Variance-Reduced Cascade Q-learning: Algorithms and Sample Complexity

Mohammad Boveiri and Peyman Mohajerin Esfahani

Abstract. We study the problem of estimating the optimal Q-function of γ-discounted Markov decision processes (MDPs) under the synchronous setting, where independent samples for all state-action pairs are drawn from a generative model at each iteration. We introduce and analyze a novel model-free algorithm called Variance-Reduced Cascade Q-learning (VRCQ). VRCQ comprises two key building blocks: (i) the established direct variance reduction technique and (ii) our proposed variance reduction scheme, Cascade Q-learning. By leveraging these techniques, VRCQ provides superior guarantees in the ℓ_∞-norm compared with the existing model-free stochastic approximation-type algorithms. Specifically, we demonstrate that VRCQ is minimax optimal. Additionally, when the action set is a singleton (so that the Q-learning problem reduces to policy evaluation), it achieves non-asymptotic instance optimality while requiring the minimum number of samples theoretically possible. Our theoretical results and their practical implications are supported by numerical experiments.

1. Introduction

Markov decision processes (MDPs) and reinforcement learning (RL) are widely used mathematical frameworks for decision-making in dynamic and uncertain environments, with an extensive history of research [Bel57; Wat89; BT96; WKR13; Cen+22; TV97]. In recent years, thanks to the surge in available data and computing power, RL techniques have enjoyed tremendous success across a wide range of applications [MKS15; SHM16; Lev+16; VBC19; Kir+22; Kau+23; Kal+25].

Generally, there are two main approaches to RL: model-based [see, e.g., AMK13; AKY20; Li+24; BR24] and model-free [see, e.g., Jin+18; Gha+11; Wai19b; Wai19c; KYV22; Li+23b; LH19; YGL23]. In model-based RL, we first learn a model of the MDP using a batch of state-transition samples and then use this model to form a control policy. On the other hand, model-free approaches directly update either the value function, representing the expected reward starting from each state, or the policy, which is the mapping from states to their actions. Given that model-free RL algorithms operate online, demand less storage space, and are more expressive, the majority of state-of-the-art RL developments have been within the model-free paradigm [MKS13; MBM16; Sch+15; VBC19].

This paper focuses on model-free algorithms for finite state-action RL problems when we have access to a generative model of MDP [Kak03; KMN02]. That is, a sampling model or a simulator that produces independent samples for all state-action pairs. Our analysis is specifically focused on infinite-horizon discounted MDPs, with the states space $\mathcal{X}$, the actions space $\mathcal{U}$, and the discount factor $\gamma \in (0, 1)$. It is worth noting that although the generative setting is generally simpler than online RL settings [LH14; Jin+18; Li+21], the results and techniques developed within this

Date: May 24, 2025.

The authors are with the Delft Center for Systems and Control ({M.Boveiri, P.MohajerinEsfahani}@tudelft.nl), Delft University of Technology, Delft, The Netherlands. This work is partially supported by the ERC grant TRUST-949796.

framework often extend to more complex settings [see, e.g., AMK13; AOM17]. With the generative setting in mind, we introduce a novel model-free algorithm named Variance-Reduced Cascade Q-learning (VRCQ, Algorithm 2) and analyze its ℓ_∞-based sample complexity, namely the number of samples required for VRCQ to yield an entrywise ϵ-accurate estimate of the optimal Q-function.

The VRCQ algorithm consists of two building blocks: (i) a direct variance reduction technique inspired by [Wai19c; JZ13; RSB12], and (ii) our novel variance reduction scheme, Cascade Q-learning (Algorithm 1). Thanks to these methods, VRCQ offers better theoretical guarantees in the ℓ_∞ norm compared with existing model-free algorithms. To demonstrate this, we examine the sample complexity of VRCQ from two perspectives: (i) the minimax viewpoint, which provides bounds that hold uniformly over large classes of models [AMK13; Gha+11; Wai19c; Li+23a; KYV22], and (ii) the instance-dependent perspective [Kha+21; Kha+24; Li+23b; PW21; LS18; MPW24], offering bounds that hold locally around each problem instance. In the latter case, to simplify the analysis and facilitate comparison with recent works [Kha+21; Kha+24; Mou+23; Li+23b], we focus on the scenario where the action set $\mathcal{U}$ is a singleton (i.e., $|\mathcal{U}| = 1$).

Main Contributions and Connection to Prior Work

The following outlines our main contributions and positions them within the relevant literature.

(i) **Cascade Q-learning (CQ) and its direct variance-reduced extension.** We introduce a novel variance reduction scheme named Cascade Q-learning (CQ, Algorithm 1). By establishing a non-asymptotic upper bound on the performance of CQ, we demonstrate that, thanks to its unique structure, it mitigates the impact of noise (Proposition 1). Moreover, employing the CQ scheme and the direct variance-reduced technique [Wai19c; Sid+18; JZ13], we propose a novel model-free RL algorithm called Variance-Reduced Cascade Q-learning (VRCQ, Algorithm 2). VRCQ follows an epoch-based structure akin to standard variance reduction schemes, such as SVRG [JZ13] and the variance-reduced Q-learning [Wai19c]. In each epoch, we run the CQ algorithm, but we recenter our updates to reduce their variance.

(ii) **Geometric convergence rate over epochs with shorter epoch lengths.** Similar to standard direct variance reduction schemes, VRCQ exhibits a geometric convergence rate as a function of the epoch number (Theorem 1). VRCQ achieves this while utilizing only the order of $(1-\gamma)^{-2}$ samples (up to a logarithmic factor) within the inner loop of each epoch (the epoch length). In contrast, the variance-reduced Q-learning proposed in [Wai19c] requires the order of $(1-\gamma)^{-3}$ samples (up to a logarithmic factor) as the epoch length to achieve a similar outcome. (see Theorem 1 and Remark 2 for more details).

(iii) **Minimax optimality.** We consider the class $\mathcal{M}(\gamma, r_{\max})$ of optimal Q-functions that can be obtained from γ-discounted MDPs with a reward function bounded by $r_{\max}$. Over this class, we show that VRCQ requires at most $\mathcal{O}(\frac{r_{\max}^2 \log\left(\frac{D}{\delta}\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\right)}{\epsilon^2(1-\gamma)^3})$ samples to return an ϵ-accurate solution with probability at least $1-\delta$, where $\epsilon \in (0,1]$ and $D = |\mathcal{X}| \times |\mathcal{U}|$. This upper bound matches the minimax lower bound $\Omega(\frac{r_{\max}^2 \log\left(\frac{D}{\delta}\right)}{\epsilon^2(1-\gamma)^3})$ [AMK13] up to a $\log\left(\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\right)$ factor. In contrast, the worst case sample complexity of Q-learning [Li+23a] and Speedy Q-learning [Gha+11] scale as $(1-\gamma)^{-4}$. Mirror descent value iteration [KYV22] has the worst-case optimal cubic dependency on $(1-\gamma)^{-1}$ when $\epsilon \in (0, \sqrt{1-\gamma}\,]$. This implies that ϵ must be sufficiently small when γ is close to 1. Variance-reduced Q-learning [Wai19c] has the worst-case sample complexity $\mathcal{O}(\frac{r_{\max}^2 \log(\frac{1}{1-\gamma}) \log\left(\frac{D}{\delta}\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\right)}{\epsilon^2(1-\gamma)^3})$ for $\epsilon \in$

$(0, 1]$. As we can see, in the worst-case scenario, VRCQ outperforms variance-reduced Q-learning by a logarithmic factor in the discount complexity $(1-\gamma)^{-1}$. This improvement is the direct consequence of the previously mentioned feature of the VRCQ algorithm, namely achieving geometric convergence over epochs with shorter epoch lengths. While this feature only results in a logarithmic improvement in the worst-case setting, we will demonstrate that by considering a stronger criterion for optimality, known as instance optimality [Kha+21; Kha+24; MPW24; PW21], the distinction in performance between these two algorithms becomes more apparent. A detailed comparison between VRCQ and other algorithms from the minimax perspective is provided in Remark 3.

(iv) **Non-asymptotic instance optimality.** We study the instance-dependent behavior of VRCQ when the action set consists of a single element ($|\mathcal{U}| = 1$), so that the problem of estimating the optimal Q-function reduces to the policy evaluation problem. In this setting, we provide an instance-dependent upper bound on the ℓ_∞-error in the non-asymptotic regime and show that the VRCQ algorithm is instant optimal when the number of samples scales as $(1-\gamma)^{-2}$ (Theorem 3). This sample size requirement matches the lower bound developed in [Kha+21]. By comparison, Polyak-Ruppert averaged Q-learning with both polynomial step size and rescaled linear step size is suboptimal in the non-asymptotic setting [Kha+21; Li+23a]. Additionally, variance-reduced Q-learning [Wai19c] is instance optimal when the number of samples scales as $(1-\gamma)^{-3}$ [Kha+21]. Remark 5 provides a more detailed comparison between VRCQ and other algorithms from the instance-dependent viewpoint.

Algorithm	worst-case sample complexity	ϵ-range
Q-learning[Li+23a]	$\epsilon^{-2}(1-\gamma)^{-4}$	$\epsilon \in (0, 1]$
Speedy Q-learning [Gha+11]	$\epsilon^{-2}(1-\gamma)^{-4}$	$\epsilon \in (0, 1]$
Mirror descent value iteration [KYV22]	$\epsilon^{-2}(1-\gamma)^{-3}$	$\epsilon \in (0, \sqrt{1-\gamma}\,]$
VRCQ (this paper)	$\epsilon^{-2}(1-\gamma)^{-3}$	$\epsilon \in (0, 1]$

Table 1. Upper bounds on the worst-case sample complexity for some algorithms. All logarithmic factors are omitted in the table to simplify the expressions.

