Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning

Eric Xia, Koulik Khamaru, Martin J. Wainwright, *Senior Member, IEEE*, and Michael I. Jordan, *Fellow, IEEE*

***Abstract*—Various algorithms in reinforcement learning exhibit dramatic variability in their convergence rates and ultimate accuracy as a function of the problem structure. Such instance-specific behavior is not captured by existing global minimax bounds, which are worst-case in nature. We analyze the problem of estimating optimal Q-state-action value functions for a discounted Markov decision process with discrete states and actions; our main result is to identify an instance-dependent functional that controls the difficulty of estimation in the ℓ_∞-norm. Using a local minimax framework, we show that this functional arises in lower bounds on the accuracy on any estimation procedure. We establish the sharpness of these lower bounds, up to factors logarithmic in the state and action spaces, by analyzing a variance-reduced version of Q-learning. Our theory provides a precise way of distinguishing "easy" problems from "hard" ones in the context of Q-learning, as illustrated by an ensemble with a continuum of difficulty.**

***Index Terms*—Reinforcement learning; Q-learning; stochastic control; minimax lower bounds; instance-dependent complexity; variance reduction.**

I. Introduction

THE need for data-driven decision-making has fueled tremendous interest in Markov decision processes and reinforcement learning (RL). Indeed, such techniques have found use cases across a wide range of application domains (e.g., [TFR+17, LFDA16, SHM+16]). An intriguing fact is that in many applications, RL algorithms behave far better than the theoretical bounds provided by worst-case analyses would suggest. This gap provides impetus for a more refined *instance-specific* analysis, one which highlights the properties of a given instance that render it "easy" or "difficult."

Instance-dependent analysis of RL algorithms has become of substantial interest in recent years [see, e.g., SJ19, ZB19, ZKB19, MMM14, PW20, KPR+21]. By now, we have a fairly refined understanding of instance-dependence for policy evaluation problems, including ℓ_2-norm bounds on temporal difference (TD) algorithms [BRS18, LS18, DSTM18], instance-dependent ℓ_2-bounds on linear stochastic approximation for Markovian data [MPWB23], as well as bounds for the least-squares temporal difference (LSTD) estimator in the ℓ_∞-norm [PW20]. For the linear problem of evaluating a given policy, a subset of the current authors [KPR+21] provided a sharper instance-dependent ℓ_∞-bounds for a variance-reduced version of the TD(0) algorithm, and showed that this algorithm is optimal in a local non-asymptotic minimax sense.

For TD and LSTD methods, the underlying structure is linear in nature—in particular, it corresponds to solving a linear system—a property which greatly facilitates the analysis. In the current paper we undertake a similar instance-dependent analysis in the more challenging setting of Q-learning, for which the underlying updates are non-linear. Our main contributions are to identify a natural functional of the problem instance and show that it controls the fundamental difficulty of estimating optimal Q-value functions. We do so by establishing non-asymptotic lower bounds within a local minimax framework and matching those bounds, up to logarithmic factors, by analyzing a version of variance-reduced Q-learning [SWW+18, SWWY18, Wai19c].

This work is done in the context of Markov decision processes (MDPs) with a finite set of states $\mathcal{S}$ and a finite

The work of EX was supported in part by a research fellowship from the NSF Graduate Research Fellowship Program. MJW and EX were partially funded by Mathematical Data Science program of the Office of Naval Research under ONR grant N00014-21-1-2842; National Science Foundation grant NSF-DMS grant 2015454, and National Science Foundation grant NSF-CCF grant 1955450. The work of KK was supported by National Science Foundation grant NSF-DMS grant 2311304. In addition, this work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-21-1-2840 to MIJ. (*Corresponding author: Eric Xia*).

Eric Xia is with the Department of EECS, MIT, Cambridge, MA 02139 USA (e-mail: ericzxia@mit.edu).

Koulik Khamaru is with the Department of Statistics at Rutgers Univeresity, Piscataway, NJ 08854.

Martin J. Wainwright is with the Department of EECS and Department of Mathematics at MIT, Cambridge, MA 02139 USA, and also with the Department of EECS and Department of Statistics, University of California Berkeley, Berkeley, CA 94720.

Michael I. Jordan is with the Department of EECS and Department of Statistics, University of California Berkeley, Berkeley, CA 94720.

set of possible actions $\mathcal{A}$. We proceed to provide some background and notation so as to be able to introduce the functional that plays a central role in our analysis, and describe our contributions in more detail.

A. Some background

In a Markov decision process, the state s evolves dynamically in time under the influence of the actions. More precisely, there is a collection of probability transition kernels, $\{\mathbf{P}_a(\cdot \mid s) \mid (s,a) \in \mathcal{S} \times \mathcal{A}\}$, where $\mathbf{P}_a(s' \mid s)$ denotes the probability of transitioning to state s' when the action a is taken at the current state s. In addition, an MDP is equipped with a reward function r that maps every state-action pair (s,a) to a real number $r(s,a)$. The reward $r(s,a)$ is the reward received upon performing an action a in the state s. Overall, a given MDP is characterized by the problem pair $(\mathbf{P}, r)$, along with a discount factor $\gamma \in (0,1)$.

A deterministic policy π is a mapping $\mathcal{S} \to \mathcal{A}$, such that $\pi(s) \in \mathcal{A}$ indicates the action to be taken in the state s. The value of a policy is defined by the expected sum of discounted rewards in an infinite sample path. For a given policy π and discount factor $\gamma \in (0,1)$, the Q-function is given by

$$Q^\pi(s,a) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \mid s_0 = s,\, a_0 = a \right], \quad \text{where } a_k = \pi(s_k) \text{ for all } k \geqslant 1. \tag{1}$$

When both the state space $\mathcal{S}$ and action space $\mathcal{A}$ are finite, the Q-function Q can be conveniently represented as an element of $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. It is possible to consider randomized policies but for the goal of estimating the optimal Q-function, the distinction between randomized and deterministic policies is irrelevant.

There are various observation models in reinforcement learning, and in this paper we study the *generative setting* in which we have the ability to draw next-state samples from the MDP when initialized with an arbitrary state-action pair (s,a). More precisely, we are given a collection of N i.i.d. samples of the form $\{(\mathbf{Z}_k, R_k)\}_{k=1}^N$, where R_k is a matrix in $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, and $\mathbf{Z}_k$ is a collection of $|\mathcal{A}|$ matrices in $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ indexed by $\mathcal{A}$. We denote by $\mathbf{Z}_{k,a}(\cdot \mid s)$ the row-vector corresponding to a transition starting from state s and action a; computed by sampling from the transition kernel $\mathbb{P}_a(\cdot \mid s)$, independently of all other state action pairs (s,a) and making the entry corresponding to the next state 1, and the remaining entries 0. Concretely, we write

$$s' \sim \mathbb{P}_a(\cdot \mid s) \qquad \text{and} \qquad \mathbf{Z}_{k,a}(\cdot \mid s) = \mathbf{1}_{s=s'}.$$

For convenience, we may drop the dependence on k when it is clear we are referring to a single sample. The entry $\mathbf{Z}_k(s,a)$ is drawn according to the transition kernel $\mathbf{P}_a(\cdot \mid s)$, whereas the entry $R_k(s,a)$ is a random variable with mean $r(s,a)$ and σ_r-sub-Gaussian tails, corresponding to a noisy observation of the reward function. Here the rewards $\{R_k(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ are independent across the all state-action pairs, and the random rewards $\{R_k\}$ are independent of the randomness in $\{\mathbf{Z}_k\}$.

Based on the observations, our goal is to estimate the optimal state-action-value function $Q^\star$, along with an optimal policy $\pi^\star$. From the classical theory of MDPs [Put14, SB18, Ber09], the optimal Q-function is a fixed point of the Bellman (optimality) operator $\mathbf{T}$, a map from $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ to itself given by

$$\mathbf{T}(Q)(s,a) := r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}_a(s' \mid s) \max_{a'\in\mathcal{A}} Q(s',a'). \tag{2}$$

An optimal policy $\pi^\star$ can be obtained from the optimal Q-function $Q^\star$ via the maximization $\pi^\star(s) \in \arg\max_{a\in\mathcal{A}} Q^\star(s,a)$. In this paper, we measure the quality of a given estimate $\widehat{Q}$ in terms of the ℓ_∞-norm error

$$\|\widehat{Q} - Q^\star\|_\infty = \max_{(s,a)} |\widehat{Q}(s,a) - Q^\star(s,a)|. \tag{3}$$

B. Contributions of this paper

The main contribution of this paper is to show that for a given MDP, the difficulty of estimating the optimal Q-value function in ℓ_∞-norm is characterized by a particular functional of the problem instance $(\mathbf{P}, r)$, defined here.

a) An instance-dependent functional: Given a sample $(\mathbf{Z}, R)$ from our observation model, we can define the single-sample empirical Bellman operator evaluated at (s,a) as

$$\widehat{\mathbf{T}}(Q)(s,a) := R(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathbf{Z}_a(s' \mid s) \max_{a'\in\mathcal{A}} Q(s',a'), \tag{4}$$

where we have introduced the notation $\mathbf{Z}_a(s' \mid s) := \mathbb{1}_{\mathbf{Z}(s,a)=s'}$.

Note that for any fixed Q-function Q, the difference $\widehat{\mathbf{T}}(Q) - \mathbf{T}(Q)$ is a zero-mean random matrix, and a key object in this paper is the matrix $\nu \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ with entries

$$\nu(\pi;\mathbf{P},r,\gamma)(s,a) := \sqrt{\mathrm{Var}\left(\left(\mathbf{I} - \gamma \mathbf{P}^\pi\right)^{-1} \left((\widehat{\mathbf{T}} - \mathbf{T})(Q^\star) \right)(s,a) \right)}. \tag{5}$$

More explicitly, the quantity $\mathbf{P}^\pi$ is a right-linear mapping of $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ to itself, given by:

$$\mathbf{P}^\pi Q(s,a) := \sum_{s'\in\mathcal{S}} \mathbf{P}_a(s' \mid s) \cdot Q(s', \pi(s')) \quad \text{for each } (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{6}$$

and the square-root and variance operators in equation (5) are applied elementwise.

Let us provide some intuition as to why $\nu(\pi; \mathbf{P}, r, \gamma)$ plays a fundamental role. The appearance of the zero mean term $\widehat{\mathbf{T}}(Q^\star) - \mathbf{T}(Q^\star)$ is natural: it reflects the noise present in the empirical Bellman operator (4) as an estimate of the population Bellman operator (2). As for the pre-factor $(\mathbf{I} - \gamma\mathbf{P}^\pi)^{-1}$, the Neumann series expansion allows us to write

$$(\mathbf{I} - \gamma\mathbf{P}^\pi)^{-1} = \sum_{k=0}^{\infty} (\gamma\mathbf{P}^\pi)^k.$$

The sum of the powers of $\gamma\mathbf{P}^\pi$ account for the compounded effect of an initial perturbation when following the Markov chain specified by the policy π.

b) Upper and lower bounds: With these definitions in place, the core of our work involves proving—via a combination of a lower and an upper bound matching up to logarithmic factors—that the instance-specific difficulty of estimating the Q-function is captured by the quantity $\max_{\pi\in\mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r, \gamma)\|_\infty$. Here $\mathbf{\Pi}^\star$ denotes the set of all optimal policies for the MDP instance $(\mathbf{P}, r)$. This functional exhibits a wide range of behaviors: in Example 1 to follow in Section II-A2, we exhibit a very simple family of MDPs $(\mathbf{P}_\lambda, r_\lambda)$, parameterized by a scalar $\lambda \geqslant 0$, such that

$$\max_{\pi\in\mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}_\lambda, r_\lambda, \gamma)\|_\infty \asymp \Big(\frac{1}{1-\gamma}\Big)^{\frac{3}{2}-\lambda}. \tag{7}$$

The setting $\lambda = 0$ recovers a "hard" instance, one for which the global minimax bound for estimation of Q-functions, known from past work [AMK13] on batched Q-learning, is sharp. For these hard instances, the sample complexity[1] grows cubically as H^3 in the effective horizon $H = 1/(1-\gamma)$. On the other hand, as λ grows, the problems in this family become progressively easier, so that the global minimax bound is no longer sharp.

In more detail, we prove a non-asymptotic lower bound, stated as Theorem 1 to follow, by adapting a particular definition of local minimax risk studied in past work on shape-constrained estimation [CL15]. The central challenge in this proof is that perturbations to the transition matrices of a given MDP change not only the transitions themselves, but also the structure of the optimal policies. Asymptotic versions of local minimax rates in statistics dates back to the seminal works of Le Cam and Hayek (see chapter 8 in the book [Vaa98]). The paper [DR21] studies such guarantees for optimization. For reinforcement learning settings, non-asymptotic guarantees for policy evaluation have been established in the paper [KPR+21]; see that paper for further motivation and details.

In order to prove matching upper bounds, given the role of the empirical operators $\widehat{\mathbf{T}}$ in our lower bound—used in the classical Q-learning algorithm [WD92, Tsi94, Sze97, JJS94]—a natural thought would be to analyze this operator directly. However, a line of past work [Wai19b, LCC+24] has revealed the interesting fact that classical Q-learning algorithm—despite its widespread use—is *actually sub-optimal*, even when assessed when using the coarser metric of global minimax. In particular, Wainwright [Wai19b] provided numerical evidence of problems for which standard Q-learning has ℓ_∞-norm sample complexity growing at least as fast as H^4 in the effective horizon $H = 1/(1-\gamma)$. This should be contrasted with the global minimax theory [AMK13], for which (as discussed following equation (7)) the optimal scaling is H^3. Subsequent work by Li et al. [LCC+24] proved that the H^4-scaling is actually unavoidable for standard Q-learning, so that it is a sub-optimal procedure even in a global minimax sense.

