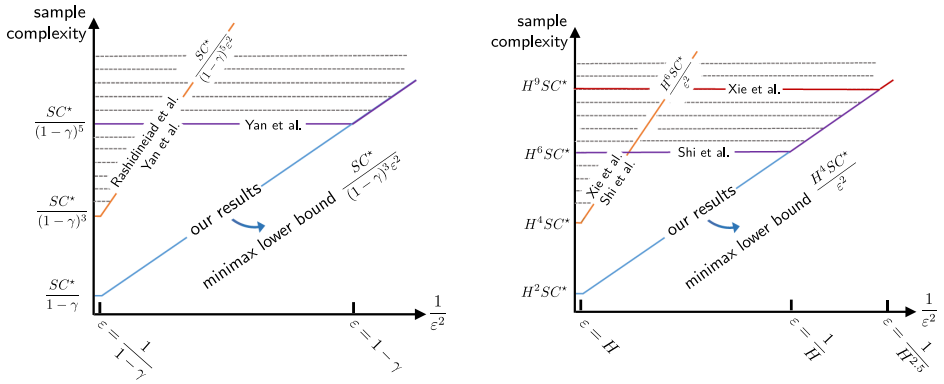
Horizon	Algorithm	Sample complexity	ε-range to attain sample optimality	Type
Infinite	VI-LCB (Rashidinejad et al. (2022))	$\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}$	–	model-based
	Q-LCB (Yan et al. (2023))	$\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}$	–	model-free
	VR-Q-LCB (Yan et al. (2023))	$\frac{SC^\star}{(1-\gamma)^3\varepsilon^2} + \frac{SC^\star}{(1-\gamma)^4\varepsilon}$	$(0, 1-\gamma]$	model-free
	VI-LCB (this paper: Theorem 1)	$\frac{SC^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}$ $(\le \frac{SC^\star}{(1-\gamma)^3\varepsilon^2})$	$(0, \frac{1}{1-\gamma}]$	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))	$\frac{H^4SC^\star}{\varepsilon^2} + \frac{H^{6.5}SC^\star}{\varepsilon}$	$(0, \frac{1}{H^{2.5}}]$	model-based
	LCB-Q-Advantage (Shi et al. (2022))	$\frac{H^4SC^\star}{\varepsilon^2} + \frac{H^5SC^\star}{\varepsilon}$	$(0, \frac{1}{H}]$	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)	$\frac{H^4SC^\star_{\mathsf{clipped}}}{\varepsilon^2}$ $(\le \frac{H^4SC^\star}{\varepsilon^2})$	$(0, H]$	model-based
	lower bound (this paper: Theorem 4)	$\frac{H^4SC^\star_{\mathsf{clipped}}}{\varepsilon^2}$	–	–

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

$$\widetilde{O}\left(\frac{H^4SC^\star_{\mathsf{clipped}}}{\varepsilon^2}\right) \tag{7}$$

  samples with high probability (see Theorem 3). A key ingredient in the algorithm design is a twofold subsampling trick that helps decouple statistical dependency along the sample rollouts. Note that the above 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 log terms).

Remarkably, our algorithms do not require sophisticated variance reduction schemes, as long as suitable confidence bounds are adopted. Detailed theoretical comparisons with prior art



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

FIG. 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* $C^\star_{\mathsf{clipped}}$ *with* $C^\star$ *in our results when drawing the plots given that* $C^\star_{\mathsf{clipped}} \le C^\star$. *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.*

can be found in Table 1 and Figure 1. Finally, we conduct a series of numerical experiments to evaluate the performance of the proposed algorithms.

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 precollected data in the face of uncertainty. As far we know, our theory is the first to identify an offline algorithm that provably attains the highest possible statistical efficiency (i.e., minimax optimality) for the entire ε-range, which in turn makes clear that no burn-in phase is needed to achieve optimal statistical accuracy, shown in Theorem 1 to Theorem 4. 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 RL as well (see 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 (2000), Wainwright (2019a)) that helps extend the frontier of classical statistics to the sample-starved regime, for which an enriched statistical toolbox is needed.

Let us single out two statistical techniques developed in this work that may of independent interest for achieving statistical efficiency. First, we have introduced a new model construction approach with twofold subsampling in Section 3.3 that permits the reuse of all samples across all iterations—compared with the previous H-fold subsampling scheme—which is an essential feature in sample-starved applications to achieve tight finite-sample guarantees. In addition, we have proposed a refined measure for the distribution coverage of offline data set, as discussed around Definition 2, which tightly determines the sample size requirement of solving offline RL problems.