Algorithm	Upper bound on the ℓ_∞-error	Sample size requirement
PR averaged Q-learning with rescaled linear step size [Kha+21; Li+23b]	Suboptimal	suboptimal even in the asymptotic regime [Li+23b, Sec. 5]
PR averaged Q-learning with polynomial step size [Kha+21; Li+23b]	$\sqrt{\log(D)}\Big(\gamma v(\mathcal{P}) + \rho(\mathcal{P})\Big)$, optimal up to a logarithmic factor in the dimension	Asymptotically optimal, but suboptimal in the non-asymptotic regime
Variance-reduced Q-learning [Wai19c; Kha+21; Kha+24]	$\sqrt{\log(D)}\Big(\gamma v(\mathcal{P}) + \rho(\mathcal{P})\Big)$, optimal up to a logarithmic factor in the dimension	$\mathcal{O}\big(\frac{\log(D)}{(1-\gamma)^3}\big)$, suboptimal
VRCQ (this paper)	$\sqrt{\log(D)}\Big(\gamma v(\mathcal{P}) + \rho(\mathcal{P})\Big)$, optimal up to a logarithmic factor in dimension	$\mathcal{O}\big(\frac{\log(D)}{(1-\gamma)^2}\big)$, optimal up to a logarithmic factor in the dimension
Lower bound [Kha+21]	$\gamma v(\mathcal{P}) + \rho(\mathcal{P})$	$\Omega\big(\frac{1}{(1-\gamma)^2}\big)$

Table 2. Instant-dependent bounds on the ℓ_∞-norm of error in the non-asymptotic regime when $|\mathcal{U}| = 1$.

The remainder of this paper is organized as follows. In Section 2, we present a review of fundamental concepts related to MDPs and RL. In Section 3, which serves as the central part of this paper, we begin by introducing a novel scheme called cascade Q-learning (CQ), designed to mitigate the impact of noise throughout the horizon. Subsequently, we present a direct variance-reduced extension of CQ, referred to as VRCQ, and examine its sample complexity from both the minimax and instance-dependent perspectives. Section 4 includes illustrative examples and simulation results. In Section 5, we discuss some conclusions and outline potential future directions. The proofs of the main results are provided in Section 6.

Notations. Throughout the paper, we use the following notations. The symbols $\mathbb{R}$ and $\mathbb{R}_{>0}$ represent the sets of real numbers and positive real numbers, respectively. The cardinality of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. Given matrices $A, B \in \mathbb{R}^{n\times m}$, $A \leq B$ means $A_{ij} \leq B_{ij}$ for all $i = 1, .., n$ and $j = 1, .., m$. The ℓ_∞-norm and span seminorm of A are, respectively, defined as $\|A\|_\infty := \max\limits_{i,j} |A_{ij}|$ and $\|A\|_{\text{span}} := \max\limits_{i,j} A_{ij} - \min\limits_{i,j} A_{ij}$. We use $\mathbb{1}$ to denote the all-ones matrix. The function $\log : \mathbb{R}_{>0} \to \mathbb{R}$ represents the natural logarithm.

2. Setting and Problem Description

This section briefly reviews some standard concepts of MDPs and RL. For further background on these topics, we refer the reader to several books (e.g., [Ber17; Put14; Sze09; SB18]). A discounted MDP is a quintuple $(\mathcal{X}, \mathcal{U}, \mathbb{P}, r, \gamma)$, where $\mathcal{X}$ is a finite set of possible states, $\mathcal{U}$ is a finite set of possible actions, $\mathbb{P} = \{\mathbb{P}_u(\,\cdot\, | x) \mid (x, u) \in \mathcal{X} \times \mathcal{U}\}$ is the collection of state-action probability transition functions (when in state x, executing an action u causes a transition to the next state drawn randomly from the transition function $\mathbb{P}_u(\,\cdot\, | x)$), $r : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$ is the reward function (i.e., $r(x, u)$ is the immediate reward collected in state $x \in \mathcal{X}$ when action $u \in \mathcal{U}$ is taken.), and $\gamma \in (0, 1)$ indicates the discount factor. A deterministic policy $\pi : \mathcal{X} \to \mathcal{U}$ is a map from the set of states $\mathcal{X}$ to the set of actions $\mathcal{U}$. The action-value function or Q-function of a given policy π is defined as

$$\Theta^\pi(x, u) := \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n r(x_n, u_n) \mid (x_0, u_0) = (x, u)\right],$$

where $u_n = \pi(x_n)$ for all $n \geq 1$, and the expectation is evaluated with respect to the randomness of the MDP trajectory. Moreover, $\Theta^\star(x, u) = \sup\limits_\pi \Theta^\pi(x, u)$ and $\pi^\star(x) = \arg\max\limits_u \Theta^\star(x, u)$ are called the optimal Q-function and optimal policy, respectively. It is well known from the theory of MDPs that $\Theta^\star$ is the unique fixed point of the Bellman operator. The Bellman operator is mapping from $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{U}|}$ to itself, whose $(x, u)-$entry is given by

$$\mathcal{T}(\Theta)(x, u) := r(x, u) + \gamma \mathbb{E}_{\bar{x}} \max_{\bar{u}} \Theta(\bar{x}, \bar{u}), \qquad \text{where} \quad \bar{x} \sim \mathbb{P}_u(\,\cdot\, | x).$$

In the context of RL, the probability transition functions $\mathbb{P} = \{\mathbb{P}_u(\,\cdot\, | x) \mid (x, u) \in (\mathcal{X}, \mathcal{U})\}$ are unknown. As a result, the Bellman operator cannot be exactly evaluated. In the generative setting of RL, at each iteration n, we observe a sample $x_n(x, u)$ drawn from the transition function $\mathbb{P}_u(\,\cdot\, | x)$ for every pair (x, u). Moreover, we assume at each iteration we have access to a noisy observation $\hat{r}_n(x, u)$ of the reward function with mean $r(x, u)$, and σ_r-sub-Gaussian tails. The rewards $\{\hat{r}(x, u)\}_{(x,u)\in(\mathcal{X},\mathcal{U})}$ are independent across all state-action pairs, as well as the randomness in $x_n(x, u)$. The objective is to compute an approximation of the optimal Q-function based on these observations. The empirical Bellman operator at iteration n is defined as

$$\widehat{\mathcal{T}}_n(\Theta)(x,u) := \hat{r}_n(x,u) + \gamma \max_{\bar{u}} \Theta(x_n, \bar{u}), \qquad \text{where} \quad x_n \equiv x_n(x,u) \sim \mathbb{P}_u(\,\cdot\,\mid x).$$

Note that the empirical Bellman operator is an unbiased estimation of the Bellman operator, i.e., $\mathbb{E}\widehat{\mathcal{T}}_n(\Theta) = \mathcal{T}(\Theta)$. Moreover, the Bellman and empirical Bellman operators are γ-contractive in the ℓ_∞-norm. The matrix $W_n := \widehat{\mathcal{T}}_n(\Theta^\star) - \mathcal{T}(\Theta^\star)$ denotes the effective noise or the Bellman noise associated with the operator $\widehat{\mathcal{T}}_n$. It indicates the failure of $\widehat{\mathcal{T}}_n$ to maintain $\Theta^\star$ as its fixed point [see Wai19b, Sec. 2.2]. It is worth noting that the (x,u)-entry of W_n can be written as

$$W_n(x,u) = \Big(\hat{r}_n(x,u) - r(x,u)\Big) + \gamma\Big(\max_{\bar{u}} \Theta^\star(x_n,\bar{u}) - \mathbb{E}_{\bar{x}} \max_{\bar{u}} \Theta^\star(\bar{x},\bar{u})\Big).$$

The first term on the right-hand side of $W_n(x,u)$ is a sub-gaussian random variable with variance at most σ_r, and the second term is bounded in absolute value by $\gamma\|\Theta^\star\|_{\text{span}}$ and has the variance

$$\sigma(\Theta^\star)(x,u) := \gamma^2 \mathbb{E}_{\tilde{x}}\Big(\max_{\bar{u}} \Theta^\star(\tilde{x},\bar{u}) - \mathbb{E}_{\bar{x}} \max_{\bar{u}} \Theta^\star(\bar{x},\bar{u})\Big)^2.$$

Here the expectations $\mathbb{E}_{\bar{x}}$ and $\mathbb{E}_{\tilde{x}}$ are both computed over $\mathbb{P}_u(\,\cdot\,\mid x)$. The matrix of variances $\sigma(\Theta^\star)$ is referred to as the effective variance matrix, and it plays a central role in the non-asymptotic analysis of stochastic approximation-type RL algorithms [Wai19b; Wai19c; Li+23a; Li+23b].

As the final preliminary point discussed in this section, we note that in the generative setting, since each iteration involves drawing $D = |\mathcal{X}| \times |\mathcal{U}|$ samples, the number of samples is a factor of D larger than the number of iterations.

3. Main Results

In this section, we first introduce a novel scheme called Cascade Q-learning (Algorithm 1) and motivate it from a variance reduction standpoint. We then integrate this scheme with the direct variance-reduced technique [RSB12; JZ13; All18; Sid+18; Wai19c] to develop an algorithm with superior theoretical guarantees (Algorithm 2).

3.1. **Cascade Q-Learning: A New Scheme to Reduce the Effect of Noise**

The pseudo-code of the Cascade Q-learning (CQ) is shown in Algorithm 1. At each iteration, CQ evaluates the empirical Bellman operator $\widehat{\mathcal{T}}_n$ at the point $Y_{n+1} = (1-\lambda)Y_n + \lambda Z_n$. It then calculate Z_{n+1} based on the update rule $Z_{n+1} = (1-\lambda)Z_n + \lambda\widehat{\mathcal{T}}_n(Y_{n+1})$, and average the iterates $\{Y_{i+1}\}_{i=1}^n$. To gain a better understanding of the algorithm, it is noteworthy that if we replace the update rule $Y_{n+1} = (1-\lambda)Y_n + \lambda Z_n$ with $Y_{n+1} = Z_n$, then CQ transforms into Polyak-Ruppert (PR) averaged [PJ92] Q-learning with the constant step size λ. Therefore, generally speaking, CQ can be seen as PR averaged Q-learning, coupled with an additional filtering step (or a momentum term), given by $Y_{n+1} = (1-\lambda)Y_n + \lambda Z_n$. As we demonstrate shortly, thanks to its underlying structure, CQ outperforms Q-learning in its ability to handle noise fluctuations. The following proposition provides an upper bound on the performance of the CQ algorithm.