Thus, in order to obtain a sharp upper bound, we need to analyze a more sophisticated procedure. In particular, we turn to the variance-reduced forms of Q-learning, as introduced in past work [SWW+18, SWWY18, Wai19c] and shown to be optimal in a globally minimax sense. Our main contribution is to show that under certain structural conditions and lower bounds on the sample size, there is a form of variance-reduced Q-learning that achieves our *instance-dependent* lower bound up to a logarithmic factor. These upper bounds, stated precisely in Theorem 2, confirm that our lower bound technique has extracted a useful form of instance dependence for estimating optimal Q-functions.

c) Other connections to the literature: The generative setting studied in this paper was first introduced and analyzed in the paper [KS02]. It has been the subject of much prior work in model-based reinforcement learning. For instance, in the generative setting, a plug-in approach is known to be optimal for estimating the Q-function [AKY20, LWC+20], where here optimality is measured in the global minimax sense. In contrast, our results apply to the model-free setting, and our

[1]The sample complexity $N(\epsilon, \gamma)$ refers to the number of samples required to achieve an ϵ-accurate solution in ℓ_∞-norm; we obtain the cubic H^3 scaling—instead of $H^{3/2}$ as appears in the bound (7) with $\lambda = 0$—since the sample size depends scales as $\sqrt{N}$.

guarantees—rather than being global over a large model class—are specific to the particular MDP instance that is given. As we illustrate in the sequel, the instance-dependent functional identified by our theory exhibits substantial variation across the space of possible MDPs. Of course, its worst-case behavior matches the global minimax theory, but there are many problems that are substantially easier, and an algorithm that adapts to problem structure will behave very differently than the global minimax prediction. We also note that there also is a vast literature on Q-learning in more general settings, with some examples including DQN [FWXY20], regret minimization in tabular setting [JAZBJ18], and regret minimization in linear MDPs [JYWJ20].

d) Notation: For a positive integer n, we use the shorthand $[n] := \{1, 2, \ldots n\}$. For a finite set S, we use $|S|$ to denote its cardinality. We use $c_1, c_2, \ldots$ to denote universal constants that may change from line to line. For any pair of vectors or matrices (v, w) with matching dimension(s), we write that $v \succeq w$ to imply $v - w$ has only non-negative entries, and $v \preceq w$ is defined similarly. We let $|u|$ denote the entrywise absolute value of a vector $u \in \mathbb{R}^n$ or a matrix $u \in \mathbb{R}^{m \times n}$; we use $|u|_+$ to denote the entry-wise positive part of u. For any vector or matrix u, we let $\|u\|_\infty$ denote the maximum absolute value taken over all entries of u, and $\|u\|_{\text{span}} = \max_j u_j - \min_j u_j$ denote the span seminorm. For a continuous operator $P : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$, its ℓ_∞-operator norm is given by $|\!|\!| P |\!|\!|_{\infty \to \infty} = \sup_{\|u\|_\infty = 1} \|Pu\|_\infty$. We often identify a Q-value function with its matrix representation and use $\|Q\|_\infty$ to denote the infinity norm (i.e., largest entry in absolute terms). In the matrix representation of Q, its rows and columns are indexed via an enumeration of the states and actions, respectively. We use the symbol $\gtrsim$ to denote a relation that holds up to logarithmic factors in the problem parameters.

II. Main results

In this section, we provide precise statements of the main results of this paper, along with a discussion of some of their consequences. In Section II-A, we define a notion of local non-asymptotic minimax risk, and then state Theorem 1, which provides such a lower bound for estimating optimal Q-value functions. In Section II-B, we turn to the complementary problem of deriving achievable results. Theorem 2 shows that under certain structural conditions on the policies, there is a form of variance-reduced Q-learning that achieves the local minimax risk up to logarithmic factors.

A. Instance-dependent lower bounds

In this section, we state a non-asymptotic lower bound for estimating optimal Q-function in the ℓ_∞-norm. This lower bound, to be stated in Theorem 1, is instance-dependent, meaning that it depends on the particular instance of the MDP $(\mathbf{P}, r)$ at hand. This dependence should be contrasted with classical global minimax bounds, which are oblivious to such local properties.

The starting point of our lower bound development is the two-point framework introduced by Cai and Low [CL15] for local minimax bounds for nonparametric shape-constrained inference; here we adapt it to our current setting. Focusing on the ℓ_∞-norm error metric, the *local non-asymptotic minimax risk* for estimating the value function $Q(\mathcal{P})$ associated with an instance $\mathcal{P} = (\mathbf{P}, r)$ is defined as

$$\mathfrak{M}_N(\mathcal{P}) := \sup_{\mathcal{P}'} \inf_{\hat{Q}_N} \max_{\mathcal{I} \in \{\mathcal{P}, \mathcal{P}'\}} \sqrt{N}\, \mathbb{E}_{\mathcal{I}}\big[\|\hat{Q}_N - Q(\mathcal{I})\|_\infty\big]. \tag{8}$$

Here the infimum is taken over all estimators $\hat{Q}_N$ that are measurable functions of the N i.i.d. observations drawn according to our observation model (see Section I-A).

The intuition underlying the definition (8) is that given an instance $\mathcal{P}$, the adversary behaves as follows: it extracts the hardest alternative $\mathcal{P}'$ relative to $\mathcal{P}$, and then measures the worst-case risk over $\mathcal{P}$ and this alternative $\mathcal{P}'$.

1) Lower bounds for Q-function estimation: We now turn to the statement of some lower bounds for estimating the optimal Q-function. Recall the definition (6) of the operator $\mathbf{P}^\pi$, along with the functional $\nu(\pi; \mathbf{P}, r, \gamma)$ from equation (5). We let $\nu^2(\pi; \mathbf{P}, r, \gamma)$ denote the matrix obtained by taking squares entrywise. Our first step is to provide a decomposition of this matrix into two separate components, corresponding to the noisiness in the reward function observation and transition matrix observations, respectively.

In order to deal with the latter source of noise, with a slight abuse of notation, we use the observed matrix $\mathbf{Z}$ to define a stochastic analog of $\mathbf{P}^\pi$—namely, the (random) right-linear operator

$$\begin{aligned} (\mathbf{Z}^\pi Q)(s, a) &:= \sum_{s' \in \mathcal{S}} \mathbf{Z}_a(s' \mid s) \cdot Q(s', \pi(s')), \\ \text{where} \quad \mathbf{Z}_a(s' \mid s) &:= \mathbb{1}_{\mathbf{Z}(s,a) = s'}. \end{aligned} \tag{9}$$

By assumption, the randomness in our observations of the reward and transitions are independent, so that for

any optimal[2] policy π, we have the decomposition

$$\nu^2(\pi; \mathbf{P}, r)(s, a) = \gamma^2 \rho^2(\pi; \mathbf{P}, r)(s, a) + \sigma^2(\pi; \mathbf{P}, r)(s, a). \tag{10a}$$

Here we define

$$\rho^2(\pi; \mathbf{P}, r) := \mathrm{Var}\left((\mathbf{I} - \gamma \mathbf{P}^\pi)^{-1}(\mathbf{Z}^\pi - \mathbf{P}^\pi)Q^\star\right), \tag{10b}$$

$$\sigma^2(\pi; \mathbf{P}, r) := \mathrm{Var}\left((\mathbf{I} - \gamma \mathbf{P}^\pi)^{-1}(R - r)\right), \tag{10c}$$

where we compute the variances in an elementwise sense.

With this notation, we have the following guarantee:

Theorem 1. *There exists a universal constant $c > 0$ such that for any instance $\mathcal{P} = (\mathbf{P}, r)$, the local non-asymptotic minimax risk is lower bounded as*

$$\mathfrak{M}_N(\mathcal{P}) \geqslant c \max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r, \gamma)\|_\infty. \tag{11a}$$

This bound is valid for all sample sizes $N \geqslant N_0$, where

$$N_0 := \max\left\{ \frac{2\gamma^2}{(1-\gamma)^2}, \frac{2\|Q^\star\|_{\mathrm{span}}^2}{(1-\gamma)^2 \|\rho^2(\pi^\star; \mathbf{P}, r)\|_\infty} \right\}, \tag{11b}$$

and $\pi^\star \in \arg\max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r)\|_\infty$.

See Section III for the proof of this claim. The main take-away is that the functional $\max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r, \gamma)\|_\infty$ controls the local minimax risk. In order to gain intuition for this claim, it is worth exploring the range of possible behaviors exhibited by the functional appearing in the lower bound (11a).

2) Exploring the range of possible behaviors: A first point of comparison is between the instance-dependent lower bound from Theorem 2 with the existing minimax lower bounds for Q-learning. Azar et al. [AMK13] proved a global minimax lower bound on the ℓ_∞-norm error for estimating the optimal Q-function. They exhibited a family of γ-discounted MDPs for which the ℓ_∞-error of any procedure is lower bounded by the quantity $\frac{1}{(1-\gamma)^{3/2}} \cdot \frac{1}{\sqrt{N}}$, up to logarithmic factors in dimension. In terms of the number of samples $N(\epsilon)$ required to achieve an ϵ-accurate solution in the ℓ_∞-norm, this worst-case result scales as H^3 in the effective horizon $H = 1/(1-\gamma)$.

This lower bound is optimal in a globally minimax sense, and it is worthwhile understanding the properties of instances that exhibit this worst-case behavior: concretely, for such worst-case problems, we must have a scaling of the form

$$\max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r, \gamma)\|_\infty \asymp \tfrac{1}{(1-\gamma)^{1.5}}.$$

On the flipside, it also worthwhile understanding the properties of problems that are much "easier" than this worst-case theory would suggest.

The following construction, which takes inspiration from the papers [PW20, KPR+21], allows us to explore this continuum in an illuminating fashion:

Example 1 (A continuum of local minimax risks)**.** Consider an MDP with two states $\{s_1, s_2\}$, two actions $\{a_1, a_2\}$, and with transition functions and reward functions given by

$$\mathbf{P}_{a_1} = \begin{bmatrix} p & 1-p \\ 0 & 1 \end{bmatrix}, \quad \mathbf{P}_{a_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad r = \begin{bmatrix} 1 & 0 \\ \tau & 0 \end{bmatrix}. \tag{12}$$

We assume that there is no randomness in the rewards. Here the pair (p, τ) along with the discount factor γ are parameters of the construction, and we consider a subfamily of these parameters indexed by a non-negative scalar λ. For any $\lambda \geqslant 0$ and discount factor $\gamma \in (\frac{1}{4}, 1)$, consider the settings

$$p = \frac{4\gamma - 1}{3\gamma}, \quad \text{and} \quad \tau = 1 - (1-\gamma)^\lambda.$$

With these choices, the optimal state-action-value function $Q^\star$ takes the form

$$Q^\star = \begin{bmatrix} \frac{1}{4} \cdot \frac{3+\tau}{1-\gamma} & \frac{\gamma}{4} \cdot \frac{3+\tau}{1-\gamma} \\ \frac{\tau}{1-\gamma} & \frac{\gamma\tau}{1-\gamma} \end{bmatrix},$$

with an unique optimal policy $\pi^\star(s_1) = \pi^\star(s_2) = a_1$. We can then compute that

$$\max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}_\lambda, r_\lambda)\|_\infty = c \cdot \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}. \tag{13}$$

See Appendix A for the details of this calculation.

Substituting the expression (13) into equation (11a) yields that the local minimax risk is lower bounded as $\mathfrak{M}_N(\mathcal{P}) \geqslant c \frac{1}{(1-\gamma)^{1.5-\lambda}}$. Consequently, for $\lambda > 0$, our lower bounds suggest it should be possible to estimate the optimal Q-function more accurately by a factor $(1-\gamma)^\lambda$. To be clear, this is *not merely* a constant factor difference in the ratio of "easy" to "hard"; it is a function depending on the discount factor γ, and it diverges to infinity as $\gamma \to 1$ from below.

[2]Optimality of π is required so that $\mathbf{T}(Q^\star) = r + \gamma \mathbf{P}^\pi Q^\star$, with a similar relation for the empirical Bellman operator.

B. Instance-dependent upper bounds

Thus far, we have stated some instance-dependent lower bounds on the sample complexity of estimating Q-value functions. As we saw in the preceding Example 1, these lower bounds exhibit a wide range of behavior depending on the structure of the transition functions, discount parameter and reward functions. However, these differences in the lower bounds are only interesting if we can show that they are optimal, meaning that there is a (hopefully practical) algorithm that matches the behavior predicted by the lower bounds.

In this section we close this gap, in particular via a careful analysis of variance-reduced Q-learning (or VR-QL for short). Variance-reduced forms of Q-learning have been proposed and shown to be globally minimax in previous work [SWW+18, SWWY18, Wai19c]; the version analyzed here is adapted from the paper [Wai19c]. In Theorem 2, we show that the VR-QL algorithm is instance-optimal up to logarithmic factors under two different sets of assumptions.

1) From standard to variance-reduced Q-learning: The classical Q-learning algorithm is a stochastic approximation algorithm for estimating the unique fixed point $Q^\star$ of the Bellman operator $\mathbf{T}$. Recall the definition (4) of the empirical Bellman operator $\widehat{\mathbf{T}}_k$. At each iteration $k = 1, 2, \ldots$, standard Q-learning performs an update of the form

$$Q_{k+1} = (1 - \alpha_k)Q_k + \alpha_k \widehat{\mathbf{T}}_k(Q_k), \tag{14}$$

where $\alpha_k \in (0, 1)$ is a stepsize parameter. Appropriately decaying choices of the stepsize ensure that the estimate Q_k converges to $Q^\star$. Unfortunately, the convergence rate is known to be non-optimal, failing to achieve the global minimax rate [Wai19b, LCC+24], let alone the finer-grained instance-dependent requirements in this paper. This non-optimality has to do with the rate at which variance accumulates as the procedure is run.

Variance reduction is a general principle that can be applied to stochastic approximation schemes so as to accelerate their convergence. Here we describe the variance-reduced version of Q-learning that we analyze here. Similar to standard variance-reduced schemes for stochastic optimization [see, e.g., JZ13], the algorithm consists of a sequence of epochs. Within each epoch, we run a re-centered version of the QL update. The re-centering is done in such a way, using a Monte Carlo approximation of the population Bellman operator $\mathbf{T}$, so that the re-centered updates have lower variance. We leave the details of the epochs and Monte Carlo to Section II-B4; here let us describe the basic form of the updates within a given epoch.