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 $\{1, \dots, H\}$ 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 $a \geq b$ (resp., $a \leq b$) means $a_i \geq b_i$ (resp., $a_i \leq b_i$) for all i. Following the convention in RL, 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

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

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 MDPs. In the following, we shall first introduce the models and assumptions, followed by algorithm design and main results.

2.1. *Models and assumptions.* Consider a discounted infinite-horizon MDP represented by a tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, \gamma, r\}$. The key components of $\mathcal{M}$ are: (i) $\mathcal{S} = \{1, 2, \ldots, S\}$: a finite state space of size S; (ii) $\mathcal{A} = \{1, 2, \ldots, A\}$: 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, $r(s, a)$ is 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

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

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:

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

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:

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

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:

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

We also introduce the *discounted occupancy distributions* associated with π as follows:

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

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

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:

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

Correspondingly, the discounted occupancy distributions associated with $\pi^\star$ is denoted by

$$\begin{aligned} \forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad & d^\star(s) := d^{\pi^\star}(s;\rho), d^\star(s,a) \\ & := d^{\pi^\star}(s,a;\rho) = d^\star(s)\,\mathbb{1}\big(a = \pi^\star(s)\big), \end{aligned} \tag{16}$$

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 Rashidinejad et al. (2022). To be precise, imagine that we observe a batch data set $\mathcal{D} = \{(s_i, a_i, s_i')\}_{1\leq i\leq N}$ 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

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

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

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 data set $\mathcal{D}$ 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 data set can be viewed as expert data when $C^\star$ approaches 1, meaning that the data set is close to the target policy in terms of the induced distributions. This coefficient $C^\star$ is referred to as the "single-policy" concentrability coefficient since it concerns 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), Munos (2007), Xie and Jiang (2021)), the latter of which assumes a uniform density-ratio bound over all policies and requires the data set 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 data set $\mathcal{D}$ is defined as

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

REMARK 1. A direct comparison of conditions (18) and (19) implies that

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

for any given batch data set $\mathcal{D}$. 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))

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

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

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

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

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 $d^\star(s,a)$.

Goal. Armed with the batch data set $\mathcal{D}$, 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

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

with high probability, using a batch data set $\mathcal{D}$ (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 $\varepsilon \in (0, \frac{1}{1-\gamma}]$).

[1] As a concrete example, suppose that

$$d^\star(s) = \begin{cases} 1 - \dfrac{S-1}{S^3} & \text{if } s = 1, \\ \dfrac{1}{S^3} & \text{else,} \end{cases} \quad \text{and} \quad d^{\mathsf{b}}(s,a) = \begin{cases} 1 - \dfrac{S-1}{S^2} & \text{if } a = \pi^\star(s) \text{ and } s = 1, \\ \dfrac{1}{S^2} & \text{if } a = \pi^\star(s) \text{ and } s \ne 1, \\ 0 & \text{else.} \end{cases}$$

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 $C^\star_{\mathsf{clipped}}$.

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 data set $\mathcal{D}$. For any given state-action pair (s, a), we denote by

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

the number of samples transitions from (s, a). We then construct an empirical transition matrix $\widehat{P}$ such that: for each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,

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

The pessimistic Bellman operator. Our 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 $Q \in \mathbb{R}^{SA}$,

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

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

$$\begin{aligned} \widehat{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a) := \max\big\{r(s,a) + \gamma \widehat{P}_{s,a} V - b(s,a;V), 0\big\} \\ \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}, \end{aligned} \tag{27}$$

where $b(s, a; V)$ denotes the penalty term employed to enforce pessimism amid uncertainty. As one can anticipate, 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 variance statistics:

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

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 ($1 - \delta$ 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 isolate several useful properties below, whose proof is deferred to Appendix B.2 in the Supplementary Material (Li et al. (2024)).

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* $Q_1(s,a), Q_2(s,a) \in [0, \frac{1}{1-\gamma}]$ *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, thus 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:

$$\begin{aligned}\widehat{Q}_\tau(s,a) &= \widehat{\mathcal{T}}_{\mathsf{pe}}(\widehat{Q}_{\tau-1})(s,a)\\ &= \max\big\{r(s,a) + \gamma \widehat{P}_{s,a}\widehat{V}_{\tau-1} - b(s,a;\widehat{V}_{\tau-1}), 0\big\}, \quad \tau = 1,2,\ldots\end{aligned} \tag{30}$$

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:

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

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.

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. Its proof is deferred to Appendix B.3 in the Supplementary Material (Li et al. (2024)).