Proposition 1 (Non-asymptotic guarantee for Cascade Q-learning)**.** *Consider an MDP with discount factor γ and optimal Q-function $\Theta^\star$. Suppose we run Algorithm 1 from the initialization Θ_0 for N_e iterations with the constant step size $\lambda = \frac{1}{\sqrt{N_e}}$. Then, we have*

$$\mathbb{E}\|\Theta_{N_e} - \Theta^\star\|_\infty \le \frac{2\|\Theta_0 - \Theta^\star\|_\infty}{(1-\gamma)\sqrt{N_e}} + \frac{2\gamma\log(2D)\|\Theta^\star\|_{span}}{3(1-\gamma)N_e} + \frac{2\sqrt{2\log(2D)}\big(\|\sigma(\Theta^\star)\|_\infty + \sigma_r\big)}{(1-\gamma)\sqrt{N_e}}. \tag{1}$$

Algorithm 1 Cascade Q-learning (CQ)

```
Require: $N_e$ and $\Theta_0$
  $Y_1 = Z_1 = \Theta_0$
  for $n = 1, ...., N_e$ do
      $Y_{n+1} = (1-\lambda)Y_n + \lambda Z_n$
      $Z_{n+1} = (1-\lambda)Z_n + \lambda\widehat{\mathcal{T}}_n(Y_{n+1})$
        $\Theta_n = \frac{1}{n}\sum_{i=1}^{n} Y_{i+1}$
  end for
```

Some comments on Proposition 1 are in order. The first term on the right-hand side of the above inequality indicates the initialization error, while the second and third terms originate from the fluctuations of the Bellman noise $W_n = \widehat{\mathcal{T}}_n(\Theta^\star) - \mathcal{T}(\Theta^\star)$ in the algorithm. Furthermore, it follows directly from the above inequality that by running CQ for

$$N_e \geq c\Big(\frac{\|\Theta_0 - \Theta^\star\|_\infty^2}{(1-\gamma)^2\epsilon^2} + \frac{\gamma\log(D)\|\Theta^\star\|_{\text{span}}}{(1-\gamma)\epsilon} + \frac{\log(D)\big(\|\sigma(\Theta^\star)\|_\infty^2 + \sigma_r^2\big)}{(1-\gamma)^2\epsilon^2}\Big) \tag{2}$$

iterations, we have $\mathbb{E}\|\Theta_{N_e} - \Theta^\star\| \leq \epsilon$, where c is a universal constant.

Cascade Q-learning versus standard Q-learning. Consider the Q-learning algorithm given by $\Theta_{n+1} = (1-\lambda_n)\Theta_n + \lambda\widehat{\mathcal{T}}_n(\Theta_n)$, with the rescaled linear step size $\lambda_n = \frac{1}{1+(1-\gamma)n}$. Moreover, assume $\sigma_r = 0$. In [Wai19b], it is shown that running Q-learning for

$$N_e \geq c\Big(\frac{\|\Theta_0 - \Theta^\star\|_\infty}{(1-\gamma)\epsilon} + \frac{\log(D)\|\Theta\|_{\text{span}}}{(1-\gamma)^2\epsilon} + \frac{\log^2(D)\big(\|\sigma(\Theta^\star)\|_\infty^2\big)}{(1-\gamma)^3\epsilon^2}\Big) \tag{3}$$

iterations results in an ϵ-accurate solution in expectation. Furthermore, [Wai19b] provides an example demonstrating that (3) is sharp, indicating that this bound is generally unimprovable. It is also worth mentioning that the last term on the right-hand side of (3), representing the variance of the Bellman noise, is the dominant term. A close examination of (2) and (3) reveals that the dependency on the horizon, $(1-\gamma)^{-1}$, in the last two terms (noise-related terms) in CQ, as compared to Q-learning, has been improved. Therefore, in general terms, CQ can be considered a form of variance reduction, where the impact of noise through the horizon $(1-\gamma)^{-1}$ has been reduced. Notably, this improvement has resulted in reducing the predominant dependence on the horizon from $(1-\gamma)^{-3}$ to $(1-\gamma)^{-2}$.

3.2. **Variance-Reduced Cascade Q-learning (VRCQ)**

The full potential of cascade Q-learning becomes apparent when combined with the direct variance reduction technique [RSB12; JZ13; All18; Sid+18; Wai19c]. This method allows us to select step sizes that are independent of the total iteration number, resulting in a faster algorithm.

In this section, we present a direct variance-reduced extension of CQ, referred to as VRCQ (Algorithm 2). VRCQ follows an epoch-based structure akin to standard direct variance reduction schemes [JZ13; All18; Wai19c]. In each epoch, we run the CQ scheme but we recenter our updates to reduce their variance. Similar to variance-reduced Q-learning [Wai19c], this recentering employs an empirical approximation $\widetilde{\mathcal{T}}$ to the population Bellman operator $\mathcal{T}$. The complete description of VRCQ is summarized in Algorithm 2. The overall algorithm is characterized by four key choices: the total number of epochs denoted as M; the sequence of step sizes $\{\lambda(m)\}_{m=0}^{M-1}$; the sequence of epoch lengths $\{N_e(m)\}_{m=0}^{M-1}$, and the sequence of recentering samples $\{N_{\mathcal{T}}(m)\}_{m=0}^{M-1}$. Our initial finding indicates that through a suitable choice of these parameters, we achieve a geometric convergence rate as a function of the epoch number.

Algorithm 2 Variance-Reduced Cascade Q-learning (VRCQ)

```
for $m = 0, ..., M-1$ do
    $\widetilde{\mathcal{T}}(\Theta_m) = \frac{1}{N_{\mathcal{T}}} \sum_{i=1}^{N_{\mathcal{T}}} \widehat{\mathcal{T}}_i(\Theta_m)$
    $Y_1 = Z_1 = \Theta_m$
    for $n = 1, ...., N_e$ do
        $Y_{n+1} = (1-\lambda) Y_n + \lambda Z_n$
        $Z_{n+1} = (1-\lambda) Z_n + \lambda \left( \widehat{\mathcal{T}}_n(Y_{n+1}) - \widehat{\mathcal{T}}_n(\Theta_m) + \widetilde{\mathcal{T}}(\Theta_m) \right)$
    end for
    $\Theta_{m+1} = \frac{1}{N_e} \sum_{n=1}^{N_e} Y_{n+1}$
end for
```

Theorem 1 (Geometric convergence over epochs)**.** *Consider an MDP with the discount factor γ and optimal Q-function $\Theta^\star$. Let $\phi \in (0,1)$ be the desired convergence rate over epochs. suppose we run Algorithm 2 from initial point $\Theta_0 = 0$ for M epochs (for $m = 0, ..., M-1$)*

(a) ***Convergence in expectation:*** *If we choose $\lambda(m) = \frac{1}{\sqrt{N_e(m)}}$, $N_{\mathcal{T}}(m) \geq \frac{32 \log(2D)}{\phi^{2m+2}(1-\gamma)^2}$, and $N_e(m) \geq \frac{13^2 \log(2D)}{\phi^2 (2-\phi^m)^2 (1-\gamma)^2}$, then we have $\mathbb{E}\|\Theta_M - \Theta^\star\|_\infty \leq \phi^M \big(\|\Theta^\star\|_\infty + \sigma_r \big)$.*

(b) ***Convergence with high probability:*** *By setting $N_{\mathcal{T}}(m) \geq \frac{32 \log(\frac{10MD}{\delta})}{\phi^{2m+2}(1-\gamma)^2}$, and $N_e(m) \geq \frac{338 \log(\frac{1690MD}{\phi^2(2-\phi^m)^2(1-\gamma)^2\delta})}{\phi^2(2-\phi^m)^2(1-\gamma)^2}$, $\|\Theta_M - \Theta^\star\|_\infty \leq \phi^M \big(\|\Theta^\star\|_\infty + \sigma_r \big)$ holds with probability at least $1-\delta$.*

Remark 1 (Behavior of the parameters over epochs)**.** As the epoch number m increases, both the recentring sample size $N_{\mathcal{T}}(m)$ and the step-size $\lambda(m)$ increase, while the epoch length $N_e(m)$ decreases. A possible explanation for this behavior is that as $N_{\mathcal{T}}$ increases, the empirical Bellman operator $\widetilde{\mathcal{T}}$ becomes a more accurate estimation of the Bellman operator $\mathcal{T}$, which allows for a more aggressive step size and, consequently, a shorter epoch length. Moreover, it is worth noting that, in practice, the universal constants in the algorithm parameters may be conservative. Typically, smaller constants for epoch lengths and recentering sample sizes, or even a larger constant for the step size, can be employed (see Section 4).

Remark 2 (Geometric convergence with shorter epoch lengths)**.** Similar to the variance-reduced Q-learning [Wai19c], VRCQ exhibits a geometric convergence rate as a function of the epoch number m. VRCQ achieves this while utilizing only $N_e = \mathcal{O}(\frac{\log(\frac{D}{(1-\gamma)\delta})}{(1-\gamma)^2})$ iterations within the inner loop of each epoch. In contrast, variance-reduced Q-learning requires $N_e = \mathcal{O}(\frac{\log(\frac{D}{(1-\gamma)\delta})}{(1-\gamma)^3})$ iterations to achieve a similar outcome [Wai19c, Sec. 3.2]. Generally speaking, this feature is the primary contributing factor behind the improved sample complexity results presented in sections 3.3 and 3.4.

3.3. Global Minimax Analysis of VRCQ

In this section, we examine the worst-case sample complexity of VRCQ and establish a bound on the total number of samples required for VRCQ to return an ϵ-accurate solution with high probability. For simplicity and consistency with the existing literature [AMK13; Wai19c; Li+23a; KYV22], we assume $\sigma_r = 0$. First, it follows directly from Theorem 1.b that running VRCQ for

$M = \log_{\frac{1}{\phi}}(\frac{\|\Theta^\star\|_\infty}{\epsilon})$ epochs results in a ϵ-accurate solution with probability at least $1-\delta$. Moreover, the total number of samples is bounded by

$$S = \sum_{m=0}^{M-1} \Big(N_e(m) + N_{\mathcal{T}}(m)\Big) \leq c\Big(\frac{M\log(\frac{MD}{(1-\gamma)\delta})}{(1-\gamma)^2} + \frac{\log(\frac{MD}{\delta})}{\phi^{2M}(1-\gamma)^2}\Big), \tag{4}$$

for some universal constant c. Taking to account that $\sup_{\Theta^\star \in \mathcal{M}(\Theta^\star,\gamma)} \|\Theta^\star\|_\infty \leq \frac{r_{\max}}{1-\gamma}$, we have

$$S \leq c\Big(\frac{\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\log\big(\frac{D}{(1-\gamma)\delta}\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\big)}{(1-\gamma)^2} + \frac{r_{\max}^2\log\big(\frac{D}{\delta}\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\big)}{\epsilon^2(1-\gamma)^4}\Big), \tag{5}$$

which do not mach the optimal cubic scaling in $\frac{1}{1-\gamma}$. Nevertheless, using a refined analysis, similar to the one done for the standard variance-reduced Q-learning [Wai19b, Sec. 3.4.3], we prove that VRCQ has the optimal sample complexity. At first, suppose that VRCQ is run for $M_{\text{Init}} = \log_{\frac{1}{\phi}}(\frac{1}{\sqrt{1-\gamma}})$ epoches, then its output $\Theta_{M_{\text{Init}}}$ satisfies $\|\Theta_{M_{\text{Init}}} - \Theta^\star\|_\infty \leq \frac{r_{\max}}{\sqrt{1-\gamma}}$ with high probability. Furthermore, based on (5), the number of iterations is bounded by

$$S_{\text{Init}} \leq c\Big(\frac{\log(\frac{1}{(1-\gamma)})\log\big(\frac{D}{(1-\gamma)\delta}\log(\frac{1}{1-\gamma})\big)}{(1-\gamma)^2} + \frac{\log\big(\frac{D}{\delta}\log(\frac{1}{(1-\gamma)})\big)}{(1-\gamma)^3}\Big).$$

The following proposition, inspired by Proposition 1 in [Wai19c], states that running VRCQ from $\Theta_{M_{\text{Init}}}$ for a further logarithmic number of epochs results in an ϵ-accurate solution.