Suppose that we run the algorithm using a total of M epochs. At epoch m, the algorithm uses a re-centering point $\overline{Q}_m$ in order to re-center the update, where $\overline{Q}_m$ acts as the current best estimate of $Q^\star$. Ideally, we should re-center the operator $\widehat{\mathbf{T}}_k$ using the quantity $\mathbf{T}(\overline{Q}_m)$, but we lack the access to it; instead, we use the Monte Carlo approximation

$$\overline{\mathbf{T}}_{N_m}(\overline{Q}_m) := \frac{1}{N_m} \sum_{i \in \mathcal{D}_m} \widehat{\mathbf{T}}_i(\overline{Q}_m). \tag{15}$$

Given the pair $(\overline{Q}_m, \overline{\mathbf{T}}_{N_m}(\overline{Q}))$ and a stepsize parameter $\alpha \in (0, 1)$, we define the *variance-reduced Q-learning update* $Q \mapsto \mathcal{V}_k\left(Q; \alpha, \overline{Q}_m, \overline{\mathbf{T}}_{N_m}\right)$, where

$$\mathcal{V}_k\left(Q; \alpha, \overline{Q}_m, \overline{\mathbf{T}}_{N_m}\right) := (1 - \alpha)Q + \alpha \widehat{\mathbf{S}}_k(Q, \overline{Q}_m) \tag{16}$$

with

$$\widehat{\mathbf{S}}_k(Q, \overline{Q}_m) := \left\{ \widehat{\mathbf{T}}_k(Q) - \widehat{\mathbf{T}}_k(\overline{Q}_m) + \overline{\mathbf{T}}_{N_m}(\overline{Q}_m) \right\}.$$

The operator $\widehat{\mathbf{T}}_k$ is independent of the set of operators $\{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{D}_m}$, used to compute the Monte Carlo approximation $\overline{\mathbf{T}}_{N_m}$. As a result, the stochastic operator $\widehat{\mathbf{T}}_k$ is independent of the re-centering quantity $\overline{\mathbf{T}}_{N_m}(\overline{Q}_m)$. See Section II-B4 for the details on how the epoch lengths and re-centering sample sizes $\mathcal{D}_m$ are chosen.

2) Non-asymptotic guarantees for variance-reduced Q-learning: In this section, we state some non-asymptotic guarantees for the VR-QL algorithm. We provide guarantees under a condition which involves the structure of the set of optimal policies. We begin by introducing some definitions that underlie this condition.

Given an MDP instance $(\mathbf{P}, r)$, we define the *optimality gap*

$$\Delta := \min_{\pi \in \mathbf{\Pi} \setminus \mathbf{\Pi}^\star} \|Q^\star - \{r + \gamma \mathbf{P}^\pi Q^\star\}\|_\infty, \tag{17}$$

where $Q^\star, \mathbf{\Pi}^\star$, and $\mathbf{\Pi}$, respectively, denote the optimal Q-function, the set of optimal policies, and the set of all policies for the MDP $(\mathbf{P}, r)$. Observe that the scalar Δ captures the difficulty in detecting the set of optimal policies. In other words, when Δ is small, it is hard to distinguish an optimal policy from a suboptimal policy.

For any Q-value function Q, we say that a policy π is greedy with respect to Q if $\pi(s) \in \arg\max_{a \in \mathcal{A}} Q(s, a)$ for all $s \in \mathcal{S}$. Note that any policy $\pi^\star$ that is greedy with respect to the optimal Q-value function $Q^\star$ is an optimal policy.

With these definitions in place, we place the following lower bound on the sample size: there is some $\beta > 0$ such that

$$\frac{N}{(\log N)^2} \geqslant c_2 \log(D/\delta) \cdot \frac{(1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^3} \cdot \max\{1, \frac{1}{\Delta^2(1-\gamma)^\beta}\}. \tag{18}$$

We assume that we are given an initial point $\bar{Q}_1$ such that

$$\|\bar{Q}_1 - Q^\star\|_\infty \leqslant \frac{\|r\|_\infty}{\sqrt{1-\gamma}}. \tag{19}$$

Such an initial condition has already been used in the literature [Wai19c], and it can be ensured by first running Algorithm VR-QL for a total of $\frac{1}{(1-\gamma)^3}$ samples (up to logarithmic factor corrections).

Theorem 2. *There is a choice of epoch parameters such that given any discount parameter $\gamma \in [\frac{1}{2}, 1)$ and an initial point $\bar{Q}_1$ satisfying the sample size requirement (18) and initialization condition (19), Algorithm VR-QL run for $M := \log_4\left(\frac{N(1-\gamma)^2}{8\log((16D/\delta)\cdot\log N)}\right)$ epochs yields an estimate $\bar{Q}_{M+1}$ such that*

$$\|\bar{Q}_{M+1} - Q^\star\|_\infty \leqslant c_0 \cdot \sqrt{\frac{\log_4(8DM|\mathbf{\Pi}^\star|/\delta)}{N}} \cdot \max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma)\|_\infty + c_1 \cdot \frac{\log_4(8DM|\mathbf{\Pi}^\star|/\delta)}{N} \cdot \frac{\|Q^\star\|_{\text{span}}}{1-\gamma}, \tag{20}$$

with probability exceeding $1 - \delta$.

See Section IV for the proof of this claim.

a) Comparing the upper and lower bounds: When both the lower bounds on the sample sizes from Theorems 1 and 2 hold, we can see that the guarantees from both theorems match up to logarithmic factors (and higher-order terms), and consequently, the VR-QL algorithm is *instance optimal*. However the guarantee in Theorem 2 holds under a more stringent condition on N which depends on the optimality gap Δ, as compared to the relatively mild sample size condition of Theorem 1. Closing this gap between the differences in restrictions on N is a worthwhile goal for future work. We conjecture that the sample size dependence on Δ in Theorem 2 can be removed.

b) Usefulness of the bounds: There exists a separate line of work [XKWJ23] that is focused on estimating functionals of the form similar to $\max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma)\|_\infty$ and using estimates for inferential purposes and early stopping. The authors exhibit substantial savings in the number of samples required to reach a target threshold by taking advantage of the fact that $\max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma)\|_\infty$ can be much smaller than the global minimax rate.

3) Confirming the theoretical predictions: Some numerical experiments are helpful in order to illustrate instance-adaptive behavior guaranteed by Theorem 2. Recall the family of MDPs (12) from Example 1. Suppose that we set $\lambda = 0.5$ and for each choice of $\gamma \in (1/2, 1)$, we collect $N = \lceil \frac{16 \cdot 32}{9} \frac{1}{(1-\gamma)^3} \rceil$ samples, and then run the VR-QL algorithm over a range of discount parameters γ, using the settings from Theorem 2 and Section II-B4, thereby obtaining an estimate $\bar{Q}_{M+1}$.

Figure 1(a) plots the evolution of log ℓ_∞-norm error of the estimate over time as the algorithm proceeds; the form of these curves show the epoch-based nature of the convergence. See Section II-B4 for more details on the parameters of the epochs, including the base parameter illustrated here. Plotted as blue circles in panel (b) of Figure 1 are the logarithm of the ℓ_∞-norm error of the final output; that is, $\log \|\bar{Q}_{M+1} - Q^\star\|_\infty$ versus $\log(H)$ of the effective horizon $H = 1/(1-\gamma)$. Each point in this plot represents an average over 1000 trials.

In terms of theory, with the settings given above, existing worst-case bounds [AMK13, Wai19c] predict that the log ℓ_∞-norm error remains constant as the log discount complexity grows; accordingly, we have plotted a dotted red line with slope zero to illustrate the worst-case guarantee. On the other hand, for the MDP instance (12) with $\lambda = 0.5$, a simple calculation yields that for the instance (12) the suboptimality gap Δ satisfies $\Delta = 1 - \frac{(1-\gamma)^\lambda}{4} \geqslant \frac{3}{4}$. In our experiment, we set the sample size to be $N = \lceil \frac{32}{(1-\gamma)^3} \cdot \frac{4^2}{3^2} \rceil \geqslant \frac{32}{(1-\gamma)^3 \cdot \Delta^2}$; as a result, the bounds from Theorems 1 and 2 are valid.

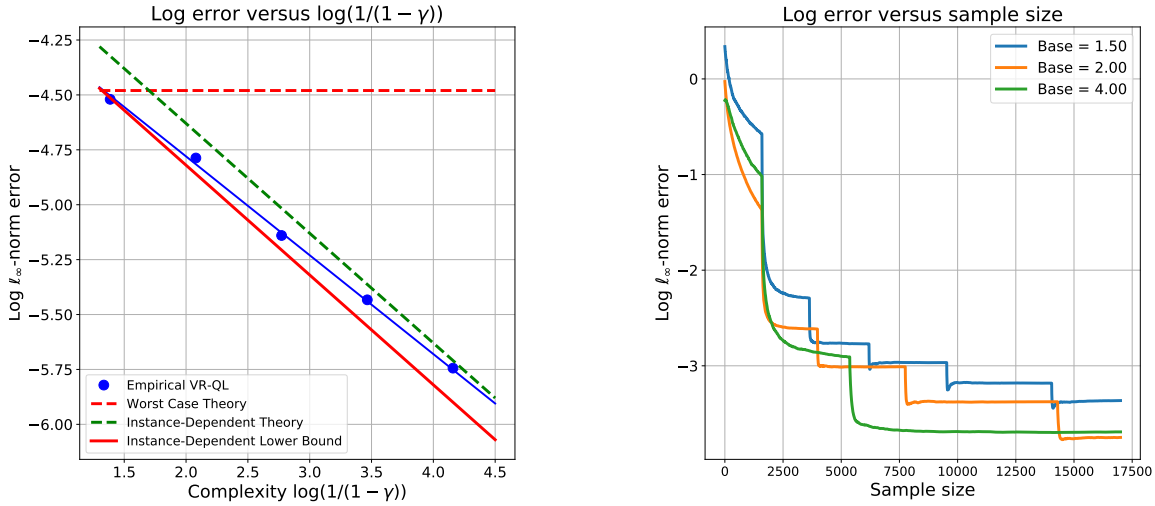
With the setting $\lambda = 0.5$, our calculations from Example 1 yield

$$\max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma)\|_\infty \asymp \left(\frac{1}{1-\gamma}\right)^{-0.5}.$$

Thus, with the choice of sample size N given above, our theory predicts that the log ℓ_∞-norm error should exhibit the scaling

$$\begin{aligned} &\log \|\bar{Q}_{M+1} - Q^\star\|_\infty \\ &\quad \asymp \log\left(\frac{1}{\sqrt{N}} \max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma)\|_\infty\right) \\ &\quad \asymp -0.5 \log\left(\frac{1}{1-\gamma}\right) + c, \end{aligned}$$

where c is a constant. In Figure 1(b), we plot the lower bound from Theorem 1 as a solid red line, and the upper



(a) (b)

Fig. 1. (a) $\lambda = 0.5, \quad N = \lceil \frac{32}{(1-\gamma)^3} \cdot \frac{4^2}{3^2} \rceil$ $\gamma = 0.9$. Illustration of the qualitative behavior of Algorithm VR-QL applied on the MDP (1) along with instance dependent and the worst case bounds. The figure plots the log ℓ_∞-error $\|Q_{M+1} - Q^\star\|_\infty$ against the log discount complexity factor $\log(\frac{1}{1-\gamma})$ with $\lambda = 0.5$. We have also plotted the least-squares fit through these points, and the instance-dependent lower bound from Theorem 1, the instance-dependent upper bound from Theorem 2, and the worst-case bound [Wai19c]. (b) Behavior of the VR-QL algorithm with different choices of the base b. The plot demonstrates that different choices of the base b yield similar behavior.

bound from Theorem 2 as a dashed green line. (While these lines both have slope -0.50, the intercept term c is different due to the additional logarithmic factors in dimension present in the upper bound.)

In order to test how the empirical behavior conforms to the theoretical prediction, we did an ordinary least-squares fit of the log ℓ_∞-norm error versus the log discount complexity; this fit yields a line with slope $\widehat{\beta} = -0.45$, and is plotted in solid blue. This test shows good agreement between the theoretical prediction and the practical behavior.

4) Details of the epochs and procedure: In this section, we provide the complete details of the algorithm used in our version of variance-reduced Q-learning.

a) A single epoch: A single epoch of the overall variance-reduced QL algorithm involves repeated applications of the basic variance-reduced update $\mathcal{V}_k$ from equation (16). The epochs are indexed with integers $m = 1, 2, \ldots, M$, where M corresponds to the total number of epochs to be run. Each epoch m requires the following four inputs:

- an element $\overline{Q}$, which is chosen to be the output of the previous epoch $m-1$;
- a positive integer K denoting the number of steps within the given epoch;
- a positive integer N_m denoting the batch size used to calculate the Monte Carlo update (15);
- a set of fresh operators $\{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{C}_m}$, with $|\mathcal{C}_m| = N_m + K$. The set $\mathcal{C}_m$ is partitioned into two subsets having sizes N_m and K, respectively. The first subset, of size N_m, which we call $\mathcal{D}_m$, is used to construct the Monte Carlo approximation (15). The second subset, of size K is used to run the K steps within the epoch.