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

input: data set $\mathcal{D}$; 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 $\tau = 1, 2, \ldots, \tau_{\max}$ **do**
  for $s \in \mathcal{S}$, $a \in \mathcal{A}$ **do**
    compute the penalty term $b(s,a;\widehat{V}_{\tau-1})$ according to (28).
    set $\widehat{Q}_\tau(s,a) = \max\{r(s,a) + \gamma \widehat{P}_{s,a}\widehat{V}_{\tau-1} - b(s,a;\widehat{V}_{\tau-1}), 0\}$.
  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 2. *Suppose* $\widehat{Q}_0 = 0$. *Then the iterates of Algorithm 1 obey*

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

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

$$\big\|\widehat{Q}_{\tau_{\max}} - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \le 1/N. \tag{33}$$

Algorithmic comparison with Rashidinejad et al. (2022). VI-LCB has been studied in Rashidinejad et al. (2022). The difference between our version and theirs is twofold:

- *Sample reuse vs.* $\widetilde{O}(\frac{1}{1-\gamma})$*-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 $\widetilde{O}(\frac{1}{1-\gamma})$ 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, Ortner and Auer (2010) put forward an algorithm called UCRL2 that computes an optimistic policy with the aid of Hoeffding-style confidence regions for the transition kernel. Later on, Azar, Osband and Munos (2017) proposed to build upper confidence bounds (UCB) for the optimal values instead, which leads to improved sample complexity; see, for example, He, Zhou and Gu (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 markedly 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, Algorithm 1 yields ε-accuracy with a minimal number of samples, as stated below.

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

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

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

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

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

REMARK 2. Regarding the numerical constants in Theorem 1, a conservative yet concrete sufficient condition is $c_{\mathsf{b}} \geq 144$ and $c_1 = 21{,}000 c_{\mathsf{b}}$, which we shall rigorize in the proof.

The proof of this theorem is postponed to Section A in the Supplementary Material (Li et al. (2024)). 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 confirm the efficacy of our algorithm, whose proof can be found in Appendix E.2 in the Supplementary Material (Li et al. (2024)).

THEOREM 2. *For any* $(\gamma, S, C^\star_{\mathsf{clipped}}, \varepsilon)$ *obeying* $\gamma \in [\frac{2}{3}, 1)$, $S \geq 2$, $C^\star_{\mathsf{clipped}} \geq \frac{8\gamma}{S}$ *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 data set with* N *independent samples and single-policy clipped concentrability coefficient* $C^\star_{\mathsf{clipped}}$ *such that*

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

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 highly conservative) condition for c_2, Theorem 2 is valid when $c_2 = 1/25{,}088$.

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

$$\widetilde{O}\Big(\frac{S C^\star_{\mathsf{clipped}}}{(1-\gamma)^3 \varepsilon^2}\Big). \tag{36}$$

  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 $C^\star$, that is, $C^\star_{\mathsf{clipped}} = O(1/S)$.
- *No burn-in cost.* The fact that the sample size bound (35) holds for the full ε-range (i.e., any given $\varepsilon \in (0, \frac{1}{1-\gamma}]$) 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 $\widetilde{O}(\frac{1}{1-\gamma})$ 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

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

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 $\widetilde{O}(\frac{S}{(1-\gamma)^3\varepsilon^2})$ even in the presence of a simulator (Gheshlaghi Azar, Munos and Kappen (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, Zhou and Gu (2021), Zhang, Zhou and Ji (2021)), 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, Zhou and Gu (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.* 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) $\mathcal{S} = \{1, \ldots, S\}$: a state space of size S; (ii) $\mathcal{A} = \{1, \ldots, A\}$: 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:

$$P_{h,s,a} := P_h(\cdot|s,a), \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{37}$$

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

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

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

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

when conditioned on the state-action (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 weighted value function of π:

$$V_1^\pi(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V_1^\pi(s)\right]. \tag{40}$$

We also introduce the following *occupancy distributions* associated with policy π at step h:

$$d_h^\pi(s;\rho) := \mathbb{P}(s_h = s | s_1 \sim \rho; \pi), \tag{41a}$$

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

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

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

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 $Q^\star = \{Q_h^\star\}_{1\le h\le H}$:

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

Further, we adopt the following notation for convenience: for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$d_h^\star(s) := d_h^{\pi^\star}(s;\rho) \quad \text{and} \quad d_h^\star(s,a) := d_h^{\pi^\star}(s,a;\rho) = d_h^\star(s)\mathbb{1}\{a = \pi^\star(s)\}, \tag{43}$$

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

Offline/batch data. Suppose we have access to a batch data set (or historical data set) $\mathcal{D}$, which comprises a collection of K i.i.d. sample trajectories generated by a behavior policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{1\le h\le H}$. The kth 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:

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

Here and throughout, ρ^{b} stands for some predetermined initial state distribution associated with the batch data set. In addition to the above data set (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 shorthand notation for the occupancy distribution w.r.t. the behavior policy π^{b}:

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

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 data set 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 data set $\mathcal{D}$ is defined as

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

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 versus 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 data set $\mathcal{D}$ is defined as

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

From the definition above, it holds trivially that

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

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 data set $\mathcal{D}$ 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, that is, for any $\varepsilon \in (0, H]$.