Proposition 2 (Minimax optimality of VRCQ)**.** *Suppose VRCQ is run from the initial point* $\Theta_0 = \Theta_{M_{Init}}$ *for* $M_{Late} = \log_{\frac{1}{\phi}}\left(\frac{\bar{c}r_{max}}{\sqrt{(1-\gamma)\epsilon}}\right)$ *iterations where* $\bar{c} \geq \frac{4\sqrt{2}\log(2)}{r_{max}} + 1$*, and the algorithm parameters are chosen according to Theorem 1.b. Then, the inequality* $\|\Theta_M - \Theta^\star\| \leq \epsilon$ *holds with probability at least* $1-\delta$*.*

Note that, based on (4), the number of iterations required in Proposition 2 is bounded by

$$S_{\text{Late}} \leq c\Big(\frac{\log(\frac{r_{\max}}{\sqrt{1-\gamma}\epsilon})\log\big(\frac{D}{(1-\gamma)\delta}\log(\frac{r_{\max}}{\sqrt{1-\gamma}\epsilon})\big)}{(1-\gamma)^2} + \frac{r_{\max}^2\log\big(\frac{D}{\delta}\log(\frac{r_{\max}}{\sqrt{1-\gamma}\epsilon})\big)}{\epsilon^2(1-\gamma)^3}\Big).$$

As a result, the total number of iterations, counting both the initial iterations required to obtain $\bar{\Theta}$ and later S_{late} iterations, used to obtain this ϵ-accurate solution is bounded by

$$S = S_{\text{init}} + S_{\text{late}} \leq c\Big(\frac{\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\log\big(\frac{D}{(1-\gamma)\delta}\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\big)}{(1-\gamma)^2} + \frac{r_{\max}^2\log\big(\frac{D}{\delta}\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\big)}{\epsilon^2(1-\gamma)^3}\Big). \tag{6}$$

The first term on the right-hand side of the above inequality represents the number of samples used as epoch lengths, i.e., $\sum_m N_e(m)$, while the second term represents the number of samples used for recentring, i.e., $\sum_m N_{\mathcal{T}}(m)$. As we can see, the second term is dominant, and it matches the lower bound $\Omega\Big(\frac{r_{\max}^2\log(\frac{D}{\delta})}{\epsilon^2(1-\gamma)^3}\Big)$, up to a $\log\Big(\log(\frac{r_{\max}}{(1-\gamma)\epsilon})\Big)$ factor. This additional logarithmic term results from using the union bound in the proof of Theorem 1.b.

Remark 3 (VRCQ versus other algorithms: Minimax viewpoint)**.** The worst-case sample complexity of various algorithms has been compiled in Table 1. As we can observe, Q-learning and Speedy Q-learning are suboptimal. Mirror Descent Value Iteration is minimax optimal up to some logarithmic factors; however, it requires ϵ to be sufficiently small when the discount factor γ is close to 1. The variance-reduced Q-learning has the following worst-case sample complexity [see Wai19c,

Sec. 3.4.3]

$$S \le c\Big(\frac{\log\left(\frac{r_{\max}}{(1-\gamma)\epsilon}\right)\log\left(\frac{D}{(1-\gamma)\delta}\log\left(\frac{r_{\max}}{(1-\gamma)\epsilon}\right)\right)}{(1-\gamma)^3} + \frac{r_{\max}^2\log\left(\frac{D}{\delta}\log\left(\frac{r_{\max}}{(1-\gamma)\epsilon}\right)\right)}{\epsilon^2(1-\gamma)^3}\Big). \tag{7}$$

Similar to (6), the first term on the right-hand side of (7) represents the number of samples used as the epoch lengths, and the second term represents the number of samples used for recentring. A comparison between (6) and (7) reveals that both VRCQ and variance-reduced Q-learning employ the same sample size for recentring. However, the number of samples used as epoch lengths in (7) is significantly larger than the corresponding term in (6). Consequently, when considered as functions of the horizon $(1-\gamma)^{-1}$, the right-hand side of (7) is dominated by its first term, resulting in the worst-case sample complexity of $\mathcal{O}\left(\frac{r_{\max}^2\log\left(\frac{1}{1-\gamma}\right)\log\left(\frac{D}{(1-\gamma)\delta}\log\left(\frac{r_{\max}}{(1-\gamma)\epsilon}\right)\right)}{\epsilon^2(1-\gamma)^3}\right)$. Thus, in the worst-case scenario, VRCQ outperforms variance-reduced Q-learning by a logarithmic factor in the discount complexity $(1-\gamma)^{-1}$. In the next section, we will demonstrate that by considering a stronger criterion for optimality, known as instance optimality, the distinction in performance between these two algorithms becomes more apparent.

3.4. Instance-Dependent Analysis of VRCQ When $|\mathcal{U}| = 1$

In this section, we explore the instance-dependent behavior of VRCQ in the non-asymptotic regime, where samples are limited. To simplify the analysis and facilitate the comparison between our algorithm and others [Kha+21; Kha+24; Mou+23; Li+23b; PW21], we focus on the specific scenario where the action space contains only a single action, i.e., $|\mathcal{U}| = 1$. In this case, the problem of estimating the optimal Q-function coincides with the policy evaluation problem. Moreover, a given problem instance can be characterized by the pair $\mathcal{P} = (\mathbb{P}, r)$ along with a discount factor γ, where $r \in \mathbb{R}^D$ represents the reward vector and $\mathbb{P} \in [0,1]^{D\times D}$ denote a row-stochastic (Markov) transition matrix. The value function of the problem instance $\mathcal{P}$ (denoted here by the vector $\Theta^\star(\mathcal{P}) \in \mathbb{R}^D$) is the sum of the infinite-horizon discounted rewards. This value function is the unique fixed point of the Bellman operator $T(\Theta) = r + \gamma\mathbb{P}\Theta$. As discussed in Section 2, in the generative setting, at each iteration, we have access to samples $(\widehat{\mathbb{P}}_n, \hat{r}_n)$, where $\hat{r}_n$ is the noisy observation of the reward, and $\widehat{\mathbb{P}}_n$ denotes a draw of a random matrix with $\{0,1\}$ entries and a single one in each row. The empirical Belman operator is $\widehat{\mathcal{T}}_n(\Theta) = \hat{r}_n + \gamma\widehat{\mathbb{P}}_n\Theta$.

The local minimax risk and complexity measures. Suppose we have N i.i.d samples $\{(\widehat{\mathbb{P}}_n, \hat{r}_n)\}_{i=1}^N$ from our observation model. The *local non-asymptotic minimax risk* for $\Theta^\star(\mathcal{P})$ at an instance $\mathcal{P} = (\mathbb{P}, r)$ is defined as [Kha+21]

$$\mathcal{M}_N(\mathcal{P}) = \sup_{\bar{\mathcal{P}}}\inf_{\widehat{\Theta}_N}\max_{\mathcal{Q}\in\{\mathcal{P},\bar{\mathcal{P}}\}} \sqrt{N}\mathbb{E}_Q\|\widehat{\Theta}_N - \Theta^\star(\mathcal{Q})\|_\infty, \tag{8}$$

where the infimum is taken over all estimators $\widehat{\Theta}_N$ that are measurable functions of N i.i.d observations. Moreover, For a given instance $\mathcal{P} = (\mathbb{P}, r)$ we define the complexity measures

$$v(\mathcal{P}) := \left\|\operatorname{diag}\Big((I-\gamma\mathbb{P})^{-1}\operatorname{cov}_{\widehat{\mathbb{P}}\sim\mathbb{P}}\big(\widehat{\mathbb{P}}\Theta^\star(\mathcal{P})\big)(I-\gamma\mathbb{P})^{-T}\Big)\right\|_\infty^{\frac{1}{2}},$$
$$\rho(\mathcal{P}) := \sigma_r\left\|\operatorname{diag}\Big((I-\gamma\mathbb{P})^{-1}(I-\gamma\mathbb{P})^{-T}\Big)\right\|_\infty^{\frac{1}{2}}.$$

The following theorem lower bounds $\mathcal{M}_N(\mathcal{P})$ using the complexity measures $v(\mathcal{P})$ and $\rho(\mathcal{P})$.

Theorem 2 ([Kha+21] Lower bound on $\mathcal{M}_N(\mathcal{P})$)**.** *There exists a universal constant $c \geq 0$ such that for any instance $\mathcal{P}$, the local non-asymptotic minimax risk is lower bounded as*

$$\mathcal{M}_N(\mathcal{P}) \geq c\big(\gamma v(\mathcal{P}) + \rho(\mathcal{P})\big).$$

This bound is valid for all sample sizes N that satisfy

$$N \geq N_0 := \frac{\max\{\gamma^2, \frac{\|\Theta^\star(\mathcal{P})\|_{span}^2}{v^2(\mathcal{P})}\}}{(1-\gamma)^2}. \tag{9}$$

The statement of Theorem 2 indicates that the local complexity of estimating the value function $\Theta^\star(\mathcal{P})$ induced by the instance $\mathcal{P}$ is governed by the quantities $v(\mathcal{P})$ and $\rho(\mathcal{P})$. The minimum sample size condition is natural since when the rewards are observed with noise (i.e., for any $\sigma_r > 0$), this condition is necessary to obtain an estimate of the value function with $\mathcal{O}(1)$ error (see [Kha+21; PW21] for more details).