We summarize a single epoch in pseudocode form in Algorithm SingleEpoch.

b) Overall algorithm: The overall algorithm, denoted by VR-QL for short, has five inputs: (a) an initialization $\bar{Q}_1$, (b) an integer M, denoting the number of epochs to be run, (c) an integer K, denoting the length of each epoch, (d) a sequence of batch sizes $\{N_m\}_{m=1}^K$, denoting the number of operators used for re-centering in the M epochs, and (e) sample batches $\{\{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{C}_m}\}_{m=1}^M$ to be used in the M epochs. Given these five inputs, the overall procedure can be summarized as in Algorithm VR-QL.

c) Settings for Theorem 2: Given a tolerance probability $\delta \in (0,1)$ and the number of available i.i.d. samples N, we run Algorithm VR-QL with a total of $M := \log_4 \left(\frac{N(1-\gamma)^2}{8 \log((16D/\delta) \cdot \log N)} \right)$ epochs, along with the

Algorithm **SingleEpoch** RunEpoch $(\overline{Q}; K, N_m, \{\widehat{\mathbf{T}}_i\}_{i\in\mathcal{C}_m})$

1: Given (a) Epoch length K, (b) Re-centering vector $\overline{Q}$, (c) Re-centering batch size N_m, (d) Operators $\{\widehat{\mathbf{T}}_i\}_{i\in\mathcal{C}_m}$
2: Compute the re-centering quantity

$$\overline{\mathbf{T}}_{N_m}(\overline{Q}) := \frac{1}{N_m}\sum_{i\in\mathcal{D}_m}\widehat{\mathbf{T}}_i(\overline{Q})$$

3: Initialize $Q_1 = \overline{Q}$
4: **for** $k = 1, 2, \ldots, K$ **do**
5: Compute the variance-reduced update:

$$Q_{k+1} = \mathcal{V}_k(Q_k; \alpha_k, \overline{Q}, \overline{\mathbf{T}}_{N_m}) \quad \text{with stepsize } \alpha_k = \frac{1}{1+(1-\gamma)k}.$$

6: **end for**
7: **return** Q_{K+1}

Algorithm VR-QL

1: Given (a) Initialization $\overline{Q}_1$, (b) Number of epochs, M, (c) Epoch length K, (d) Re-centering sample sizes $\{N_m\}_{m=1}^M$, (e) Sample batches $\{\widehat{\mathbf{T}}_i\}_{i\in\mathcal{C}_m}$ for $m = 1, \ldots, M$
2: Initialize at $\overline{Q}_1$
3: **for** $m = 1, 2, \ldots, M$ **do**
4: $\overline{Q}_{m+1} = \text{RunEpoch}(\overline{Q}_m; K, N_m, \{\widehat{\mathbf{T}}_i\}_{i\in\mathcal{C}_m})$
5: **end for**
6: **return** $\overline{Q}_{M+1}$ as final estimate

following parameter choices:

`Re-centering sizes:`

$$N_m = c_1 \frac{4^m}{(1-\gamma)^2}\cdot \log_4(16MD/\delta) \tag{21a}$$

`Sample batches:`

Partition the N samples to obtain $\{\widehat{\mathbf{T}}_i\}_{i\in\mathcal{C}_m}$ for $m = 1, \ldots, M$ (21b)

`Epoch length:`

$$K = \frac{N}{2M}. \tag{21c}$$

III. Proof of Theorem 1

Given an MDP instance $\mathcal{P} = (\mathbf{P}, r)$, we start by introducing the following two classes of alternative MDPs:

$$\mathcal{S}_1 = \{\mathcal{P}' = (\mathbf{P}', r') \mid r' = r\}, \quad \text{and} \quad \mathcal{S}_2 = \{\mathcal{P}' = (\mathbf{P}', r') \mid \mathbf{P}' = \mathbf{P}\}. \tag{22}$$

We consider the restricted version of the local minimax risk at the instance $\mathcal{P}'$ to the classes $\mathcal{S}_i$:

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_i) = \sup_{\mathcal{P}'\in\mathcal{S}_i}\inf_{\hat{Q}_N}\max_{\mathcal{I}\in\{\mathcal{P},\mathcal{P}'\}}\sqrt{N}\cdot L(\hat{Q}_N, \mathcal{I}), \tag{23}$$

where we have defined

$$L(\hat{Q}_N, \mathcal{I}) := \mathbb{E}_{\mathcal{I}}\|\hat{Q}_N - Q(\mathcal{I})\|_\infty.$$

The main part of the proof involves showing that there exists a universal constant $c > 0$ such that

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_1) \geqslant c\cdot\max_{\pi\in\mathbf{\Pi}^\star}\|\gamma\rho(\pi; \mathbf{P}, r)\|_\infty, \quad \text{and} \tag{24a}$$

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_2) \geqslant c\cdot\max_{\pi\in\mathbf{\Pi}^\star}\|\sigma(\pi; \mathbf{P}, r)\|_\infty, \tag{24b}$$

where $\mathbf{\Pi}^\star$ denotes the optimal policy set for $(\mathbf{P}, r)$. We can then conclude

$$\begin{aligned}\mathfrak{M}_N(\mathcal{P}) &\geqslant \max\{\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_1), \mathfrak{M}_N(\mathcal{P}; \mathcal{S}_2)\}\\ &\geqslant \frac{1}{2}\left(\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_1) + \mathfrak{M}_N(\mathcal{P}; \mathcal{S}_2)\right)\\ &\geqslant \frac{c}{2}\max_{\pi\in\mathbf{\Pi}^\star}\|\gamma\rho(\pi; \mathbf{P}, r)\|_\infty\\ &\qquad + \frac{c}{2}\max_{\pi\in\mathbf{\Pi}^\star}\|\sigma(\pi; \mathbf{P}, r)\|_\infty\\ &\geqslant \frac{c}{2}\max_{\pi\in\mathbf{\Pi}^\star}\|\nu(\pi; \mathbf{P}, r)\|_\infty.\end{aligned}$$

The last inequality above follows from the decomposition (10a). It remains to prove the claims (24a) and (24b). More precisely, the core of our proof involves proving the following two lemmas:

Lemma 1. *For all $\mathcal{S} \in \{\mathcal{S}_1, \mathcal{S}_2\}$, we have that $\mathfrak{M}_N(\mathcal{P}; \mathcal{S}) \geqslant \frac{1}{8}\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S})$ where we define*

$$\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S}) := \sup_{\mathcal{P}'\in\mathcal{S}}\left\{\sqrt{N}\cdot\|Q(\mathcal{P}) - Q(\mathcal{P}')\|_\infty \;\middle|\; d_{Hel}(\mathcal{P}, \mathcal{P}') \leqslant \frac{1}{2\sqrt{N}}\right\}.$$

This lemma follows as a fairly straightforward consequence of the standard reduction from estimation to testing; see Appendix B-A for the details.

Our next lemma requires more effort to prove, and leverages the specific structure of the problem at hand:

Lemma 2. *Given any MDP instance $\mathcal{P} = (\mathbf{P}, r)$:*

(a) There exists an instance $\mathcal{P}_1 = (\mathbf{P}', r) \in \mathcal{S}_1$ such that $d_{Hel}(\mathcal{P}, \mathcal{P}_1) \leqslant \frac{1}{2\sqrt{N}}$ and

$$\sqrt{N}\cdot\|Q(\mathcal{P}) - Q(\mathcal{P}_1)\|_\infty \geqslant c\cdot\max_{\pi\in\mathbf{\Pi}^\star}\|\gamma\rho(\pi; \mathbf{P}, r)\|_\infty.$$

(b) There exists an instance $\mathcal{P}_2 = (\mathbf{P}, r') \in \mathcal{S}_2$ such that $d_{Hel}(\mathcal{P}, \mathcal{P}_2) \leqslant \frac{1}{2\sqrt{N}}$ and

$$\sqrt{N}\cdot\|Q(\mathcal{P}) - Q(\mathcal{P}_2)\|_\infty \geqslant c\cdot\max_{\pi\in\mathbf{\Pi}^\star}\|\sigma(\pi; \mathbf{P}, r)\|_\infty.$$

Note that the bounds (24a)–(24b) stated in Theorem 1 follow by combining the claims of Lemmas 1 and 2. The remainder of our proof focuses on establishing Lemma 2.

A. Proof of Lemma 2

In this section, we prove the two parts of Lemma 2.

1) Proof of Lemma 2(a): Throughout the proof, we use z to denote a generic element of the state-action set $\mathcal{S} \times \mathcal{A}$. Let Q be the true Q-function for the MDP $\mathcal{P} = (\mathbf{P}, r)$. We adopt the shorthands

$$\pi_1 \in \arg\max_{\pi \in \mathbf{\Pi}^\star} \|\rho(\pi; \mathbf{P}, r)\|_\infty, \tag{25a}$$

$$\bar{z} \in \arg\max_{z \in \mathcal{S} \times \mathcal{A}} \rho(\pi_1; \mathbf{P}, r), \tag{25b}$$

$$\tilde{\rho}(z) := \rho(\pi_1; \mathbf{P}, r)(z), \tag{25c}$$

$$\mathbf{U} := (\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1}, \quad \text{and} \tag{25d}$$

$$\varphi^2(z) := \mathrm{Var}\left(\mathbf{Z}^{\pi_1} Q(z)\right). \tag{25e}$$

To explain this notation, we choose π_1 to be the optimal policy that achieves the largest ℓ_∞-norm across $\rho(\pi^*; \mathbf{P}, r)$ for optimal policies π^*, we let $\bar{z}$ is the state-action pair index that achieves the maximal entry of $\rho(\pi_1; \mathbf{P}, r)$, and we use $\tilde{\rho}$ as convenient shorthand to refer to the values of $\rho(\pi_1; \mathbf{P}, r)$. This choice of notation implies that

$$\tilde{\rho}(\bar{z}) = \max_{\pi \in \mathbf{\Pi}^\star} \|\rho(\pi; \mathbf{P}, r)\|_\infty.$$

Additionally, note that $\mathbf{U}$ is a linear transformation from $\mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ to itself, so we can express the action of $\mathbf{U}$ on Q as

$$(\mathbf{U}Q)(z) = \sum_{z' \in \mathcal{S} \times \mathcal{A}} \mathbf{U}_{z,z'} Q(z').$$

Note moreover that

$$\begin{aligned} \varphi^2(z) &= \sum_{s'} \mathbf{P}_{s',z} (Q(s', \pi_1(s')) - (\mathbf{P}^{\pi_1} Q)(z))^2 \\ \text{and} \quad \tilde{\rho}^2(z) &= \sum_{z'} (\mathbf{U}_{z,z'})^2 \varphi^2(z'). \end{aligned} \tag{26}$$

With these definitions, we now define $\bar{\mathbf{P}}_{y,z}$ as follows[3]:

$$\bar{\mathbf{P}}_{y,z} = \mathbf{P}_{y,z} \left(1 + \frac{\mathbf{U}_{\bar{z},z}}{\tilde{\rho}(\bar{z})\sqrt{2N}} (Q(y, \pi_1(y)) - (\mathbf{P}^{\pi_1} Q)(z))\right) \tag{27}$$

Here we have used the shorthand $\mathbf{P}_{y,z} \equiv \mathbf{P}_a(y \mid s)$, where $z = (s, a) \in \mathcal{S} \times \mathcal{A}$. Let $Q := Q(\mathbf{P}, r)$, and $\bar{Q} := Q(\bar{\mathbf{P}}, r)$ be the optimal Q functions for MDP instances $(\mathbf{P}, r)$ and $(\bar{\mathbf{P}}, r)$ respectively. In the rest of the proof, we use the following properties of $\bar{\mathbf{P}}$.

[3]We will prove that this choice is a valid probability transition kernel shortly.

Lemma 3. *For any MDP $\mathcal{P} = (\mathbf{P}, r)$ and the optimal policy π_1 defined in equation (25), the following properties hold:*

(a) The operator $\bar{\mathbf{P}}$ is a probability transition kernel.

(b) The MDP instances $\mathcal{P} = (\mathbf{P}, r)$ and $\mathcal{P}_1 = (\bar{\mathbf{P}}, r)$ satisfy $d_{Hel}(\mathcal{P}, \mathcal{P}_1) \leqslant \frac{1}{2\sqrt{N}}$ and $\|\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}\|_{\infty \to \infty} \leqslant \frac{1}{\sqrt{2N}}$.

(c) Each entry of $(\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1}[\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}]Q$ is non-negative.

See Appendix B-B for a proof of this lemma.

Equipped with these tools, we are now ready to lower bound the norm $\|Q - \bar{Q}\|_\infty$. The optimal Q-functions Q and $\bar{Q}$ satisfy the following Bellman equations:

$$Q = r + \gamma \mathbf{P}^{\pi_1} Q \qquad \text{and} \qquad \bar{Q} = r + \gamma \bar{\mathbf{P}}^{\bar{\pi}} \bar{Q}, \tag{28}$$

where $\pi_1 \in \mathbf{\Pi}^\star$ is the optimal policy that achieves $\max_{\pi \in \mathbf{\Pi}^\star} \|\rho(\pi; \mathbf{P}, r)\|_\infty$, and $\bar{\pi}$ is an optimal policy for $(\bar{\mathbf{P}}, r)$. By the optimality of policy $\bar{\pi}$ and the Q-function $\bar{Q}$, we have the entrywise inequality $\bar{\mathbf{P}}^{\bar{\pi}} \bar{Q} \geq \bar{\mathbf{P}}^{\pi_1} \bar{Q}$, which implies $(\mathbf{I} - \gamma \bar{\mathbf{P}}^{\pi_1}) \bar{Q} \geq (\mathbf{I} - \gamma \bar{\mathbf{P}}^{\bar{\pi}}) \bar{Q} = r$. Define the quantity

$$\Delta(\pi_1) = (\mathbf{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1} - (\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1}.$$

Using the identity $A_1^{-1} - A_0^{-1} = A_1^{-1}(A_0 - A_1)A_0^{-1}$ for invertible operators A_0 and A_1,

$$\begin{aligned} \bar{Q} - Q &\geq \left[(\mathbf{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1} - (\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1}\right] r \\ &= (\mathbf{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1} \left[(\mathbf{I} - \gamma \mathbf{P}^{\pi_1}) - (\mathbf{I} - \gamma \bar{\mathbf{P}}^{\pi_1})\right] \\ &\qquad \cdot (\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1} r \\ &= \gamma (\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1} \left[\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}\right] (\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1} r \\ &\quad + \gamma \cdot \Delta(\pi_1) (\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}) (\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1} r \\ &= \gamma (\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1} \left[\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}\right] Q \\ &\quad + \gamma \cdot \Delta(\pi_1) (\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}) Q, \end{aligned}$$

where the final equation follows from the Bellman optimality condition (28). Lemma 3(c) guarantees that the entries of $(\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1}[\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}]Q$ are non-negative, and therefore we conclude

$$\begin{aligned} \|\bar{Q} - Q\|_\infty \geqslant \gamma \|(\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1} [\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}] Q\|_\infty \\ - \gamma \|\Delta(\pi_1)(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}) Q\|_\infty. \end{aligned} \tag{29}$$

Consider the second term $T_2 := \|\Delta(\pi_1)(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})Q\|_\infty$. We have

$$\begin{aligned} T_2 &\leqslant \|\!|(\mathbf{I} - \gamma\bar{\mathbf{P}}^{\pi_1})^{-1}(\mathbf{I} - \gamma\mathbf{P}^{\pi_1}) - \mathbf{I}|\!\|_{\infty\to\infty} \\ &\quad \cdot \|(\mathbf{I} - \gamma\mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})Q\|_\infty \\ &= \gamma\|\!|(\mathbf{I} - \gamma\bar{\mathbf{P}}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})|\!\|_{\infty\to\infty} \\ &\quad \cdot \|(\mathbf{I} - \gamma\mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})Q\|_\infty \\ &\leqslant \frac{\gamma}{1-\gamma}\|\!|\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}|\!\|_{\infty\to\infty} \\ &\quad \cdot \|(\mathbf{I} - \gamma\mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})Q\|_\infty \\ &\leqslant \frac{\gamma}{2}\|(\mathbf{I} - \gamma\mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})Q\|_\infty, \end{aligned}$$

where the last inequality uses Lemma 3(b) and the first part of the minimum sample size assumption (11b). Combining this result with the bound (29) we conclude

$$\|\bar{Q} - Q\|_\infty \geqslant \frac{\gamma}{2}\|(\mathbf{I} - \gamma\mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})Q\|_\infty.$$

With this result in hand, substituting the value of the transition kernel $\bar{\mathbf{P}}$ from equation (27) and recalling the definition of state-action pair z from equation (25) we have

$$\begin{aligned} \sqrt{N}\cdot\|\bar{Q} - Q\|_\infty &\geqslant \frac{\gamma\sqrt{N}}{2}(\mathbf{I} - \gamma\mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})Q(\bar{z}) \\ &= \frac{\gamma\sqrt{N}}{2\sqrt{2}}\sum_z \mathbf{U}_{\bar{z},z}\cdot(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})Q(z) \\ &\overset{(i)}{\geqslant} \frac{\gamma}{4\tilde{\rho}(\bar{z})}\sum_z(\mathbf{U}_{\bar{z},z})^2\varphi^2(z) \\ &\overset{(ii)}{=} \frac{\gamma\tilde{\rho}(\bar{z})}{4} = \frac{1}{4}\cdot\max_{\pi\in\mathbf{\Pi}^\star}\|\gamma\rho(\pi;\mathbf{P},r)\|_\infty, \end{aligned}$$

where step (i) follows by substituting the value of the transition kernel $\bar{\mathbf{P}}$ (cf. Proof of Lemma 3 part (c)), and step (ii) follows by using the expression (26). This completes the proof of part (a) of Lemma 2.