3.2. *A model-based offline RL algorithm*: *VI-LCB*. Suppose for the moment that we have a data set $\mathcal{D}_0$ 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

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

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 P as follows:

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

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, Yang and Wang (2021), Xie et al. (2021)). Specifically, the algorithm adopts the following update rule:

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

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

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

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

$$b_h(s,a) = \min\left\{\sqrt{\frac{c_{\mathsf{b}} \log\frac{NH}{\delta}}{N_h(s,a)} \mathsf{Var}_{\widehat{P}_{h,s,a}}(\widehat{V}_{h+1})} + c_{\mathsf{b}} H \frac{\log\frac{NH}{\delta}}{N_h(s,a)}, H\right\} \tag{55}$$

for some universal constant $c_{\mathsf{b}} > 0$ (e.g., $c_{\mathsf{b}} = 16$). Here, $\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widehat{V}_{h+1})$ 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., $\mathsf{Var}_{P_{h,s,a}}(\widehat{V}_{h+1})$) in the variance term, mainly because we have no access

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

input: data set $\mathcal{D}_0$; reward function r; target success probability $1-\delta$.
initialization: $\widehat{V}_{H+1} = 0$.
for $h = H, \ldots, 1$ **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 $\widehat{Q}_h(s,a) = \max\{r_h(s,a) + \widehat{P}_{h,s,a}\widehat{V}_{h+1} - b_h(s,a), 0\}$.
  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: $\widehat{\pi} = \{\widehat{\pi}_h\}_{1\le h\le H}$.

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, Osband and Munos (2017), Fruit, Pirotta and Lazaric (2020), Jin et al. (2018), Li et al. (2021), Talebi and Maillard (2018), Zhang, Zhou and Ji (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 twofold subsampling*. Given that the batch data set $\mathcal{D}$ 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.

To circumvent this technical difficulty, we propose a twofold subsampling trick that allows one to exploit desired statistical independence. Informally, we propose the steps below:

- First of all, we randomly split the data set 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.
- For each $(s,h) \in \mathcal{S}\times[H]$, we use the data set $\mathcal{D}^{\mathsf{aux}}$ 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 data set $\mathcal{D}^{\mathsf{trim}}$.
- Run VI-LCB on the subsampled data set $\mathcal{D}^{\mathsf{trim}}$ (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 data set $\mathcal{D}^{\mathsf{aux}}$ and that $\mathcal{D}^{\mathsf{trim}}$ is subsampled from another data set $\mathcal{D}^{\mathsf{main}}$, 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 see in the proof (i.e., Appendix C.2 in the Supplementary Material (Li et al. (2024))), 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.

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

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

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

(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

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

(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 $\widehat{\pi}$.

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 D.1 in the Supplementary Material (Li et al. (2024)).

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*

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

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

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 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).

The proof of this result is postponed to Appendix C in the Supplementary Material (Li et al. (2024)). 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 E.3 in the Supplementary Material (Li et al. (2024)).

THEOREM 4. *For any $(H, S, C^\star_{\text{clipped}}, \varepsilon)$ obeying $H \geq 12$, $C^\star_{\text{clipped}} \geq 8/S$ and $\varepsilon \leq c_3 H$, one can construct a collection of MDPs $\{\mathcal{M}_\theta | \theta \in \Theta\}$, an initial state distribution ρ and a batch data set with K independent sample trajectories each of length H, such that*

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

provided that the total sample size

$$N = KH \leq \frac{c_4 C^\star_{\text{clipped}} S H^4}{\varepsilon^2}. \tag{61}$$

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 $\mathcal{M}_\theta$.