The next theorem provides an upper bound on the performance of VRCQ in terms of the local complexity measures $v(\mathcal{P})$ and $\rho(\mathcal{P})$. Similar to Theorem 2, we provide this bound on the expected error. A high-probability bound can also be derived by selecting the algorithm parameters according to Theorem 1.b and following a similar line of reasoning.

Theorem 3 (Non-asymptotic optimality of VRCQ)**.** *Suppose that the input parameters of Algorithm 2 are chosen according to Theorem 1.a. Moreover, suppose that the total sample size N satisfies*

$$N \geq \frac{c\gamma \log(D)}{(1-\gamma)^2}, \tag{10}$$

where c is a sufficiently large universal constant. Then, by running Algorithm 2 from any initial point Θ_0 for $M = \log_{\frac{1}{\phi}}(\frac{\sqrt{1-\phi^2}(1-\gamma)\sqrt{N}}{8\sqrt{\gamma \log 2D}})$ epochs, the resulting output Θ_M satisfies

$$\mathbb{E}\|\Theta_M - \Theta^\star(\mathcal{P})\|_\infty \leq \Big(\frac{8\sqrt{\gamma\log(2D)}}{\sqrt{N}\sqrt{1-\phi^2}(1-\gamma)}\Big)^{1+\log_{\frac{1}{\phi}}(\frac{9}{6+\phi})} \|\Theta_0 - \Theta^\star(\mathcal{P})\|_\infty + \frac{2\|\Theta^\star(\mathcal{P})\|_{span}\log(2D)}{(\frac{4}{3}-\frac{1}{\phi})(1-\phi^2)(1-\gamma)N}$$
$$+ \frac{13}{\sqrt{1-\phi^2}}\big(\rho(\mathcal{P}) + \gamma v(\mathcal{P})\big)\sqrt{\frac{\log 2D}{N}}. \tag{11}$$

Remark 4 (Instance-dependent upper and lower bounds)**.** The first term on the right-hand side of the upper bound (11) depends on the initialization Θ_0. When viewed as a function of the sample size N, this initialization-dependent term can be made smaller than other terms by choosing a suitable convergence rate ϕ. For instance, by setting $\phi = 0.875$, we have $(\log_{\frac{1}{\phi}}(\frac{9}{6+\phi}) \geq 2)$. It should be noted that a careful look at the proof of Theorem 3 reveals that another way to make the initialization-dependent term small is by increasing the epoch length N_e by a constant factor. This indicates that the performance of VRCQ does not depend on the initialization Θ_0 in a significant way. The second and the third terms are the dominating terms. Furthermore, assuming that the minimum sample size requirement (10) is met, the upper bound matches the lower bound up to a logarithmic term in the dimension. Similarly, up to a logarithmic factor in dimension, the minimum sample size requirement in Theorem 3 matches the lower bound (9).

Remark 5 (VRCQ versus other algorithms: Instance-dependent behavior)**.** The instance-dependent guarantees for various algorithms have been collected in Table 2. In [Kha+21], it is shown that while Q-learning with PR averaging is instance-optimal in the asymptotic setting (i.e., when the sample size N converges to infinity), it fails to achieve the correct rate in the non-asymptotic regime, even when the sample size is quite large (see also [Li+23b, Sec. 5] for further discussions about the instance optimally of Q-learning with PR averaging). Moreover, [Kha+21] demonstrate that the variance-reduced Q-learning proposed in [Wai19c] is instance optimal when sample size satisfies the condition $N \geq \frac{c\log(D)}{(1-\gamma)^3}$. This sample size requirement, unlike VRCQ, does not match the lower bound (9). Another notable algorithm is Root-SA proposed in [Mou+23]. This algorithm provides an instance-dependent bound on the span seminorm of the error, assuming that the sample size meets the condition $N \geq \frac{c\log(\frac{D}{1-\gamma})}{(1-\gamma)^2}$ [Mou+23, Corollary 6]. Nevertheless, since $\frac{1}{2}\|\cdot\|_{\text{span}} \leq \|\cdot\|_\infty$, meaning the span seminorm is dominated by the ℓ_∞-norm, bounds on the span seminorm of the error are inherently weaker than those on the ℓ_∞-error. Therefore, since the result is not expressed in the ℓ_∞-norm, we have not included Root-SA in Table 2.

4. Numerical Results

In this section, we validate our results through two numerical simulations: one demonstrating the instance optimality of VRCQ, and the other illustrating its performance on random Garnet MDPs [AMT95; Bha+09].

Example 1 (Instance optimality)**.** Consider the 2-state MDP illustrated in Figure 1, where $p = \frac{4\gamma-1}{3\gamma}$. This sub-family of MDPs is fully parameterized by the pair (γ, β). Assuming $\sigma_r = 0$, straightforward calculations show that $v(\mathcal{P}) = \frac{c}{(1-\gamma)^{1.5-\beta}}$ and $\rho(\mathcal{P}) = 0$, where c is a constant. Hence, according to Theorem 2 the local minimax risk satisfies

$$\mathcal{M}_N(\mathcal{P}) \geq \frac{c}{(1-\gamma)^{1.5-\beta}}. \tag{12}$$

For numerical results, we generate a range of MDPs with different discount factors $\gamma \in [0.96, 0.997]$, keeping the value of β fixed. For each γ, we consider the problem of estimating $\Theta^\star(\mathcal{P})$ using

$$N = \frac{100}{(1-\gamma)^2} \tag{13}$$

samples. It follows directly from (12) and (13) that for an instance-optimal algorithm, a linear relationship is expected between the log ℓ_∞-error and the log complexity $\frac{1}{1-\gamma}$, with a slope of $\frac{1}{2} - \beta$.

Figure 2 shows log-log plots of the ℓ_∞-error as a function of the complexity parameter $\frac{1}{1-\gamma}$ for VRCQ (yellow) and PR averaged Q-learning with four different step sizes, $\lambda_n = n^\eta$, where $\eta \in \{-0.8, -0.7, -0.6, -0.5\}$, along with the theoretical local lower bound (dashed). Each data point is obtained by averaging 500 independent trials. For VRCQ, we utilize the following parameters: $M = 15$, $\phi = 0.95$, $N_{\mathcal{T}}(m) = \frac{0.738}{\phi^{2m}(1-\gamma)^2}$, and $N_e(m) = \frac{5}{(1-\gamma)^2}$. Note that we have $\sum_{m=0}^{14} N_{\mathcal{T}}(m) + N_e(m) \leq N$. Moreover, constant factors in N_e and $N_{\mathcal{T}}$ are smaller than those suggested by Theorem 1.a. Nevertheless, as shown in Figure 2, we observe that VRCQ achieves the instance-optimal rate. In contrast, PR averaged Q-learning with polynomial step size fails to achieve the correct rate, even though it is optimal for sufficiently large sample sizes [Li+23b, Theorem 5.1]. Finally, it is worth noting that the variance-reduced Q-learning is not applicable in this setting because it requires $N_e = \mathcal{O}\left(\frac{1}{(1-\gamma)^3}\right)$ samples as the epoch length whereas in this problem, only $\mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right)$ samples are available [see Kha+21, Theorem 3.3].

Example 2 (Performance on random Garnet MDPs)**.** To compare the practical performance of VRCQ with the standard variance-reduced Q-learning [Wai19c; Kha+24; Kha+21], we consider the problem of estimating the optimal Q-function of randomly generated Garnet MDPs [AMT95; Bha+09]. A Garnet MDP is characterized by three integer parameters: (i) the number of states $|\mathcal{X}|$,

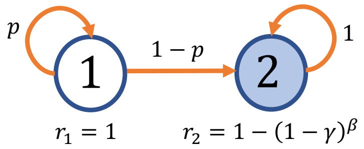
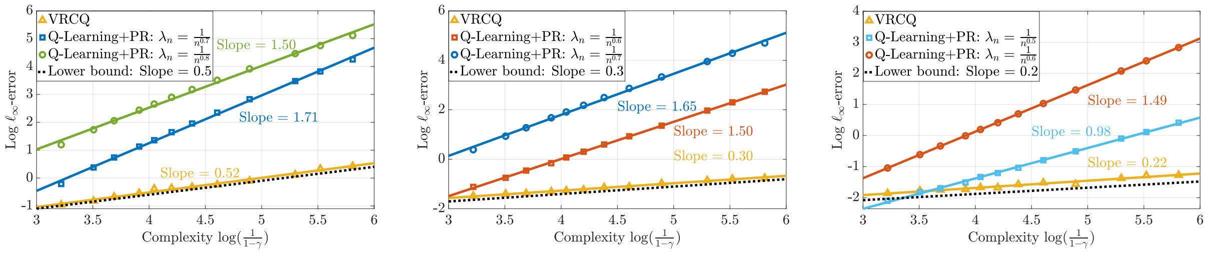


Figure 1. Transition diagram a class of MDP, adopted from [Kha+21]. The scalers $\beta \geq 0$, and $0 < p < 1$ are parameters of the construction. The chain remains in state 1 with probability p and transitions to state 2 with probability $1-p$. State 2 is absorbing.



(a) $\beta = 0$ (b) $\beta = 0.2$ (c) $\beta = 0.3$

Figure 2. Log-log plots of the ℓ_∞-error versus complexity parameter $\frac{1}{1-\gamma}$ for different algorithms. Each data point is an average of 500 independent trials.

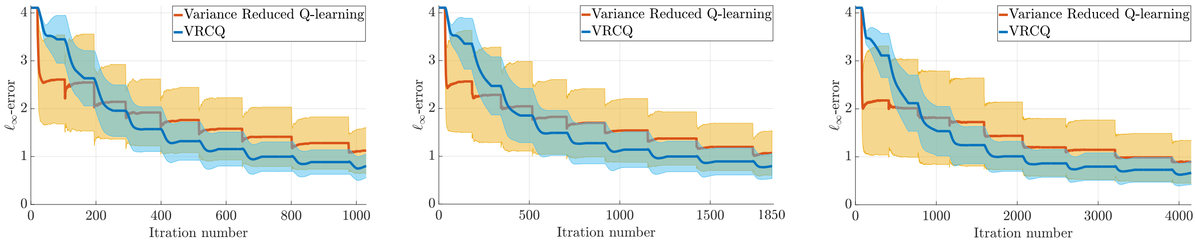
(ii) the number of actions $|\mathcal{U}|$, and (iii) the branching factor b, specifying the number of possible next states for each state-action pair. The next states are selected randomly from the state set without replacement. Moreover, the probability of going to each next state is generated by partitioning the unit interval at $b-1$ randomly chosen cut points between 0 and 1.