2) Proof of Lemma 2(b): Borrowing the notation from part (a) of the proof, let z denote a generic element of the state-action set $\mathcal{S}\times\mathcal{A}$. Let $\pi_2 \in \arg\max_{\pi\in\mathbf{\Pi}^\star}\|\sigma(\pi;\mathbf{P},r)\|_\infty$. We use the shorthands

$$\begin{aligned} \sigma^2(\bar{z}) &:= \max_{\pi\in\mathbf{\Pi}^\star}\|\sigma(\pi;\mathbf{P},r)\|_\infty^2 = \|\sigma(\pi_2;\mathbf{P},r)\|_\infty^2, \\ &\text{and} \quad \mathbf{U} := (\mathbf{I} - \gamma\mathbf{P}^{\pi_2})^{-1}. \end{aligned} \tag{30}$$

We define our perturbed reward function to be

$$\bar{r}(z) = r(z) + \frac{1}{\sigma(\bar{z})\sqrt{2N}}\mathbf{U}_{\bar{z},z}\sigma_r^2 \quad \text{for} \quad z\in\mathcal{S}\times\mathcal{A}. \tag{31}$$

For $\mathcal{P}_2 := (\mathbf{P},\bar{r})$, a short computation shows that the Hellinger distance between the components of the instance pair $(\mathcal{P},\mathcal{P}_2)$ takes the form

$$\begin{aligned} d_{\text{Hel}}^2(\mathcal{P},\mathcal{P}_2) &\leqslant D_{KL}(\mathcal{N}(r,\sigma_r^2\mathbf{I}) \mid \mathcal{N}(\bar{r},\sigma_r^2\mathbf{I})) \\ &= \frac{1}{2\sigma_r^2}\|r - \bar{r}\|_2^2. \end{aligned}$$

Substituting the value of the reward $\bar{r}$ from equation (31) yields

$$\begin{aligned} d_{\text{Hel}}^2(\mathcal{P},\mathcal{P}_2) &\leqslant \frac{1}{2\sigma_r^2}\|\bar{r} - r\|_2^2 \\ &= \frac{1}{\sigma^2(\bar{z})\cdot 4N}\sum_z(\mathbf{U}_{\bar{z},z})^2\sigma_r^2 \\ &= \frac{1}{4N}, \end{aligned}$$

where the last equality uses the definition

$$\sigma^2(\bar{z}) = \sum_{z'}(\mathbf{U}_{\bar{z},z'})^2\sigma_r^2. \tag{32}$$

It remains to prove a lower bound on the ℓ_∞-norm between the optimal Q-functions for instances $\mathcal{P}$ and $\mathcal{P}_2$.

Let $Q := Q(\mathbf{P},r)$, and $\bar{Q} := Q(\mathbf{P},\bar{r})$ be the optimal Q functions for MDP instances $\mathcal{P} := (\mathbf{P},r)$ and $\mathcal{P}_2 := (\mathbf{P},\bar{r})$, respectively. Note that Q and $\bar{Q}$ satisfy the Bellman equations

$$Q = r + \gamma\mathbf{P}^{\pi_2}Q, \quad \text{and} \quad \bar{Q} = \bar{r} + \gamma\mathbf{P}^{\bar{\pi}}\bar{Q}, \tag{33}$$

where $\bar{\pi}$ is an optimal policy for the MDP instance $(\mathbf{P},\bar{r})$. By the optimality of policy $\bar{\pi}$, we have the entrywise inequality $\mathbf{P}^{\bar{\pi}}\bar{Q} \geq \mathbf{P}^{\pi_2}\bar{Q}$; as a result, we have

$$(\mathbf{I} - \gamma\mathbf{P}^{\pi_2})\bar{Q} \geq \bar{r} \implies \bar{Q} \geq (\mathbf{I} - \gamma\mathbf{P}^{\pi_2})^{-1}\bar{r},$$

where the last step uses the fact that $(\mathbf{I} - \gamma\mathbf{P}^{\pi_2})^{-1}$ is entry-wise non-negative. Combining the last inequality with the Bellman equation (33) we have that

$$\bar{Q} - Q \geq (\mathbf{I} - \gamma\mathbf{P}^{\pi_2})^{-1}(\bar{r} - r) \tag{34}$$

and that

$$\begin{aligned} \|(\mathbf{I} - \gamma\mathbf{P}^{\pi_2})^{-1}(\bar{r} - r)\|_\infty &\geqslant (\mathbf{I} - \gamma\mathbf{P}^{\pi_2})^{-1}(\bar{r} - r)(\bar{z}) \\ &= \frac{1}{\sigma(\bar{z})\sqrt{2N}}\sum_z(\mathbf{U}_{\bar{z},z})^2\sigma_r^2 \\ &= \frac{\sigma(\bar{z})}{\sqrt{2N}}, \end{aligned}$$

where the last equality uses the relation (32). Putting together the pieces, we have shown that

$$\|\bar{Q} - Q\|_\infty \geqslant \frac{\sigma(\bar{z})}{\sqrt{2N}} = \frac{1}{\sqrt{2N}} \cdot \max_{\pi \in \mathbf{\Pi}^\star} \|\sigma(\pi; \mathbf{P}, r)\|_\infty,$$

as desired.

IV. Proof of Theorem 2

In this section, we provide a proof of the upper bounds stated in Theorem 2. Throughout the proof, we adopt the following shorthands

$$\begin{aligned} \tau^* &= \max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r)\|_\infty \cdot \sqrt{\log(8DM|\mathbf{\Pi}^\star|/\delta)}, \\ \tau_{\max} &= \frac{1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}}{(1-\gamma)^{1.5}} \cdot \sqrt{\log(8DM/\delta)}. \\ \text{and} \quad \kappa &= \frac{\|Q^\star\|_{\text{span}}}{(1-\gamma)} \cdot \log(8DM/\delta). \end{aligned} \tag{35}$$

Our proof is based on the following two lemmas characterizing the behavior of VR-QL across epochs.

Lemma 4. *Under the assumptions of Theorem 2, for each epoch $m = 1, \ldots, M$, we have*

$$\|\bar{Q}_{m+1} - Q^\star\|_\infty \leqslant \frac{\|\bar{Q}_m - Q^\star\|_\infty}{16} + c\left(\frac{\tau_{\max}}{\sqrt{N_m}} + \frac{\kappa}{N_m}\right), \tag{36}$$

with probability at least $1 - \frac{\delta}{M}$.

Lemma 4 follows by an argument similar to that used in the proof of Theorem 1 of the paper [Wai19c], so we omit the details here. See also the proof of Lemma 5 for some relevant arguments. We remark that the paper [Wai19c] uses the lemma to establish the minimax optimality of VR-QL; however, this is insufficient for our purposes, given our goal of proving instance optimality.

Lemma 5. *Under the assumptions of Theorem 2, for epochs m large enough such that the re-centering sample size N_m satisfies the bound $N_m \geqslant \log_4(8DM/\delta)\frac{(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{\Delta^2(1-\gamma)^3}$, we have*

$$\|\bar{Q}_{m+1} - Q^\star\|_\infty \leqslant \frac{\|\bar{Q}_m - Q^\star\|_\infty}{16} + c \cdot \left(\frac{\tau^*}{\sqrt{N_m}} + \frac{\kappa}{N_m}\right), \tag{37}$$

with probability at least $1 - \frac{\delta}{M}$.

See Section IV-B for the proof of Lemma 5.

A. Completing the proof

Using the two lemmas above, we can now complete the proof of Theorem 2(a). Recalling the epoch sample size formula (21a)

$$N_m = c_1 \frac{4^m}{(1-\gamma)^2} \cdot \log_4(16MD/\delta),$$

we see that the bound (37) holds for all epochs

$$m \geqslant m^* := \log_2 \frac{1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}}{\Delta\sqrt{1-\gamma}}.$$

Observe that the minimum sample size requirement from Theorem 2 ensures that $M \geqslant m^*$. Now, applying the recursions (37) and (36) we obtain

$$\begin{aligned} &\|\bar{Q}_{M+1} - Q^\star\|_\infty \\ &\leqslant \frac{\|\bar{Q}_M - Q^\star\|_\infty}{16} + c\left(\frac{\tau^*}{\sqrt{N_M}} + \frac{\kappa}{N_M}\right) \\ &\overset{(i)}{\leqslant} \frac{\|\bar{Q}_{m*} - Q^\star\|_\infty}{16^{M-m*}} \\ &\quad + c\left(\sum_{k=0}^{M-m^*} \frac{\tau^*}{16^k\sqrt{N_{M-k}}} + \frac{\kappa}{16^k \cdot N_{M-k}}\right) \\ &\overset{(ii)}{\leqslant} \frac{\|\bar{Q}_1 - Q^\star\|_\infty}{16^M} + c \cdot \sum_{k=0}^{M} \frac{\kappa}{16^k N_{M-k}} \\ &\quad + c\left(\sum_{k=0}^{M-m^*} \frac{\tau^*}{16^k\sqrt{N_{M-k}}}\right) \\ &\quad + c\left(\sum_{k=M-m^*+1}^{M} \frac{\tau_{\max}}{16^k\sqrt{N_{M-k}}}\right) \\ &\overset{(iii)}{\leqslant} \frac{\|\bar{Q}_1 - Q^\star\|_\infty}{16^M} + c\left(\frac{\tau_{\max}}{8^{M-m^*} \cdot \sqrt{N_M}}\right) \\ &\quad + c\left(\frac{\tau^*}{\sqrt{N_M}} + \frac{\kappa}{N_M}\right). \end{aligned}$$

Inequality (i) follows via repeated application of the recursion (37); inequality (ii) follows via repeated application of the recursion (36); and inequality (iii) utilizes the relation $N_{M-k} \cdot 4^k = N_M$ (cf. definition (21a)). Via the union bound, the above inequalities hold simultaneously with probability at least $1 - \delta$. Next, note that by our choice of N_m, we have the inequality $2N_M \leqslant N \leqslant \frac{8}{3}N_M$. Putting together the pieces, we

conclude that

$$\|\bar{Q}_{M+1} - Q^\star\|_\infty \leqslant c\|\bar{Q}_1 - Q^\star\|_\infty \frac{\log^2((8D/\delta)\log N)}{N^2(1-\gamma)^4} + \frac{c(1+\|r\|_\infty + \sigma_r\sqrt{1-\gamma})^4}{(1-\gamma)^{1.5}\sqrt{N}} \cdot \frac{\log^2((8D/\delta)\cdot\log N)}{N^{3/2}(1-\gamma)^{\frac{9}{2}}\Delta^3} + c\left(\sqrt{\frac{\log_4(8DM|\mathbf{\Pi}^\star|/\delta)}{N}} \max_{\pi^\star\in\mathbf{\Pi}^\star} \|\nu(\pi^\star;\mathbf{P},r,\gamma)\|_\infty + \frac{\log_4(8DM|\mathbf{\Pi}^\star|/\delta)}{N} \cdot \frac{\|Q^\star\|_{\text{span}}}{1-\gamma}\right). \quad (38)$$

Substituting the lower bound condition

$$\frac{N}{\log^2(N)} \geqslant c\log(D/\delta)\frac{(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^3} \cdot \max\left\{1, \frac{1}{\Delta^2\cdot(1-\gamma)^\beta}\right\}$$

yields the claimed bound. All that remains is to verify the choice of batch sizes $\{N_m\}_{M=1}^M$ is a valid choice, i.e., we need to verify that the algorithm VR-QL with parameter choices (21) uses at most N samples. Recall that the total number of samples used in the M epochs is given by $KM + \sum_{m=1}^M N_m$. Substituting the values of N_m and M from equations (21) we obtain

$$\begin{aligned} KM + \sum_{m=1}^M N_m &\leqslant c\log_4(8DM/\delta)\sum_{m=1}^M \frac{4^m}{(1-\gamma)^2} + \frac{N}{2} \\ &\leqslant c'\log_4(8DM/\delta)\frac{4^M}{(1-\gamma)^2} \\ &\leqslant \frac{N}{2} + \frac{N}{2} \leqslant N. \end{aligned}$$

This completes the proof of Theorem 2(a).

a) Comment on the lower-order terms:: Here, we argue that the first two terms in the right-hand side of the bound (38) are of lower order. A careful look at the proof reveals that for any $p \geqslant 1$ by increasing our choice of N_m by a constant factor depending on p, we can bound the first term by

$$c_1 \cdot \frac{\|\bar{Q}_1 - Q^\star\|_\infty}{N^p} \cdot \frac{\log^p((8D/\delta)\cdot\log N)}{(1-\gamma)^{2p}},$$

and the second term by

$$c_2 \cdot \frac{(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^{3q+1}}{(1-\gamma)^{1.5}\sqrt{N}} \cdot \frac{\log^{2q}((8D/\delta)\cdot\log N)}{N^{3q/2}(1-\gamma)^{\frac{9q}{2}}\Delta^{3q}},$$

where $q = \frac{2}{3}p - \frac{1}{3}$, and (c_1, c_2) are universal constants only depending on (p, q). The number of samples satisfies $N \gtrsim \frac{(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{\Delta^2(1-\gamma)^{3+\beta}}$ by assumption, and consequently, the two terms can be made arbitrarily small by increasing (p, q) appropriately. The equation (38) displays the bound for the pair $(p, q) = (2, 1)$.