Implications. 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

$$\widetilde{O}\Big(\frac{H^4 S C^\star_{\text{clipped}}}{\varepsilon^2} \Big). \tag{62}$$

  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 $\frac{H^4 S C^\star_{\text{clipped}}}{\varepsilon^2}$ 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 ($H^9 S C^\star$ and $H^6 S C^\star$, respectively).
- *Sample reduction and model compressibility when $C^\star_{\text{clipped}} < 1$.* Given that $C^\star_{\text{clipped}}$ might drop below 1, the sample complexity of our algorithm might be as low as $\widetilde{O}(\frac{H^4 S}{\varepsilon^2})$. In fact, recognizing that $C^\star_{\text{clipped}}$ can be as small as $\frac{1+o(1)}{S}$, we see that the sample complexity can sometimes be reduced to

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

  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 $C^\star_{\text{clipped}}$. For instance, $C^\star_{\text{clipped}} = O(1/S)$ might

happen when a small number of states accounts for a dominant fraction of probability mass in $d_h^\star(s)$, 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 data set. 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

$$r_1(0,0) = 0, \qquad r_1(0,1) = 0, \qquad r_2(0,0) = 0, \qquad r_2(1,0) = 1,$$
$$P_1(0|0,0) = 0.5, \qquad P_1(1|0,0) = 0.5, \qquad P_1(0|0,1) = p, \qquad P_1(1|0,1) = 1-p$$

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

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

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., $(1 - d_1^{\mathsf{b}}(0,1))^K \approx 1/e$ for large K) such that the data set 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 data set.

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 will not 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:

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

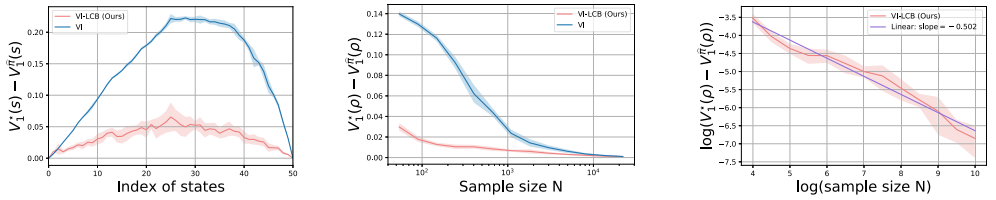
with the proviso that $K \ge 100 c_{\mathsf{b}} H S C^\star \log \frac{NH}{\delta}$. Encouragingly, the dominate term (i.e., the first term in the bound (65)) matches the instance-dependent lower bound established in Yin and Wang ((2021), Theorem 4.3), thus confirming the instance optimality of the proposed algorithm for a large enough sample size. The proof of (65) can be found in Appendix D.2 in the Supplementary Material (Li et al. (2024)).

Comparisons with prior statistical analysis. Let us discuss the novelty of our statistical analysis. 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, Zhou and Ji (2020)). In contrast, our theory makes apparent that variance reduction is unnecessary, which leads to both simpler algorithm and tighter analysis. This also confirms 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 $\{0, 1, \ldots, 50\}$ and the associated accessible actions obey $a \in \{0, 1, \ldots, \min\{s, 50 - s\}\}$, $H = 100$ 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 data set 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 VI-LCB method (cf. Algorithm 2) with comparisons to the well-known value iteration (VI) method without the pessimism principle. To begin with, Figure 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. Second, we evaluate the



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

FIG. 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.

performance gap $V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho)$ with varying sample size $N \in \{54, 90, 148, \ldots, 22{,}026\} \approx \{e^4, e^{4.5}, e^5, \ldots, e^{10}\}$, over 10 independent trials. Note that throughout the experiments, we fix $c_{\mathsf{b}} = 0.05$, which determines the level of the pessimism penalty of VI-LCB (cf. (55)). Figure 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 suboptimality performance gap of VI-LCB w.r.t. the sample size on a log-log scale in Figure 2(c). Fitting using linear regression leads to a slope estimate of -0.502, with the corresponding fitted line plotted also in Figure 2(c). 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.

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 differs from the behavior policy used in data collection; and offline policy learning, where the goal is to identify a near-optimal policy. Our work falls under the second category. Note that off-policy evaluation has been extensively studied (Duan, Jia and Wang (2020), Jiang and Li (2016), Kallus and Uehara (2020), Li, Munos and Szepesvári (2014)); we excuse ourselves from enumerating the works in that space.

Offline RL with the pessimism principle. The prior works that are the most relevant to this paper are Jin, Yang and Wang (2021), Rashidinejad et al. (2022), Shi et al. (2022), 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. (2021), 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, Wainwright and Brunskill (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, for example, Jin, Yang and Wang (2021), Nguyen-Tang, Gupta and Venkatesh (2021), Uehara and Sun (2021), Yin et al. (2021), Zhan et al. (2022), which are beyond the scope of the current paper.