In our experiments, we set $|\mathcal{X}| = 20$, $|\mathcal{U}| = 2$, and $b = 2$. We implement VRCQ and variance-reduced Q-learning using identical recentering sample sizes and epoch lengths. Figure 3 plots the ℓ_∞-error of both algorithms versus the number of iterations for three different discount factors. As we can see, in the early iterations, when the error due to initialization is typically larger than the error due to noise fluctuations in the algorithm, the averaged error of variance-reduced Q-learning decreases faster compared to VRCQ. Nevertheless, after a certain number of iterations, VRCQ outperforms variance-reduced Q-learning by achieving a lower averaged error and variance.

5. Conclusion and Future Directions

We studied the problem of estimating the optimal Q-function for γ-discounted MDPs in the synchronous setting. We introduced a novel model-free algorithm, VRCQ, and demonstrated that it is not only minimax optimal, but also, when the action set is a singleton, it achieves non-asymptotic instance optimality while requiring the theoretically minimum number of samples. We conclude this article with the following potential future directions.

(i) **Generalization to cone-contractive operators.** Although our findings were primarily presented within the Q-learning framework, many techniques and results in this study extend beyond the Q-learning problem. For example, by adapting a methodology similar to [Wai19b], Algorithm



(a) $\gamma = 0.80$ (b) $\gamma = 0.85$ (c) $\gamma = 0.90$

Figure 3. Comparison of the convergence behavior of VRCQ and variance-reduced Q-learning. For a given algorithm and value of γ, we run the algorithm for a certain number of epochs, thereby obtaining a path of ℓ_∞-errors at each iteration. We averaged these paths over a total of 500 independent trials. The radius of the shaded area at each iteration represents the standard deviation of the ℓ_∞-error.

1, and Proposition 1 can be generalized to the broader task of finding the fixed point of an operator that is monotone and quasi-contractive with respect to an underlying cone.

(ii) **Studying the instance-dependent behavior of VRCQ for $|\mathcal{U}| > 1$.** Recently, [Kha+24] provided an instance-dependent lower bound similar to Theorem 3 for MDPs with $|\mathcal{U}| > 1$, assuming that the sample size scales as $(1-\gamma)^{-2}$. Given that VRCQ is already instance-optimal for this sample size when $|\mathcal{U}| = 1$, we conjecture that it remains optimal for $|\mathcal{U}| > 1$. Proving this conjecture is an interesting direction for future research, especially considering that, to the best of our knowledge, none of the existing algorithms in the literature matches this lower bound, e.g., Root-SA is instance optimal when the sample size scales as $(1-\gamma)^{-4}$ [Mou+23, Corollary 5], while variance-reduced Q-learning, under certain conditions, is instance optimal when the sample size scales as $(1-\gamma)^{-3}$ [Kha+24, Theorem 2].

(iii) **Generalization to other RL settings.** Another potential future direction is to explore the applicability of our proposed methods in other reinforcement learning settings, such as the episodic setting [Jin+18; AOM17], the asynchronous setting [Li+21; QW20], and Q-learning with function approximation in large (possibly continuous) state-action space [Mey24; YW19].

6. Proofs of the Main Results

6.1. Preliminaries

In this section, we provide the proofs of the theorems and propositions presented in the paper. These proofs are based on the following preparatory lemmas. The first lemma (Lemma 1) establishes a non-asymptotic upper bound on the performance of cascade Q-learning for a given sequence of γ-contractive operators.

Lemma 1. *Consider a sequence of operators $\{\widehat{H}_n : \mathbb{R}^{|\mathbb{X}|\times|\mathbb{U}|} \to \mathbb{R}^{|\mathbb{X}|\times|\mathbb{U}|}\}_{n=1}^{N_e}$ that are γ-contractive in the ℓ_∞-norm and the sequences $\{Y_n\}_{n=1}^{N_e}$ and $\{Z_n\}_{n=1}^{N_e}$ generated by the recursions*

$$Y_{n+1} = (1-\lambda)Y_n + \lambda Z_n, \tag{14a}$$

$$Z_{n+1} = (1-\lambda)Z_n + \lambda \widehat{H}_n(Y_{n+1}), \tag{14b}$$

with the step size $\lambda \in (0,1)$ and initial conditions $Z_1 = Y_1 = \Theta_0$. Then, for any $\Theta_H \in \mathbb{R}^{|\mathbb{X}|\times|\mathbb{U}|}$, we have

$$\Big\|\frac{1}{N_e}\sum_{n=1}^{N_e} Y_{n+1} - \Theta_H\Big\|_\infty \leq \frac{2\|\Theta_0 - \Theta_H\|_\infty}{(1-\gamma)\lambda N_e} + \frac{1}{1-\gamma}\left(\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty\right), \tag{15}$$

where the random matrix sequences $\{P_n^y\}_{n=1}^{N_e}$ and $\{P_n^z\}_{n=1}^{N_e}$ are defined via the recursions

$$P_{n+1}^y = (1-\lambda)P_n^y + \lambda P_n^z, \tag{16a}$$

$$P_{n+1}^z = (1-\lambda)P_n^z + \lambda W_n, \tag{16b}$$

$$W_n = \widehat{H}_n(\Theta_H) - \Theta_H, \tag{16c}$$

and initialized at $P_1^z = P_1^y = 0$.

The next lemma provides upper bounds, both in expectation and with high probability, over the expression $\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty$, the second term on the right-hand side of (15), in some scenarios.

Lemma 2. *Consider the stochastic processes (16a) and (16b) with stepsize $\lambda = \frac{1}{\sqrt{N_e}}$. We denote the fixed points of the Bellman operator $\mathcal{T}$ and the operator $H(\Theta) := \mathcal{T}(\Theta) - \mathcal{T}(\Theta_m) + \widetilde{\mathcal{T}}(\Theta_m)$ by $\Theta^\star$ and $\widehat{\Theta}$, respectively.*

(a) Suppose $\widehat{H}_n = \widehat{\mathcal{T}}_n$, and $\Theta_H = \Theta^\star$, then we have

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^z\|_\infty \leq \frac{2\gamma}{3}\lambda^2\log(2D)\|\Theta^\star\|_{span} + 2\lambda\sqrt{2\log(2D)}\big(\|\sigma(\Theta^\star)\|_\infty + \sigma_r\big), \tag{17a}$$

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty \leq \frac{2\gamma}{3}\lambda^2\log(\frac{8N_eD}{\delta})\|\Theta^\star\|_{span} + 2\lambda\sqrt{2\log(\frac{8N_eD}{\delta})}\big(\|\sigma(\Theta^\star)\|_\infty + \sigma_r\big), \tag{17b}$$

with probability at least $1-\delta$.

(b) Suppose $\widehat{H}_n(\Theta) = \widehat{\mathcal{T}}_n(\Theta) - \widehat{\mathcal{T}}_n(\Theta_m) + \widetilde{\mathcal{T}}(\Theta_m)$ and $\Theta_h = \Theta^\star$. Define the function $C(x,\delta) = \frac{2x}{3}\log(\frac{2D}{\delta}) + \sqrt{2x\log(\frac{2D}{\delta})}$, we have

$$\begin{aligned}\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^z\|_\infty \leq& \big(2\gamma C(\lambda^2,1) + \gamma C(\frac{1}{N_{\mathcal{T}}},1)\big)\|\Theta_m - \Theta^\star\|_\infty\\ &+ \gamma C(\frac{1}{N_{\mathcal{T}}},1)\|\Theta^\star\|_\infty + \sqrt{\frac{2\log(2D)}{N_{\mathcal{T}}}}\sigma_r,\end{aligned} \tag{18a}$$

$$\begin{aligned}\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty \leq& \big(2\gamma C(\lambda^2,\frac{\delta}{5N_eM}) + \gamma C(\frac{1}{N_{\mathcal{T}}},\frac{\delta}{5M})\big)\|\Theta_m - \Theta^\star\|_\infty\\ &+ \gamma C(\frac{1}{N_{\mathcal{T}}},\frac{\delta}{5M})\|\Theta^\star\|_\infty + \sqrt{\frac{2\log(\frac{10MD}{\delta})}{N_{\mathcal{T}}}}\sigma_r,\end{aligned} \tag{18b}$$

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

(c) Suppose $\widehat{H}_n(\Theta) = \widehat{\mathcal{T}}_n(\Theta) - \widehat{\mathcal{T}}_n(\Theta_m) + \widetilde{\mathcal{T}}(\Theta_m)$ and $\Theta_h = \widehat{\Theta}$, then we have

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^z\|_\infty \leq 2\gamma C(\lambda^2,1)\|\Theta_m - \widehat{\Theta}\|_\infty, \tag{19a}$$

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty \le 2\gamma C(\lambda^2, \frac{\delta}{5MN_e})\|\Theta_m - \widehat{\Theta}\|_\infty, \tag{19b}$$

with probability at least $1 - \frac{2\delta}{5M}$.

The following simple lemma becomes extremely useful when attempting to derive an explicit closed-form formula for some parameters of Algorithm 2.

Lemma 3. *Let* $\alpha, \beta, N \in \mathbb{R}_{>0}$. *The inequality* $N \ge \alpha\log(\beta N)$ *holds if* $N \ge \max\{\alpha, 2\alpha\log(\alpha\beta)\}$.

6.2. Proof of Proposition 1

The empirical Bellman operators $\{\widehat{\mathcal{T}}_n\}_{n=1}^{N_e}$ are γ-contractive. Hence, by employing Lemma 1 and setting $\Theta_H = \Theta^\star$, we have

$$\mathbb{E}\Big\|\frac{1}{N_e}\sum_{n=1}^{N_e} Y_{n+1} - \Theta^\star\Big\|_\infty \le \frac{2\|\Theta_0 - \Theta^\star\|_\infty}{(1-\gamma)\lambda N_e} + \frac{1}{1-\gamma}\left(\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^z\|_\infty\right).$$

Substituting (17a) from Lemma 2 into the above inequality gives (1). It is worth noting that by using Lemma 1 and (17b), one can also derive a high probability bound on the ℓ_∞-error.