The only remaining detail is to prove Lemma 5.

B. Proof of Lemma 5

Recall that the update within an epoch takes the form (cf. SingleEpoch)

$$Q_{k+1} = (1-\alpha_k)Q_k + \alpha_k\left\{\hat{\mathbf{T}}_k(Q) - \hat{\mathbf{T}}_k(\bar{Q}_m) + \overline{\mathbf{T}}_{N_m}(\bar{Q}_m)\right\},$$

where $\bar{Q}_m$ represents the input into epoch m. We define the shifted operators and noisy shifted operators for epoch m by

$$\mathbf{J}(Q) = \mathbf{T}(Q) - \mathbf{T}(\bar{Q}_m) + \overline{\mathbf{T}}_{N_m}(\bar{Q}_m) \quad \text{and} \quad \hat{\mathbf{J}}_k(Q) = \hat{\mathbf{T}}_k(Q) - \hat{\mathbf{T}}_k(\bar{Q}_m) + \overline{\mathbf{T}}_{N_m}(\bar{Q}_m). \quad (39)$$

Since both of the operators $\mathbf{T}$ and $\hat{\mathbf{T}}_k$ are γ-contractive in the ℓ_∞-norm, the operators $\mathbf{J}$ and $\hat{\mathbf{J}}_k$ are also γ-contractive operators in the same norm. Let $\widehat{Q}_m$ denote the unique fixed point of the operator $\mathbf{J}$. The roadmap of the proof is to show that at the end of epoch m, the estimate Q_{K+1} is close to the fixed point $\widehat{Q}_m$ for sufficiently large value of the epoch length K and that the fixed point $\widehat{Q}_m$ is closer to $Q^\star$ than the epoch initialization $\bar{Q}_m$ for sufficiently large N_m.

The proof of Lemma 5 relies on two auxiliary lemmas that formalize this intuition. Lemma 6 characterizes the progress of Algorithm VR-QL within an epoch, and Lemma 7 addresses the progress of Algorithm VR-QL over the epochs.

Lemma 6. *Given an epoch length K lower bounded as $K \geqslant c_2\frac{\log(ND/\delta)}{(1-\gamma)^3}$, we have*

$$\|Q_{K+1} - \widehat{Q}_m\|_\infty \leqslant \frac{1}{33}\|\bar{Q}_m - Q^\star\|_\infty + \frac{1}{33}\|\widehat{Q}_m - Q^\star\|_\infty,$$

with probability exceeding $1 - \frac{\delta}{2M}$.

Lemma 6 is borrowed from Khamaru et al. [KPR+21]; see the proof of Lemma 2 in that paper for details.

Our next lemma bounds the difference between the epoch fixed point $\widehat{Q}_m$ and the optimal value function $Q^\star$.

Lemma 7. *Assume that N_m satisfies the bound $N_m \geqslant c \log_4(8DM/\delta) \frac{(1+\|r\|_\infty + \sigma_r\sqrt{1-\gamma})^2}{\Delta^2(1-\gamma)^3}$. Then we have*

$$\|\widehat{Q}_m - Q^\star\|_\infty \leqslant \frac{\|\bar{Q}_m - Q^\star\|_\infty}{33} + c_4 \left\{ \frac{\tau^*}{\sqrt{N_m}} + \frac{\kappa}{N_m} \right\},$$

with probability exceeding $1 - \frac{\delta}{2M}$.

See Appendix C-A for a proof of this lemma.

With these two auxiliary results in hand, completing the proof of Lemma 5 is relatively straightforward. By the triangle inequality, we have

$$\begin{aligned}
\|\bar{Q}_{m+1} - Q^\star\|_\infty &= \|Q_{K+1} - Q^\star\|_\infty \\
&\leqslant \|Q_{K+1} - \widehat{Q}_m\|_\infty + \|\widehat{Q}_m - Q^\star\|_\infty \\
&\overset{(i)}{\leqslant} \left\{ \frac{1}{32}\|\bar{Q}_m - Q^\star\|_\infty + \frac{1}{32}\|\widehat{Q}_m - Q^\star\|_\infty \right\} \\
&\quad + \|\widehat{Q}_m - Q^\star\|_\infty \\
&= \frac{1}{32}\|\bar{Q}_m - Q^\star\|_\infty + \frac{33}{32}\|\widehat{Q}_m - Q^\star\|_\infty \\
&\overset{(ii)}{\leqslant} \frac{1}{32}\|\bar{Q}_m - Q^\star\|_\infty + \frac{1}{32}\left\{\|\bar{Q}_m - Q^\star\|_\infty\right\} \\
&\quad + c\left(\frac{\tau^*}{\sqrt{N_m}} + \frac{\kappa}{N_m} \right) \\
&\leqslant \frac{1}{16}\|\bar{Q}_m - Q^\star\|_\infty + c\left(\frac{\tau^*}{\sqrt{N_m}} + \frac{\kappa}{N_m} \right).
\end{aligned} \tag{40}$$

Here inequality (i) follows from Lemma 6, whereas inequality (ii) follows from Lemma 7. Finally, the two bounds hold jointly with probability at least $1 - \frac{\delta}{M}$ via a union bound.

V. Discussion

The main contribution of this paper was to analyze the fundamental limits of estimating optimal Q-functions using the lens of instance dependence. Our analysis provides upper and lower bounds on the sample size required to estimate the optimal Q-function of a given instance up to a given accuracy; both these bounds involve a functional of the instance that incorporates both the noise in the observation model, and the compounded effects of the noise via the Markovian dynamics. The upper bounds are achieved by an efficient algorithm based on applying a variance reduction scheme to the classical Q-learning algorithm (which is itself a sub-optimal procedure). While our analysis is sharp in terms of its instance dependence, there remain some minor gaps between our upper and lower bounds. In particular, our current techniques lead to a logarithmic gap in the state-action spaces, and our upper bounds depend on sample size conditions that are stronger than those required for the upper bounds.

More broadly, the current analysis was performed in a relatively benign setting, in which the state and action spaces are both finite, there is no function approximation, and the sampling model is i.i.d. (as an instance of the so-called generative model). In related work [MPW23], a subset of the current authors analyzed the instance-dependence of policy evaluation problems in a setting that does involve function approximation and Markovian noise. The evaluation problem is linear in nature, as opposed to the non-linear policy optimization problem studied here. However, it is an interesting direction for future work to develop such extensions for non-linear problems, such as the optimal Q-estimation problem analyzed in this paper, as well as policy gradient methods.

Appendix A
Calculations for Example 1

Here we derive the bound (13). Letting $V^\star$ denote the value function of the optimal policy $\pi^\star$, we have

$$(\mathbf{Z}^{\pi^\star} - \mathbf{P}^{\pi^\star})Q = \begin{bmatrix} | & | \\ (\mathbf{Z}_{a_1} - \mathbf{P}_{a_1})V^\star & 0 \\ | & | \end{bmatrix}. \tag{41}$$

Letting $\mathbf{W} = (\mathbf{I} - \gamma \mathbf{P}_{a_1})^{-1}(\mathbf{Z}_{a_1} - \mathbf{P}_{a_1})Q_{\pi^\star}$ and solving for $(\mathbf{I} - \gamma \mathbf{P}^{\pi^\star})\mathbf{Y} = \gamma(\mathbf{Z}^{\pi^\star} - \mathbf{P}^{\pi^\star})Q$ gives

$$\mathbf{Y} = \gamma \cdot \begin{bmatrix} | & | \\ \mathbf{W} & \gamma\mathbf{W} \\ | & | \end{bmatrix}. \tag{42}$$

Thus, we have

$$\begin{aligned}
\|\nu(\pi^\star; \mathbf{P}_\lambda, r_\lambda)\|_\infty &:= \max_{(s,a)} |\nu(\pi^\star; \mathbf{P}_\lambda, r_\lambda)(s,a)| \\
&= \max_{(s,a)} |\sqrt{\mathrm{Var}(\mathbf{Y})(s,a)}| \\
&\leqslant c \cdot \frac{1}{(1-\gamma)^{1.5-\lambda}}.
\end{aligned}$$

The second equality above follows from the definition (5) of the matrix $\nu(\pi^\star; \mathbf{P}_\lambda, r_\lambda)$, and the last step via some simple calculations.

Appendix B
Auxiliary lemmas for Theorem 1

In this section, we prove the auxiliary lemmas that are the used in the proof of Theorem 1.

A. Proof of Lemma 1

This proof uses standard arguments, in particular following the usual avenue of reducing estimation to testing [Bir83, Wai19a]. For completeness, we provide the details here. We use Q and Q' to denote the optimal Q-functions for problem $\mathcal{P}$ and $\mathcal{P}'$ respectively. We first lower bound the minimax risk over $\mathcal{P}, \mathcal{P}'$ by the averaged risk as follows:

$$\inf_{\hat{Q}_N} \max_{\mathcal{I} \in \{\mathcal{P}, \mathcal{P}'\}} \mathbb{E}_{\mathcal{P}} \left[\|Q - Q(\mathcal{I})\|_\infty\right] \geqslant \frac{1}{2}\left(\mathbb{E}_{\mathcal{P}^N}\|\hat{Q}_N - Q\|_\infty + \mathbb{E}_{(\mathcal{P}')^N}\|\hat{Q}_N - Q'\|_\infty\right).$$

Here $\mathcal{P}^N$ is a product measure that yields N i.i.d. samples from $\mathcal{P}$. Then, for any $\delta \geqslant 0$, we have by Markov's inequality

$$\mathbb{E}_{\mathcal{P}^N}\left[\|\hat{Q}_N - Q\|_\infty\right] + \mathbb{E}_{(\mathcal{P}')^N}\left[\|\hat{Q}_N - Q'\|_\infty\right] \geqslant \delta\left[\mathcal{P}^N\left(\|\hat{Q}_N - Q\|_\infty \geqslant \delta\right) + (\mathcal{P}')^N\left(\|\hat{Q}_N - Q'\|_\infty \geqslant \delta\right)\right].$$

Define $\delta_{01} := \frac{1}{2}\|Q - Q'\|_\infty$, we have

$$\|\hat{Q}_N - Q\|_\infty < \delta_{01} \implies \|\hat{Q}_N - Q'\|_\infty > \delta_{01},$$

yielding

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}^N}\left[\|\hat{Q}_N - Q\|_\infty\right] + \mathbb{E}_{(\mathcal{P}')^N}\left[\|\hat{Q}_N - Q\|_\infty\right] \\
&\geqslant \delta_{01}\left[1 - \mathcal{P}^N\left(\|\hat{Q}_N - Q\|_\infty < \delta_{01}\right) + (\mathcal{P}')^N\left(\|\hat{Q}_N - Q'\|_\infty \geqslant \delta_{01}\right)\right] \\
&\geqslant \delta_{01}\left[1 - \mathcal{P}^N\left(\|\hat{Q}_N - Q'\|_\infty \geqslant \delta_{01}\right) + (\mathcal{P}')^N\left(\|\hat{Q}_N - Q'\|_\infty \geqslant \delta_{01}\right)\right] \\
&\geqslant \delta_{01}\left[1 - \|\mathcal{P}^N - (\mathcal{P}')^N\|_{\mathrm{TV}}\right] \\
&\geqslant \delta_{01}\left[1 - \sqrt{2} d_{\mathrm{Hel}}\left(\mathcal{P}^N, (\mathcal{P}')^N\right)^2\right],
\end{aligned}$$

where $d_{\mathrm{Hel}}(\mathbb{P}, \mathbb{Q})$ denotes the Hellinger distance between distributions $\mathbb{P}$ and $\mathbb{Q}$. Via the tensorization property of the Hellinger distance for independent random variables, we have

$$\begin{aligned}
d^2_{\mathrm{Hel}}\left(\mathcal{P}^N, (\mathcal{P}')^N\right) &= 1 - \left(1 - d^2_{\mathrm{Hel}}\left(\mathcal{P}, \mathcal{P}'\right)\right)^N \\
&\leqslant N d^2_{\mathrm{Hel}}\left(\mathcal{P}, \mathcal{P}'\right).
\end{aligned}$$

Putting together the pieces, we have that

$$\inf_{\hat{Q}_N} \max_{\mathcal{I} \in \{\mathcal{P}, \mathcal{P}'\}} \mathbb{E}_{\mathcal{I}}\left[\|Q - Q(\mathcal{I})\|_\infty\right] \geqslant \frac{\|Q(\mathcal{P}) - Q(\mathcal{P}')\|_\infty}{4}\left(1 - \sqrt{2} N d^2_{\mathrm{Hel}}\left(\mathcal{P}, \mathcal{P}'\right)\right)_+.$$

The desired result then follows from taking a supremum over all positive alternative $\mathcal{P}' \in \mathcal{S}$ and a simple calculation.

B. Proof of Lemma 3

We devote a subsection to each of the three parts of this lemma.

1) Proof of Lemma 3(a): In order to establish that $\bar{\mathbf{P}}$ is a transition kernel, we observe that

$$\sum_{s'} \mathbf{P}_{s',z}(Q(s', \pi_1(s')) - (\mathbf{P}^{\pi_1} Q)(z)) = 0$$

by noting that $(\mathbf{P}^{\pi_1} Q)(z) = \sum_{s'} \mathbf{P}_{s',z} Q(s', \pi_1(s'))$. Thus we conclude $\sum_{s'} \bar{\mathbf{P}}_{s',z} = 1$, establishing that the rows of $\bar{\mathbf{P}}$ sum up to 1. To check non-negativity of entries of $\bar{\mathbf{P}}$ note we have $|\mathbf{U}_{z,z'}| \leqslant \frac{1}{1-\gamma}$, and $2\|Q\|_{\mathrm{span}} \geqslant |Q(s', \pi_1(s')) - (\mathbf{P}^{\pi_1} Q)(z)|$. Combining the last two observation along with the sample size requirement (11b) implies

$$\bar{\mathbf{P}}_{s',z} \geqslant 1 - \frac{1}{\tilde{\rho}(\bar{z})\sqrt{2N}} \cdot \frac{\|\theta\|_{\mathrm{span}}}{1 - \gamma} \geqslant 0,$$

establishing that $\bar{\mathbf{P}}$ defines a valid set of transition kernels.