Online RL and the optimism principle. The optimism principle in the face of uncertainty has received widespread adoption from bandits to online RL (Lai and Robbins (1985), Lattimore and Szepesvári (2020)). In the context of online RL, Jaksch, Ortner and Auer (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 Auer and Ortner (2006); see also Bourel, Maillard and

Talebi (2020), Filippi, Cappé and Garivier (2010), Talebi and Maillard (2018) for other variants of UCRL. When applied to episodic finite-horizon MDPs, the regret bound in Jaksch, Ortner and Auer (2010) was suboptimal by a factor of at least $\sqrt{H^2S}$; see the discussion in Azar, Osband and Munos (2017), Jin et al. (2018). Fruit, Pirotta and Lazaric (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, Osband and Munos (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, Zhou and Ji (2020)). Our investigation of offline RL has inspired new algorithms for both online RL and hybrid RL (Li et al. (2023a, 2023b)).

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, Kakade and Yang (2020), Gheshlaghi Azar, Munos and Kappen (2013), Li et al. (2023c), Li et al. (2020)) and the online exploratory setting (Azar, Osband and Munos (2017), Jin et al. (2020), Zhang et al. (2023), Zhang, Ji and Du (2021)). 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, Kakade and Yang (2020), Cui and Yang (2021), Li et al. (2023c), Pananjady and Wainwright (2021), and has recently been shown to be effective for multiagent offline RL (Yan et al. (2022)) and distributionally robust RL (Shi et al. (2023)) as well.

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. (2023a), Li et al. (2022), Qu and Wierman (2020), Shi et al. (2022), 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. (2021), Li et al. (2023a), Li et al. (2022), Qu and Wierman (2020)), which is stronger than the single-policy concentrability considered herein. Moreover, Q-learning alone is known to be suboptimal in terms of the sample complexity in various settings (Bai et al. (2019), Jin et al. (2018), Li et al. (2023a), Shi et al. (2022)). This motivates the incorporation of the variance reduction in order to further improve the sample complexity (Du et al. (2017), Li et al. (2021), Li et al. (2022), Shi et al. (2022), Wainwright (2019b), Yan et al. (2023), Zhang, Zhou and Ji (2020), Zhang, Zhou and Ji (2021)). 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. Discussion. Our primary contribution has been to pin down the sample complexity of model-based offline RL for the tabular settings, by establishing its 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 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 suboptimality 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 confidence bounds are carefully employed to capture the estimation variance in each iteration.

Turning to future directions, we first note that the twofold 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 orderwise sense, the potential gain in terms of the preconstants as well as the algorithmic 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.

Acknowledgments. Y. Wei is supported by the Google Research Scholar Award, and NSF Grants CCF-2106778, DMS-2147546/2015447 and CAREER award DMS-2143215.

Y. Chen is supported by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grants FA9550-22-1-0198, the ONR Grant N00014-22-1-2354 and NSF Grants CCF-2221009, CCF-1907661, DMS-2014279, IIS-2218713 and IIS-2218773.

L. Shi and Y. Chi are supported by the grants ONR N00014-19-1-2404, NSF CCF-2106778 and DMS-2134080, and CAREER award ECCS-1818571. L. Shi was also gratefully supported by the Leo Finzi Memorial Fellowship, Wei Shen and Xuehong Zhang Presidential Fellowship and Liang Ji-Dian Graduate Fellowship at CMU.

Y. Wei is the corresponding author.

SUPPLEMENTARY MATERIAL

Supplement to "Settling the sample complexity of model-based offline reinforcement learning" (DOI: 10.1214/23-AOS2342SUPP; .pdf). Supplementary information.

REFERENCES

AGARWAL, A., KAKADE, S. and YANG, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory* 67–83.

AUER, P. and ORTNER, R. (2006). Logarithmic online regret bounds for undiscounted reinforcement learning. *Adv. Neural Inf. Process. Syst.* **19**.

AZAR, M. G., OSBAND, I. and MUNOS, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning* 263–272.

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* 8002–8011.

BECK, C. L. and SRIKANT, R. (2012). Error bounds for constant step-size Q-learning. *Systems Control Lett.* **61** 1203–1208. MR2998204 https://doi.org/10.1016/j.sysconle.2012.08.014

BERTSEKAS, D. P. (2017). *Dynamic Programming and Optimal Control. Vol. I*, 4th ed. Athena Scientific, Belmont, MA. MR3644954

BOUREL, H., MAILLARD, O. and TALEBI, M. S. (2020). Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning* 1056–1066.

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* 1042–1051.

CHEN, M., LI, Y., WANG, E., YANG, Z., WANG, Z. and ZHAO, T. (2021). Pessimism meets invariance: Provably efficient offline mean-field multi-agent RL. *Adv. Neural Inf. Process. Syst.* **34**.