6.3. Proof of Theorem 1

The operator $\widehat{H}_n(\Theta) = \widehat{\mathcal{T}}_n(\Theta) - \widehat{\mathcal{T}}_n(\Theta_m) + \widetilde{\mathcal{T}}(\Theta_m)$ is γ-contractive. Applying Lemma 1 with $\Theta_H = \Theta^\star$, and taking into account that $Z_0 = Y_0 = \Theta_m$, we find that

$$\Big\|\underbrace{\frac{1}{N_e}\sum_{n=1}^{N_e} Y_{n+1}}_{\Theta_{m+1}} - \Theta^\star\Big\|_\infty \le \frac{2\|\Theta_m - \Theta^\star\|_\infty}{(1-\gamma)\lambda N_e} + \frac{1}{1-\gamma}\left(\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty\right). \tag{20}$$

Proof of Theorem 1.a. By employing (18a) along with the inequality (20), we have

$$\mathbb{E}\|\Theta_{m+1} - \Theta^\star\|_\infty \le \left(\underbrace{\frac{2}{(1-\gamma)\lambda N_e} + \frac{2\gamma}{1-\gamma}C(\lambda^2, 1)}_{\Psi_{N_e}} + \underbrace{\frac{\gamma}{1-\gamma}C(\frac{1}{N_{\mathcal{T}}}, 1)}_{\Psi_{N_{\mathcal{T}}}}\right)\|\Theta_m - \Theta^\star\|_\infty$$
$$+ \underbrace{\frac{\gamma}{1-\gamma}C(\frac{1}{N_{\mathcal{T}}}, 1)}_{\Psi_{N_{\mathcal{T}}}}\|\Theta^\star\|_\infty + \underbrace{\frac{1}{1-\gamma}\sqrt{\frac{2\log(2D)}{N_{\mathcal{T}}}}}_{\Psi_r}\sigma_r.$$

A simple calculation shows that choosing the algorithm parameters according to Theorem 1.a leads to $\Psi_{N_{\mathcal{T}}} \le \frac{\phi^{m+1}}{3}$, $\Psi_{N_e} \le \frac{\phi}{3}(2 - \phi^m)$, and $\Psi_r \le \frac{\phi^{m+1}}{3}$. As a result, we find that

$$\mathbb{E}\|\Theta_{m+1} - \Theta^\star\|_\infty \le \frac{2\phi}{3}\|\Theta_m - \Theta^\star\|_\infty + \frac{\phi^{m+1}}{3}\big(\|\Theta^\star\|_\infty + \sigma_r\big). \tag{21}$$

Using the above inequality, we prove that $\mathbb{E}\|\Theta_M - \Theta^\star\|_\infty \le \phi^M(\|\Theta^\star\|_\infty + \sigma_r)$ via an inductive argument. The base case is trivial. Now, suppose that Θ_m satisfies the bound $\mathbb{E}\|\Theta_m - \Theta^\star\|_\infty \le \phi^m(\|\Theta^\star\|_\infty + \sigma_r)$. By substituting this inequality into (21), we have $\mathbb{E}\|\Theta_{m+1} - \Theta^\star\|_\infty \le \frac{2}{3}\phi^{m+1}(\|\Theta^\star\|_\infty + \sigma_r) + \frac{\phi^{m+1}}{3}(\|\Theta^\star\|_\infty + \sigma_r) \le \phi^{m+1}(\|\Theta^\star\|_\infty + \sigma_r)$, as claimed.

Proof of Theorem 1.b. By substituting (18a) into (20), we find that

$$\|\Theta_{m+1}-\Theta^\star\|_\infty \leq \left(\underbrace{\frac{2}{(1-\gamma)\lambda N_e}+\frac{2\gamma}{1-\gamma}C(\lambda^2,\frac{\delta}{5MN_e})}_{\bar{\Psi}_{N_e}}+\underbrace{\frac{\gamma}{1-\gamma}C(\frac{1}{N_{\mathcal{T}}},\frac{\delta}{5M})}_{\bar{\Psi}_{N_{\mathcal{T}}}}\right)\|\Theta_m-\Theta^\star\|_\infty$$

$$+\underbrace{\frac{\gamma}{1-\gamma}C(\frac{1}{N_{\mathcal{T}}},\frac{\delta}{5M})}_{\bar{\Psi}_{N_{\mathcal{T}}}}\|\Theta^\star\|_\infty+\underbrace{\frac{1}{1-\gamma}\sqrt{\frac{2\log(\frac{10MD}{\delta})}{N_{\mathcal{T}}}}}_{\bar{\Psi}_r}\sigma_r,$$

with probability at least $1-\frac{\delta}{M}$. Selecting the algorithm parameters according to Theorem 1.b gives $\bar{\Psi}_{N_{\mathcal{T}}}\leq\frac{\phi^{m+1}}{3}$, $\bar{\Psi}_r\leq\frac{\phi^{m+1}}{3}$, and $\bar{\Psi}_{N_e}\leq\frac{\phi}{3}(2-\phi^m)$, where the last inequality is obtained by applying Lemma 3. Consequently, we have $\|\Theta_{m+1}-\Theta^\star\|_\infty\leq\frac{2\phi}{3}\|\Theta_m-\Theta^\star\|_\infty+\frac{\phi^{m+1}}{3}\big(\|\Theta^\star\|_\infty+\sigma_r\big)$, with probability at least $1-\frac{\delta}{M}$. The remainder of the proof relies on an inductive argument identical to that used in the last part of the proof of Theorem 1.a, along with the union bound.

6.4. **Proof of Proposition 2**

Following a similar analysis as in [Wai19c, Sec. 4.3], we show that under the stated conditions in Proposition 2, the iterates $\{\Theta_m\}_{m=0}^M$ satisfy

$$\|\Theta_m-\Theta^\star\|_\infty\leq\bar{c}\phi^m\frac{r_{\max}}{\sqrt{1-\gamma}}. \tag{22}$$

with high probability. Note that the above inequality immediately implies $\|\Theta_M-\Theta^\star\|_\infty\leq\epsilon$ for $M=\log_{\frac{1}{\phi}}(\frac{\bar{c}r_{\max}}{\sqrt{(1-\gamma)}\epsilon})$. We prove (22) via an inductive argument. for $m=0$ is trivial. Now, suppose $\|\Theta_m-\Theta^\star\|_\infty\leq\bar{c}\phi^m\frac{r_{\max}}{\sqrt{1-\gamma}}:=b_m$, we show that $\|\Theta_{m+1}-\Theta^\star\|_\infty\leq\phi b_m=b_{m+1}$ with probability at least $1-\frac{\delta}{M}$. First, applying Lemma 1 with $\Theta_H=\widehat{\Theta}$, and taking into account that $Z_0=Y_0=\Theta_m$, we have

$$\|\Theta_{m+1}-\widehat{\Theta}\|_\infty\leq\frac{2\|\Theta_m-\widehat{\Theta}\|_\infty}{(1-\gamma)\lambda N_e}+\frac{1}{1-\gamma}\left(\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty+\frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty\right). \tag{23}$$

Substituting (19b) into (23) and using the triangle inequality gives

$$\|\Theta_{m+1}-\widehat{\Theta}\|_\infty\leq\Big(\frac{2}{(1-\gamma)\lambda N_e}+\frac{2\gamma}{1-\gamma}C(\lambda^2,1)\Big)\|\Theta_m-\widehat{\Theta}\|_\infty\leq\frac{2-\phi^m}{3}b_{m+1}+\frac{\phi(2-\phi^m)}{3}\|\Theta^\star-\widehat{\Theta}\|_\infty,$$

with probability at least $1-\frac{2\delta}{5M}$. Moreover, according to Lemma 4 in [Wai19c], we have

$$\begin{aligned}\|\widehat{\Theta}-\Theta^\star\|_\infty&\leq\frac{4}{3}b_{m+1}\Big(\frac{\gamma}{\phi(1-\gamma)}C(\frac{1}{N_{\mathcal{T}}},\frac{\delta}{5M})+\frac{2\log(\frac{10MD}{\delta})}{3\bar{c}\phi^{m+1}(1-\gamma)^{1.5}N_{\mathcal{T}}}+\frac{(1+\frac{2\log(2)}{r_{\max}})\sqrt{2\log(\frac{10MD}{\delta})}}{\bar{c}\phi^{m+1}(1-\gamma)\sqrt{N_{\mathcal{T}}}}\Big)\\&\leq b_{m+1}\Big(\frac{13}{36}\phi^m+\frac{\phi^{m+1}}{144}+\frac{1}{48}\Big),\end{aligned}$$

with probability at least $1-\frac{3\delta}{5M}$. Finally, by employing the triangle inequality and the union bound, we have

$$\|\Theta_{m+1}-\Theta^\star\|_\infty\leq\|\Theta_{m+1}-\widehat{\Theta}\|_\infty+\|\widehat{\Theta}-\Theta^\star\|_\infty\leq b_{m+1}\underbrace{\Big(\frac{2-\phi^m}{3}+(\frac{\phi(2-\phi^m)}{3}+1)(\frac{13}{36}\phi^m+\frac{\phi^{m+1}}{144}+\frac{1}{48})\Big)}_{\leq1},$$

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

6.5. **Proof of Theorem 3**

Applying Lemma 1 with $\Theta_H = \widehat{\Theta}$, and taking into account that $Z_0 = Y_0 = \Theta_m$, we find that

$$\mathbb{E}\|\Theta_{m+1} - \widehat{\Theta}\|_\infty \leq \frac{2\mathbb{E}\|\Theta_m - \widehat{\Theta}\|_\infty}{(1-\gamma)\lambda N_e} + \frac{1}{1-\gamma}\left(\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^z\|_\infty\right).$$

By substituting (19a) into the above inequality and applying the triangle inequality, we find that

$$\mathbb{E}\|\Theta_{m+1} - \widehat{\Theta}\|_\infty \leq \Big(\frac{2}{(1-\gamma)\lambda N_e} + \frac{2\gamma}{1-\gamma}C(\lambda^2,1)\Big)\Big(\mathbb{E}\|\Theta_m - \Theta^\star(\mathcal{P})\|_\infty + \mathbb{E}\|\Theta^\star(\mathcal{P}) - \widehat{\Theta}\|_\infty\Big). \quad (24)$$