2) Proof of Lemma 3(b): The proof of part (b) follows by first providing a general upper bound on the Hellinger distance $d_{\mathrm{Hel}}(\mathcal{P}, \mathcal{P}_1)$, and then substituting the values of instances $\mathcal{P}$ and $\mathcal{P}_1$. Concretely, we prove

$$d^2_{\mathrm{Hel}}(\mathcal{P}, \mathcal{P}_1) \overset{(a)}{\leqslant} \frac{1}{2} \cdot \sum_{z,s'} \frac{(\mathbf{P}_{s',z} - \bar{\mathbf{P}}_{s',z})^2}{\mathbf{P}_{s',z}} \overset{(b)}{\leqslant} \frac{1}{4N}. \tag{43}$$

With this result in hand, the claimed bound on $\|\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}\|_{\infty \to \infty}$ is immediate. Indeed,

$$\begin{aligned}
\|\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}\|^2_{\infty \to \infty} &\leqslant \sum_{z,s'} (\mathbf{P}_{s',z} - \bar{\mathbf{P}}_{s',z})^2 \\
&\leqslant \sum_{z,s'} \frac{(\mathbf{P}_{s',z} - \bar{\mathbf{P}}_{s',z})^2}{\mathbf{P}_{s',z}} \leqslant \frac{1}{2N}.
\end{aligned}$$

It remains to prove the bounds (43a) and (43b).

a) Proof of equation (43)a: We use $(\mathbf{Z}, R)$ (respectively $(\mathbf{Z}', R')$) to denote a sample drawn from the distribution P (respectively P'), and $P_{\mathbf{Z}}, P_R$ (respectively $P'_{\mathbf{Z}}, P'_R$) to denote the marginal distribution of $\mathbf{Z}, R$ (respectively $\mathbf{Z}', R'$). By independence of $\mathbf{Z}$ and R (and likewise for $\mathbf{Z}', R'$) we have

$$P = P_{\mathbf{Z}} \otimes P_R, \quad \text{and} \quad P' = P'_{\mathbf{Z}} \otimes P'_R. \tag{44}$$

Let $\mathcal{P}' = (\mathbf{P}', r') \in \mathcal{S}_1$ (so $r' = r$). Via the independence between $\mathbf{Z}$ and R, we have

$$d^2_{\mathrm{Hel}}(P, P') = d^2_{\mathrm{Hel}}(P_{\mathbf{Z}}, P'_{\mathbf{Z}}). \tag{45}$$

For state-action pairs (s,a), $\mathbf{Z}(s,a)$ are independent (and likewise for $\mathbf{Z}'$) so

$$d^2_{\text{Hel}}(P_{\mathbf{Z}}, P'_{\mathbf{Z}}) = 1 - \prod_{s,a}\left(1 - d_{\text{Hel}}(P_{\mathbf{Z}(s,a)}, P_{\mathbf{Z}'(s,a)})\right)^2 \leqslant \sum_{s,a} d^2_{\text{Hel}}(P_{\mathbf{Z}(s,a)}, P_{\mathbf{Z}'(s,a)}).$$

Note that $\mathbf{Z}(s,a)$ and $\mathbf{Z}'(s,a)$ have multinomial distribution with parameters $\mathbf{P}_a(\cdot \mid s)$ and $\mathbf{P}'_a(\cdot \mid s)$ respectively. Therefore,

$$d^2_{\text{Hel}}(P_{\mathbf{Z}(s,a)}, P_{\mathbf{Z}'(s,a)}) \leqslant \frac{1}{2} D_{\chi^2}\left(P_{\mathbf{Z}'(s,a)} \| P_{\mathbf{Z}(s,a)}\right) = \frac{1}{2}\sum_{s'} \frac{(\mathbf{P}_{s',z} - \bar{\mathbf{P}}_{s',z})^2}{\mathbf{P}_{s',z}}.$$

b) Proof of equation (43)b: We have

$$\begin{aligned}
&(2N\tilde{\rho}^2(\bar{z})) \cdot \sum_{z,s'} \frac{(\mathbf{P}_{s',z} - \bar{\mathbf{P}}_{s',z})^2}{\mathbf{P}_{s',z}} \\
&= \sum_z \sum_{s'} \mathbf{P}_{s',z} (\mathbf{U}_{\bar{z},z})^2 (Q(s', \pi_1(s')) - (\mathbf{P}^{\pi_1} Q)(z))^2 \\
&= \sum_z \mathbf{U}^2_{\bar{z},z} \left(\sum_{s'} \mathbf{P}_{s',z} (Q(s', \pi_1(s')) - (\mathbf{P}^{\pi_1} Q)(z))^2 \right) \\
&\overset{(i)}{=} \sum_z (\mathbf{U}_{\bar{z},z})^2 \varphi^2(z) \overset{(ii)}{=} \tilde{\rho}^2(\bar{z})
\end{aligned}$$

Equality (i) follows from the definition

$$\varphi^2(z) = \text{Var}\left(\mathbf{Z}^{\pi_1} Q(z)\right) = \sum_{s'} \mathbf{P}_{s',z} (Q(s', \pi_1(s')) - (\mathbf{P}^{\pi_1} Q)(z))^2, \quad (46)$$

whereas equality (ii) follows from the definition (25), which ensures that

$$\tilde{\rho}^2(\bar{z}) = \text{Var}\left((\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1} \mathbf{Z}^{\pi_1} Q(\bar{z})\right) = \sum_{z'} (\mathbf{U}_{z,z'})^2 \varphi^2(z').$$

3) Proof of Lemma 3(c): The entries of the matrix $\mathbf{U} := (\mathbf{I} - \gamma \mathbf{P}^{\pi_1})^{-1}$ are positive, so that it suffices to show that the vector $(\bar{\mathbf{P}}^{\pi} - \mathbf{P}^{\pi})Q$ is entry-wise positive. We have

$$\begin{aligned}
&(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}) Q(z) \\
&= \sum_{s'} (\bar{\mathbf{P}}_{s',z} - \mathbf{P}_{s',z}) Q(s', \pi_1(s')) \\
&= \sum_{s'} (\bar{\mathbf{P}}_{s',z} - \mathbf{P}_{s',z}) (Q(s', \pi_1(s')) - (\mathbf{P}^{\pi_1} Q)(z)) \\
&= \frac{\mathbf{U}_{\bar{z},z}}{\tilde{\rho}(\bar{z})\sqrt{2N}} \sum_{s'} \mathbf{P}_{s',z} (Q(s', \pi_1(s')) - (\mathbf{P}^{\pi_1} Q)(z))^2 \\
&= \frac{1}{\tilde{\rho}(\bar{z})\sqrt{2N}} \mathbf{U}_{\bar{z},z} \varphi^2(z) \geqslant 0,
\end{aligned}$$

where the second equality follows from the fact that $\sum_{s'} \bar{\mathbf{P}}_{s',z} = \sum_{s'} \mathbf{P}_{s',z} = 1$, the third equality follows by substituting the value of $\bar{\mathbf{P}}$ from equation (27), and the equality in the last line follows from the definition (46). This completes the proof of part (c).

APPENDIX C
AUXILIARY LEMMAS FOR THEOREM 2

In this section, we prove the auxiliary lemmas that are used in the proof of Theorem 2.

A. Proof of Lemma 7

This section is devoted to the proof of Lemma 7 which underlies the proof of Theorem 2. In order to simplify notation, we drop the epoch number m from $\widehat{Q}_m$ and $\bar{Q}_m$ throughout the remainder of the proof. Let $\widehat{\pi}$ and $\pi^\star$ denote the greedy policies with respect to the Q functions $\widehat{Q}$ and $Q^\star$, respectively. Concretely,

$$\pi^\star(s) = \arg\max_{a \in \mathcal{A}} \; Q^\star(s,a) \quad \text{and} \quad \widehat{\pi}(s) = \arg\max_{a \in \mathcal{A}} \; \widehat{Q}(s,a). \quad (47)$$

Ties in the $\arg\max$ are broken by choosing the action a with smallest index.

By the optimality of the policies $\widehat{\pi}$ and $\pi^\star$ for the Q-functions $\widehat{Q}$ and $Q^\star$, respectively, we have

$$Q^\star = r + \gamma \mathbf{P}^{\pi^\star} Q^\star \quad \text{and} \quad \widehat{Q} = \widetilde{r} + \gamma \mathbf{P}^{\widehat{\pi}} \widehat{Q}, \quad \text{where} \quad \widetilde{r} := r + \overline{\mathbf{T}}_{N_m}(\bar{Q}) - \mathbf{T}(\bar{Q}). \quad (48)$$

In order to complete the proof, we require the following auxiliary result.

Lemma 8. *We have, for any optimal policy*

$$\|(\mathbf{I} - \gamma \mathbf{P}^{\pi^\star})^{-1} (\widetilde{r} - r)\|_\infty \leqslant \frac{\|\bar{Q} - Q^\star\|_\infty}{33} + \frac{4 \log_4(8DM/\delta)}{N_m} \cdot \frac{\|Q^\star\|_{\text{span}}}{(1-\gamma)} + 4\sqrt{\frac{\log_4(8DM/\delta)}{N_m}} \max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r)\|_\infty,$$

with probability exceeding $1 - \frac{\delta}{8M}$.

See Appendix C-B for the proof.

By a union bound over the set of optimal policy $\mathbf{\Pi}^\star$, we have that

$$\begin{aligned}&\max_{\pi\in\mathbf{\Pi}^\star}\|(\mathbf{I}-\gamma\mathbf{P}^{\pi^\star})^{-1}(\widetilde{r}-r)\|_\infty\\&\leqslant\frac{\|\bar{Q}-Q^\star\|_\infty}{33}+\frac{4\log_4(8DM|\mathbf{\Pi}^\star|/\delta)}{N_m}\cdot\frac{\|Q^\star\|_{\mathrm{span}}}{(1-\gamma)}\\&\quad+4\sqrt{\frac{\log_4(8DM|\mathbf{\Pi}^\star|/\delta)}{N_m}}\cdot\max_{\pi\in\mathbf{\Pi}^\star}\|\nu(\pi;\mathbf{P},r)\|_\infty\end{aligned}$$

with probability exceeding $1-\frac{\delta}{8M}$. It remains to prove that under the assumptions of Lemma 5, the following bound holds with probability $1-\frac{\delta}{M}$:

$$\|\widehat{Q}-Q^\star\|_\infty\leqslant\max_{\pi^\star\in\mathbf{\Pi}^\star}\|(\mathbf{I}-\gamma\mathbf{P}^{\pi^\star})^{-1}(\widetilde{r}-r)\|_\infty.\quad(49)$$

If $\widehat{\pi}$ is an optimal policy, then the above claim is immediate. We have $Q^\star=r+\gamma\mathbf{P}^{\widehat{\pi}}Q^\star$ by the Bellman optimality condition, and so we have

$$\|\widehat{Q}-Q^\star\|_\infty=\|(\mathbf{I}-\gamma\mathbf{P}^{\widehat{\pi}})^{-1}(\widetilde{r}-r)\|_\infty.$$

The following lemma establishes that $\widehat{\pi}$ is an optimal policy.

Lemma 9. *Given two Q-functions $Q^\star$ and $\widehat{Q}$ and the associated optimal policies $\pi^\star$ and $\widehat{\pi}$, we have*

$$\mathbf{P}^{\widehat{\pi}}Q^\star(s,a)\geqslant\mathbf{P}^{\pi^\star}Q^\star(s,a)-2\|\widehat{Q}-Q^\star\|_\infty$$

for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Moreover, if the batch size satisfies the lower bound $N_m\geqslant c_3\frac{(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^3}\cdot\frac{\log(DM^2/\delta)}{\Delta^2}$, then $\widehat{\pi}$ is an optimal policy with probability at least $1-\frac{\delta}{M}$.

We prove this lemma in Appendix C-C.

B. Proof of Lemma 8

Recall the definition $\widetilde{r}:=\widehat{R}+\gamma(\widehat{\mathbf{Z}}^{\overline{\pi}}-\mathbf{P}^{\overline{\pi}})\bar{Q}$, where $\overline{\pi}$ a policy greedy with respect to $\bar{Q}$; that is, $\overline{\pi}(x)=\arg\max_{u\in\mathcal{A}}\bar{Q}(s,a)$, where we break ties in the $\arg\max$ by choosing the action a with smallest index. We have, using the shorthand $\mathbf{U}=(\mathbf{I}-\gamma\mathbf{P}^{\pi^\star})^{-1}$,

$$\begin{aligned}&\|(\mathbf{I}-\gamma\mathbf{P}^{\pi^\star})^{-1}(\widetilde{r}-r)\|_\infty\\&\leqslant\|\mathbf{U}\big\{(\widehat{\mathbf{Z}}^{\overline{\pi}}\bar{Q}-\widehat{\mathbf{Z}}^{\pi^\star}Q^\star)-(\mathbf{P}^{\overline{\pi}}\bar{Q}-\mathbf{P}^{\pi^\star}Q^\star)\big\}\|_\infty\\&\quad+\|\mathbf{U}\big\{(\widehat{R}-r)+\gamma(\widehat{\mathbf{Z}}^{\pi^\star}-\mathbf{P}^{\pi^\star})Q^\star\big\}\|_\infty.\end{aligned}$$