CHEN, Z., MAGULURI, S. T., SHAKKOTTAI, S. and SHANMUGAM, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *NeurIPS* **33** 8223–8234.

CHEN, Z., MAGULURI, S. T., SHAKKOTTAI, S. and SHANMUGAM, K. (2021). A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. arXiv preprint. Available at arXiv:2102.01567.

CUI, Q. and YANG, L. F. (2021). Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence* 1496–1504.

DIEHL, C., SIEVERNICH, T., KRÜGER, M., HOFFMANN, F. and BERTRAN, T. (2021). Umbrella: Uncertainty-aware model-based offline reinforcement learning leveraging planning. arXiv preprint. Available at arXiv:2111.11097.

DONOHO, D. (2000). *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*. AMS Math Challenges Lecture.

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* 1049–1058.

DUAN, Y., JIA, Z. and WANG, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning* 2701–2709.

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. Available at arXiv:1812.00568.

EVEN-DAR, E. and MANSOUR, Y. (2003). Learning rates for Q-learning. *J. Mach. Learn. Res.* **5** 1–25.

FILIPPI, S., CAPPÉ, O. and GARIVIER, A. (2010). Optimism in reinforcement learning and Kullback–Leibler divergence. In *Allerton Conference on Communication, Control, and Computing* 115–122. IEEE, Los Alamitos.

FRUIT, R., PIROTTA, M. and LAZARIC, A. (2020). Improved analysis of UCRL2 with empirical Bernstein inequality. arXiv preprint. Available at arXiv:2007.05456.

GHESHLAGHI AZAR, M., MUNOS, R. and KAPPEN, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.* **91** 325–349. MR3064431 https://doi.org/10.1007/s10994-013-5368-1

HE, J., ZHOU, D. and GU, Q. (2021). Nearly minimax optimal reinforcement learning for discounted MDPs. *Adv. Neural Inf. Process. Syst.* **34** 22288–22300.

JAKSCH, T., ORTNER, R. and AUER, P. (2010). Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.* **11** 1563–1600. MR2645461

JIANG, N. and LI, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning* 652–661.

JIN, C., ALLEN-ZHU, Z., BUBECK, S. and JORDAN, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems* 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* 2137–2143.

JIN, Y., YANG, Z. and WANG, Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning* 5084–5096.

JOHNSON, R. and ZHANG, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems* 315–323.

KALLUS, N. and UEHARA, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Mach. Learn. Res.* **21** 167. MR4209453

KIDAMBI, R., RAJESWARAN, A., NETRAPALLI, P. and JOACHIMS, T. (2020). MOReL: Model-based offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **33** 21810–21823.

KUMAR, A., ZHOU, A., TUCKER, G. and LEVINE, S. (2020). Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems* **33** 1179–1191.

LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. MR0776826 https://doi.org/10.1016/0196-8858(85)90002-8

LANGE, S., GABEL, T. and RIEDMILLER, M. (2012). Batch reinforcement learning. In *Reinforcement Learning* 45–73. Springer, Berlin.

LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit Algorithms*. Cambridge Univ. Press, Cambridge.

LEVINE, S., KUMAR, A., TUCKER, G. and FU, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint. Available at arXiv:2005.01643.

LI, G., CAI, C., CHEN, Y., WEI, Y. and CHI, Y. (2023a). Is Q-learning minimax optimal? A tight sample complexity analysis. *Oper. Res.*

LI, G., SHI, L., CHEN, Y., CHI, Y. and WEI, Y. (2024). Supplement to "Settling the sample complexity of model-based offline reinforcement learning." https://doi.org/10.1214/23-AOS2342SUPP

LI, G., SHI, L., CHEN, Y., GU, Y. and CHI, Y. (2021). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34**.

LI, G., WEI, Y., CHI, Y. and CHEN, Y. (2023c). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Oper. Res.*

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* **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 Trans. Inf. Theory* **68** 448–473. MR4395423 https://doi.org/10.1109/tit.2021.3120096

LI, G., YAN, Y., CHEN, Y. and FAN, J. (2023a). Minimax-optimal reward-agnostic exploration in reinforcement learning. arXiv preprint. Available at arXiv:2304.07278.

LI, G., ZHAN, W., LEE, J. D., CHI, Y. and CHEN, Y. (2023b). Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. *Adv. Neural Inf. Process. Syst.*

LI, L., MUNOS, R. and SZEPESVÁRI, C. (2014). On minimax optimal offline policy evaluation. Available at arXiv:1409.3653.

MUNOS, R. (2007). Performance bounds in L_p-norm for approximate value iteration. *SIAM J. Control Optim.* **46** 541–561. MR2309039 https://doi.org/10.1137/040614384

NGUYEN-TANG, T., GUPTA, S. and VENKATESH, S. (2021). Sample complexity of offline reinforcement learning with deep ReLU networks. arXiv preprint. Available at 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* 9582–9602.

PANANJADY, A. and WAINWRIGHT, M. J. (2021). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Trans. Inf. Theory* **67** 566–585. MR4231973 https://doi.org/10.1109/TIT.2020.3027316

PRUDENCIO, R. F., MAXIMO, M. R. and COLOMBINI, E. L. (2022). A survey on offline reinforcement learning: Taxonomy, review, and open problems. arXiv preprint. Available at arXiv:2203.01387.

QU, G. and WIERMAN, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Conference on Learning Theory* 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 Trans. Inf. Theory* **68** 8156–8196. MR4544936 https://doi.org/10.1109/tit.2022.3185139

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

SHI, C., LUO, S., LE, Y., ZHU, H. and SONG, R. (2022). Statistically efficient advantage learning for offline reinforcement learning in infinite horizons. *J. Amer. Statist. Assoc.* 1–14.

SHI, L. and CHI, Y. (2022). Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. arXiv preprint. Available at arXiv:2208.05767.

SHI, L., LI, G., WEI, Y., CHEN, Y. and CHI, Y. (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning* 19967–20025.

SHI, L., LI, G., WEI, Y., CHEN, Y., GEIST, M. and CHI, Y. (2023). The curious price of distributional robustness in reinforcement learning with a generative model. *Adv. Neural Inf. Process. Syst.*

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 *NeurIPS* 5186–5196.

SILVER, D., SCHRITTWIESER, J., SIMONYAN, K., ANTONOGLOU, I., HUANG, A., GUEZ, A., HUBERT, T., BAKER, L., LAI, M. et al. (2017). Mastering the game of Go without human knowledge. *Nature* **550** 354–359.

SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning*: *An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

SZEPESVÁRI, C. (1998). The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems* 1064–1070.

TALEBI, M. S. and MAILLARD, O.-A. (2018). Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *Algorithmic Learning Theory* 2018. *Proc. Mach. Learn. Res.* (*PMLR*) **83** 770–805. MR3857329

TANG, S. and WIENS, J. (2021). Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference* 2–35.

UEHARA, M. and SUN, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. arXiv preprint. Available at arXiv:2107.06226.

WAINWRIGHT, M. J. (2019a). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge Univ. Press, Cambridge.

WAINWRIGHT, M. J. (2019b). Variance-reduced Q-learning is minimax optimal. Available at 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. *Mach. Learn.* **8** 279–292.

XIE, T. and JIANG, N. (2021). Batch value-function approximation with only realizability. In *International Conference on Machine Learning* 11404–11413.

XIE, T., JIANG, N., WANG, H., XIONG, C. and BAI, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 27395–27407.

XIONG, H., ZHAO, L., LIANG, Y. and ZHANG, W. (2020). Finite-time analysis for double Q-learning. *Adv. Neural Inf. Process. Syst.* **33**.

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. Available at arXiv:2206.04044.

YAN, Y., LI, G., CHEN, Y. and FAN, J. (2023). The efficacy of pessimism in asynchronous Q-learning. *IEEE Trans. Inf. Theory* **69** 7185–7219. MR4660859

YIN, M., DUAN, Y., WANG, M. and WANG, Y.-X. (2021). 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. *Adv. Neural Inf. Process. Syst.* **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. *Adv. Neural Inf. Process. Syst.* **33** 14129–14142.

ZANETTE, A., WAINWRIGHT, M. J. and BRUNSKILL, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34**.

ZHAN, W., HUANG, B., HUANG, A., JIANG, N. and LEE, J. (2022). Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory* 2730–2775.

ZHANG, Z., CHEN, Y., LEE, J. D. and DU, S. S. (2023). Settling the sample complexity of online reinforcement learning. arXiv preprint, 2307.13586.

ZHANG, Z., JI, X. and DU, S. (2021). Is reinforcement learning more difficult than bandits? A near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory* 4528–4531.

ZHANG, Z., ZHOU, Y. and JI, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Adv. Neural Inf. Process. Syst.* **33**.

ZHANG, Z., ZHOU, Y. and JI, X. (2021). Model-free reinforcement learning: From clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning* 12653–12662.

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 *AISTATS* 3331–3339.