Observe that the random variable $\widehat{R}$ and $\widehat{\mathbf{Z}}$ are averages of N_m i.i.d. random variables $\{R_i\}$ and $\{\widehat{\mathbf{Z}}_i\}$, respectively. Additionally, entrywise, the random reward is a Gaussian random variable with variance σ_r^2, and by the generative model assumption, the randomness in the random rewards $\{R_i\}$ is independent of the randomness in $\{\widehat{\mathbf{Z}}_i\}$. Consequently, applying Hoeffding's bound on the term involving $\{R_i\}$, a Bernstein bound on the term involving $\{\widehat{\mathbf{Z}}_i\}$ and a union bound yields the following result which holds with probability at least $1-\frac{\delta}{4M}$:

$$\begin{aligned}&\|(\mathbf{I}-\gamma\mathbf{P}^{\pi^\star})^{-1}\left\{(\widehat{R}-r)+\gamma(\widehat{\mathbf{Z}}^{\pi^\star}-\mathbf{P}^{\pi^\star})Q^\star\right\}\|_\infty\\&\leqslant\frac{4}{\sqrt{N_m}}\|\nu(\pi^\star;\mathbf{P},r)\|_\infty\sqrt{\log_4(8DM/\delta)}\\&\quad+\frac{4\|Q^\star\|_{\mathrm{span}}}{(1-\gamma)N_m}\cdot\log_4(8DM/\delta)\\&\leqslant\frac{4}{\sqrt{N_m}}\max_{\pi\in\mathbf{\Pi}^\star}\|\nu(\pi;\mathbf{P},r)\|_\infty\sqrt{\log_4(8DM/\delta)}\\&\quad+\frac{4\|Q^\star\|_{\mathrm{span}}}{(1-\gamma)N_m}\cdot\log_4(8DM/\delta).\end{aligned}$$

Finally, for each state-action pair (s,a) the random variable $(\widehat{\mathbf{Z}}^{\overline{\pi}}\bar{Q}-\widehat{\mathbf{Z}}^{\pi^\star}Q^\star)(s,a)$ has expectation $(\mathbf{P}^{\overline{\pi}}\bar{Q}-\mathbf{P}^{\pi^\star})(s,a)$ with entries uniformly bounded by $2\|\bar{Q}-Q^\star\|_\infty$. Consequently, by a standard application of Hoeffding's inequality combined with the lower bound $N_m\geqslant c_3\frac{4^m}{(1-\gamma)^2}\log_4(8DM/\delta)$, we have

$$\begin{aligned}&\frac{\gamma}{1-\gamma}\cdot\|(\widehat{\mathbf{Z}}^{\overline{\pi}}\bar{Q}-\widehat{\mathbf{Z}}^{\pi^\star}Q^\star)-(\mathbf{P}^{\overline{\pi}}\bar{Q}-\mathbf{P}^{\pi^\star}Q^\star)\|_\infty\\&\qquad\leqslant\frac{\|\bar{Q}-Q^\star\|_\infty}{33},\end{aligned}$$

with probability exceeding $1-\frac{\delta}{4M}$. The statement then follows from combining these two high-probability statements with a union bound.

C. Proof of Lemma 9

We require the following auxiliary result:

Lemma 10. *Given a batch size N_m lower bounded as $N_m\geqslant c_3\frac{\log_4(8DM/\delta)}{(1-\gamma)^2}$, we have*

$$\begin{aligned}\|\widehat{Q}_m-Q^\star\|_\infty\leqslant c_1\cdot&\frac{1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma}}{\sqrt{N_m}(1-\gamma)^{1.5}}\\&\cdot\log_4(8DM^2/\delta)\end{aligned}$$

with probability at least $1-\frac{\delta}{4M}$.

The proof of this lemma exploits the optimality of the policies $\pi^\star$ and $\widehat{\pi}$ with respect to the Q-functions $Q^\star$

and $\widehat{Q}$, respectively. Accordingly, we have for all state action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\begin{aligned}\mathbf{P}^{\hat{\pi}} Q^\star(s,a) &= \mathbf{P}^{\hat{\pi}} \widehat{Q}(s,a) + \mathbf{P}^{\hat{\pi}} Q^\star(s,a) - \mathbf{P}^{\hat{\pi}} \widehat{Q}(s,a) \\ &\geqslant \mathbf{P}^{\pi^\star} \widehat{Q}(s,a) - \|Q^\star - \widehat{Q}\|_\infty \\ &= \mathbf{P}^{\pi^\star} Q^\star(s,a) + \mathbf{P}^{\pi^\star} \widehat{Q}(s,a) \\ &\qquad - \mathbf{P}^{\pi^\star} Q^\star(s,a) - \|Q^\star - \widehat{Q}\|_\infty \\ &\geqslant \mathbf{P}^{\pi^\star} Q^\star(s,a) - 2\|Q^\star - \widehat{Q}\|_\infty. \end{aligned} \tag{50}$$

The first inequality follows from the optimality of the policy $\hat{\pi}$ with respect to the Q-function $\widehat{Q}$. This completes the proof of the first part of the lemma.

Turning to the second part, invoking Lemma 10 with a batch size $N_m \geqslant \frac{(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^3} \cdot \frac{\log(DM^2/\delta)}{\Delta^2}$ guarantees that

$$2\|Q^\star - \widehat{Q}\|_\infty < \Delta.$$

This inequality, combined with the bound (50) and the definition of the optimality gap Δ, implies that $\hat{\pi}$ is an optimal policy.

Proof of Lemma 10: This proof exploits the result of Lemma 4, that with probability at least $1 - \frac{\delta}{M^2}$, we have

$$\begin{aligned}&\|\widehat{Q}_m - Q^\star\|_\infty \\ &\quad\leqslant \frac{\|\bar{Q}_m - Q^\star\|_\infty}{33} + \frac{\log(8DM^2/\delta)}{N_m} \cdot \frac{\|Q^\star\|_{\text{span}}}{1-\gamma} \\ &\quad + \frac{1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma}}{(1-\gamma)^{1.5}} \sqrt{\frac{\log(8DM^2/\delta)}{N_m}}. \end{aligned} \tag{51}$$

For convenience, we use the shorthand

$$b = 1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}.$$

Following an argument similar to the proof of Theorem 2, we have

$$\begin{aligned}&\|\bar{Q}_{m+1} - Q^\star\|_\infty \\ &\leqslant \frac{\|\bar{Q}_m - Q^\star\|_\infty}{16} + c\left\{\frac{b}{(1-\gamma)^{1.5}}\sqrt{\frac{\log(8DM^2/\delta)}{N_m}}\right. \\ &\qquad\left. + \frac{\log(8DM^2/\delta)}{N_m} \cdot \frac{\|Q^\star\|_{\text{span}}}{1-\gamma}\right\} \\ &\overset{(i)}{\leqslant} \frac{\|\bar{Q}_1 - Q^\star\|_\infty}{16^m} + 2c\left\{\frac{b}{(1-\gamma)^{1.5}}\sqrt{\frac{\log(8DM^2/\delta)}{N_m}}\right. \\ &\qquad\left. + \frac{\log(8DM^2/\delta)}{N_m} \cdot \frac{\|Q^\star\|_{\text{span}}}{1-\gamma}\right\} \\ &\overset{(ii)}{\leqslant} \frac{\|r\|_\infty}{\sqrt{1-\gamma}} \cdot \frac{1}{(1-\gamma)\sqrt{N_m}} \cdot \frac{1}{4^m} \\ &\qquad + 2c\left\{\frac{b}{(1-\gamma)^{1.5}}\sqrt{\frac{\log(8DM^2/\delta)}{N_m}}\right. \\ &\qquad\left. + \frac{\log(8DM^2/\delta)}{N_m} \cdot \frac{\|Q^\star\|_{\text{span}}}{1-\gamma}\right\}, \end{aligned} \tag{52}$$

with probability at least $1 - \frac{\delta}{4M}$. Inequality (ii) follows by recursing the first inequality; the last inequality uses the initialization condition $\|\bar{Q}_1 - Q^\star\|_\infty \leqslant \frac{\|r\|_\infty}{\sqrt{1-\gamma}}$, and $N_m \geqslant \frac{4^m}{(1-\gamma)^2}$. Combining the bounds (51) and (52) and using the bounds $\|Q^\star\|_\infty \leqslant \frac{\|r\|_\infty}{1-\gamma}$ and $\|Q^\star\|_{\text{span}} \leqslant 2\|Q^\star\|_\infty$, we find that

$$\|\widehat{Q}_m - Q^\star\|_\infty \leqslant 8c \cdot \frac{b}{\sqrt{N_m}(1-\gamma)^{1.5}} \cdot \log(8DM^2/\delta),$$

with probability at least $1 - \frac{\delta}{4M}$. This completes the proof.

References

[AKY20] Alekh Agarwal, Sham Kakade, and Lin F Yang, *Model-based reinforcement learning with a generative model is minimax optimal*, Conference on Learning Theory, PMLR, 2020, pp. 67–83.

[AMK13] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen, *Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model*, Machine Learning **91** (2013), no. 3, 325–349.

[Ber09] Dimitri P. Bertsekas, *Neuro-dynamic programming*, Springer US, Boston, MA, 2009.

[Bir83] Lucien Birgé, *Approximation dans les espaces métriques et théorie de l'estimation*,

Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **65** (1983), no. 2, 181–237.

[BRS18] Jalaj Bhandari, Daniel Russo, and Raghav Singal, *A finite time analysis of temporal difference learning with linear function approximation*, Conference On Learning Theory, PMLR, 2018, pp. 1691–1692.

[CL15] T Cai and Mark Low, *A framework for estimation of convex functions*, Statistica Sinica **25** (2015), 423–456.

[DR21] John C. Duchi and Feng Ruan, *Asymptotic optimality in stochastic optimization*, The Annals of Statistics **49** (2021), no. 1, 21 – 48.

[DSTM18] Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor, *Finite sample analyses for TD (0) with function approximation*, AAAI Conference on Artificial Intelligence, vol. 32, 2018, pp. 6144–6153.

[FWXY20] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang, *A theoretical analysis of deep q-learning*, Learning for Dynamics and Control, PMLR, 2020, pp. 486–489.

[JAZBJ18] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan, *Is q-learning provably efficient?*, Advances in neural information processing systems **31** (2018).

[JJS94] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh, *On the convergence of stochastic iterative dynamic programming algorithms*, Neural Computation **6** (1994), no. 6, 1185–1201.

[JYWJ20] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan, *Provably efficient reinforcement learning with linear function approximation*, Conference on Learning Theory, PMLR, 2020, pp. 2137–2143.

[JZ13] Rie Johnson and Tong Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in Neural Information Processing Systems, vol. 26, 2013, pp. 315–323.

[KPR+21] K. Khamaru, A. Pananjady, F. Ruan, M. J. Wainwright, and M. I. Jordan, *Is temporal difference learning optimal? An instance-dependent analysis*, SIAM J. Math. Data Science **3** (2021), no. 4, 1013–1040.

[KS02] Michael Kearns and Satinder Singh, *Near-optimal reinforcement learning in polynomial time*, Machine learning **49** (2002), 209–232.

[LCC+24] Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi, *Is q-learning minimax optimal? a tight sample complexity analysis*, Operations Research **72** (2024), no. 1, 222–236.

[LFDA16] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel, *End-to-end training of deep visuomotor policies*, Journal of Machine Learning Research **17** (2016), no. 1, 1334–1373.

[LS18] Chandrashekar Lakshminarayanan and Csaba Szepesvari, *Linear stochastic approximation: How far does constant step-size and iterate averaging go?*, AISTATS: Conference on AI and Statistics, vol. 21, PMLR, 2018, pp. 1347–1355.

[LWC+20] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen, *Breaking the sample size barrier in model-based reinforcement learning with a generative model*, Advances in neural information processing systems **33** (2020), 12861–12872.

[MMM14] Odalric-Ambrym Maillard, Timothy A Mann, and Shie Mannor, *How hard is my MDP? "The distribution-norm to the rescue"*, Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 1835–1843.

[MPW23] Wenlong Mou, Ashwin Pananjady, and Martin J. Wainwright, *Optimal oracle inequalities for projected fixed-point equations, with applications to policy evaluation*, Math. Oper. Res. **48** (2023), no. 4, 2308–2336.

[MPWB23] Wenlong Mou, Ashwin Pananjady, Martin J. Wainwright, and Peter L. Bartlett, *Optimal and instance-dependent guarantees for Markovian linear stochastic approximation*, Mathematical Statistics and Learning (2023), To appear, Originally posted as arXiv:2112.12770.

[Put14] Martin L Puterman, *Markov Decision Processes: Discrete stochastic dynamic programming*, John Wiley & Sons, 2014.

[PW20] A. Pananjady and M. J. Wainwright, *Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning*, IEEE Transactions on Information Theory **67** (2020), no. 1, 566–585.

[SB18] Richard S. Sutton and Andrew G. Barto, *Reinforcement learning: An introduction*, second ed., The MIT Press, 2018.

[SHM+16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot, *Mastering the game of Go with deep neural networks and tree search*, Nature **529** (2016), no. 7587, 484–489.

[SJ19] Max Simchowitz and Kevin Jamieson, *Non-asymptotic gap-dependent regret bounds for tabular MDPs*, Advances in Neural Information Processing Systems, vol. 33, 2019, pp. 1153–1162.

[SWW+18] Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye, *Near-optimal time and sample complexities for solving markov decision processes with a generative model*, Advances in Neural Information Processing Systems, vol. 33, 2018, pp. 5192–5202.

[SWWY18] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye, *Variance reduced value iteration and faster algorithms for solving markov decision processes*, ACM-SIAM Symposium on Discrete Algorithms, vol. 29, SIAM, 2018, pp. 770–787.

[Sze97] Csaba Szepesvári, *The asymptotic convergence-rate of Q-learning*, Advances in Neural Information Processing Systems, vol. 10, 1997, pp. 1064–1070.

[TFR+17] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel, *Domain randomization for transferring deep neural networks from simulation to the real world*, International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 23–30.

[Tsi94] John N Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Machine Learning **16** (1994), no. 3, 185–202.

[Vaa98] A. W. van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.

[Wai19a] Martin J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.

[Wai19b] Martin J Wainwright, *Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning*, Tech. report, 2019, arXiv preprint arXiv:1905.06265.

[Wai19c] ———, *Variance-reduced Q-learning is minimax optimal*, Tech. report, 2019, arXiv preprint arXiv:1906.04697.

[WD92] Christopher JCH Watkins and Peter Dayan, *Q-learning*, Machine Learning **8** (1992), no. 3-4, 279–292.

[XKWJ23] Eric Xia, Koulik Khamaru, Martin J Wainwright, and Michael I Jordan, *Instance-dependent confidence and early stopping for reinforcement learning*, Journal of Machine Learning Research **24** (2023), no. 392, 1–43.

[ZB19] Andrea Zanette and Emma Brunskill, *Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds*, International Conference on Machine Learning, PMLR, 2019, pp. 7304–7312.

[ZKB19] Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill, *Almost horizon-free structure-aware best policy identification with a generative model*, Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 5625–5634.