Moreover, according to Lemma 4.9 in [Kha+21], we have

$$\mathbb{E}\|\widehat{\Theta} - \Theta^\star(\mathcal{P})\|_\infty \leq C(\frac{1}{N_\mathcal{T}},1)\|\Theta_m - \Theta^\star(\mathcal{P})\|_\infty + \Big(\rho(\mathcal{P}) + \gamma v(\mathcal{P})\Big)\sqrt{\frac{2\log 2D}{N_\mathcal{T}}} + \frac{\|\Theta^\star(\mathcal{P})\|_{\text{span}}}{3(1-\gamma)N_\mathcal{T}}\log(2D). \quad (25)$$

From the triangle inequality, (24), and (25), we deduce that

$$\begin{aligned}\mathbb{E}\|\Theta_{m+1} - \Theta^\star(\mathcal{P})\|_\infty &\leq \mathbb{E}\|\Theta_{m+1} - \widehat{\Theta}\|_\infty + \mathbb{E}\|\widehat{\Theta} - \Theta^\star(\mathcal{P})\|_\infty \leq \Big(\underbrace{\frac{2}{(1-\gamma)\lambda N_e} + \frac{2\gamma}{1-\gamma}C(\lambda^2,1)}_{\Psi_{N_e}} + \underbrace{\frac{\gamma}{1-\gamma}C(\frac{1}{N_\mathcal{T}},1)}_{\Psi_{N_\mathcal{T}}} \\ &\quad + (\underbrace{\frac{2}{(1-\gamma)\lambda N_e} + \frac{2\gamma}{1-\gamma}C(\lambda^2,1)}_{\Psi_{N_e}})\underbrace{\frac{\gamma}{1-\gamma}C(\frac{1}{N_\mathcal{T}},1)}_{\Psi_{N_\mathcal{T}}}\Big)\|\Theta_m - \Theta^\star(\mathcal{P})\|_\infty + \Big(\underbrace{\frac{2}{(1-\gamma)\lambda N_e} + \frac{2\gamma}{1-\gamma}C(\lambda^2,1)}_{\Psi_{N_e}} \\ &\quad + 1\Big) \times \Big(\big(\rho(\mathcal{P}) + \gamma v(\mathcal{P})\big)\sqrt{\frac{2\log 2D}{N_\mathcal{T}}} + \frac{2\|\Theta^\star(\mathcal{P})\|_{\text{span}}}{3(1-\gamma)N_\mathcal{T}}\log(2D)\Big).\end{aligned}$$

Selecting algorithm parameters according to Theorem 1.a results in $\Psi_{N_\mathcal{T}} \leq \frac{\phi^{m+1}}{3}$, and $\Psi_{N_e} \leq \frac{\phi}{3}(2-\phi^m)$. as a result, we have

$$\mathbb{E}\|\Theta_{m+1} - \Theta^\star(\mathcal{P})\|_\infty \leq \Big(\frac{2\phi}{3} + \frac{\phi^2}{9}\Big)\|\Theta_m - \Theta^\star(\mathcal{P})\|_\infty + 2\Big(\big(\rho(\mathcal{P}) + \gamma v(\mathcal{P})\big)\sqrt{\frac{2\log 2D}{N_\mathcal{T}}} + \frac{\|\Theta^\star\|_{\text{span}}}{3(1-\gamma)N_\mathcal{T}}\log(2D)\Big).$$

As a direct consequence of the above inequality, we have

$$\begin{aligned}\mathbb{E}\|\Theta_M - \Theta^\star\|_\infty \leq &\Big(\frac{2\phi}{3} + \frac{\phi^2}{9}\Big)^M\|\Theta_0 - \Theta^\star\|_\infty + \sum_{m=0}^{M-1}2\Big(\frac{2\phi}{3} + \frac{\phi^2}{9}\Big)^{M-1-m}\Big(\big(\rho(\mathcal{P}) + \gamma v(\mathcal{P})\big)\sqrt{\frac{2\log 2D}{N_\mathcal{T}(m)}} \\ &+ \frac{\|\Theta^\star(\mathcal{P})\|_{\text{span}}\log(2D)}{3(1-\gamma)N_\mathcal{T}(m)}\Big).\end{aligned}$$

Note that $\frac{\phi^{M-1-m}}{\sqrt{N_\mathcal{T}(m)}} = \frac{1}{\sqrt{N_\mathcal{T}(M-1)}}$, and $\frac{\phi^{2(M-1-m)}}{N_\mathcal{T}(m)} = \frac{1}{N_\mathcal{T}(M-1)}$. Hence, we have

$$\mathbb{E}\|\Theta_M - \Theta^\star\|_\infty \leq \Big(\frac{2\phi}{3} + \frac{\phi^2}{9}\Big)^M\|\Theta_0 - \Theta^\star\|_\infty + 9\big(\rho\mathcal{P}) + \gamma v(\mathcal{P})\big)\sqrt{\frac{2\log 2D}{N_\mathcal{T}(M-1)}} + \frac{1}{\frac{4}{3} - \frac{1}{\phi}}\frac{\|\Theta^\star\|_{\text{span}}\log(2D)}{(1-\gamma)N_\mathcal{T}(M-1)}. \quad (26)$$

It remains to express the quantities $\Big(\frac{2\phi}{3} + \frac{\phi^2}{9}\Big)^M$, and $N_\mathcal{T}(M-1)$ in terms of the total number of available samples N, and show that the total number of used samples is bounded by N. We have

$$\Big(\frac{2\phi}{3} + \frac{\phi^2}{9}\Big)^M = \Big((\frac{1}{\phi})^{-M}\Big)^{1+\log_{\frac{1}{\phi}}(\frac{9}{6+\phi})} = \Big(\frac{8\sqrt{\gamma}\ \log 2D}{\sqrt{N}\sqrt{1-\phi^2(1-\gamma)}}\Big)^{1+\log_{\frac{1}{\phi}}(\frac{9}{6+\phi})}. \quad (27)$$

Also, the number of samples used for recentering is

$$\sum_{m=0}^{M-1} N_{\mathcal{T}}(m) \leq \frac{N_{\mathcal{T}}(M-1)}{(1-\phi^2)} = \frac{32\gamma \log(2D)}{(1-\phi^2)\phi^{2M}(1-\gamma)^2} = \frac{N}{2}, \tag{28}$$

where the last equality is obtained via substituting $M = \log_{\frac{1}{\phi}}(\frac{\sqrt{N}\sqrt{1-\phi^2}(1-\gamma)}{8\sqrt{\gamma \log(2D)}})$. Substituting (27) and (28) into (26) gives (11). In addition, the number of samples used as epoch lengths is

$$\sum_{m=0}^{M-1} N_e(m) \leq M N_e(0) = \log_{\frac{1}{\phi}}(\frac{\sqrt{N}\sqrt{1-\phi^2}(1-\gamma)}{8\sqrt{\gamma \ \log(2D)}}) N_e(0) \leq \frac{N}{2},$$

where the second inequality is an immediate consequence of inequality (10) and Lemma 3. As a result, the total number of samples used by VRCQ is

$$\sum_{m=0}^{M-1} N_e(m) + \sum_{m=0}^{M-1} N_{\mathcal{T}}(m) \leq \frac{N}{2} + \frac{N}{2} = N.$$

A. Proofs of the Auxiliary Lemmas

In this appendix, we provide the proofs of the auxiliary lemmas used in Section 6.

A.1. Proof of Lemma 1

We begin by presenting a "sandwich result" that provides both lower and upper bounds for the error sequences $\{Y_n - \Theta_H\}_{n=1}^{N_e}$ and $\{Z_n - \Theta_H\}_{n=1}^{N_e}$.

Lemma 4. *The sequences $\{Y_n\}_{n=1}^{N_e}$ and $\{Z_n\}_{n=1}^{N_e}$ generated by the recursion (14a) and (14b) satisfy the following sandwich inequality*

$$\begin{bmatrix} P_n^y - a_n^y \mathbb{1} \\ P_n^z - a_n^z \mathbb{1} \end{bmatrix} \leq \begin{bmatrix} Y_n - \Theta_H \\ Z_n - \Theta_H \end{bmatrix} \leq \begin{bmatrix} a_n^y \mathbb{1} + P_n^y \\ a_n^z \mathbb{1} + P_n^z \end{bmatrix}, \tag{29}$$

where the non-negative scalar sequences a_n^y, a_n^z are generated by the linear dynamics

$$\begin{bmatrix} a_{n+1}^y \\ a_{n+1}^z \end{bmatrix} = \underbrace{\begin{bmatrix} 1-\lambda & \lambda \\ (1-\lambda)\lambda\gamma & 1-\lambda+\lambda^2\gamma \end{bmatrix}}_{H} \begin{bmatrix} a_n^y \\ a_n^z \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & 0 \\ (1-\lambda)\lambda\gamma & \lambda^2\gamma \end{bmatrix}}_{F} \begin{bmatrix} \|P_n^y\|_\infty \\ \|P_n^z\|_\infty \end{bmatrix},$$

with initial conditions $a_1^y = \|Y_1 - \Theta^\star\|_\infty$ and $a_1^z = \|Z_1 - \Theta^\star\|_\infty$.

By employing Lemma 4, we have

$$\frac{1}{N_e}\sum_{n=1}^{N_e} \begin{bmatrix} P_n^y - a_n^y \mathbb{1} \\ P_n^z - a_n^z \mathbb{1} \end{bmatrix} \leq \frac{1}{N_e}\sum_{n=1}^{N_e} \begin{bmatrix} Y_n - \Theta_H \\ Z_n - \Theta_H \end{bmatrix} \leq \frac{1}{N_e}\sum_{n=1}^{N_e} \begin{bmatrix} P_n^y + a_n^y \mathbb{1} \\ P_n^z + a_n^z \mathbb{1} \end{bmatrix},$$

which implies

$$\begin{bmatrix} \|\frac{1}{N_e}\sum_{n=1}^{N_e}(Y_n - \Theta_H - P_n^y)\|_\infty \\ \|\frac{1}{N_e}\sum_{n=1}^{N_e}(Z_n - \Theta_H - P_n^z)\|_\infty \end{bmatrix} \leq \frac{1}{N_e}\sum_{n=1}^{N_e} \begin{bmatrix} a_n^y \\ a_n^z \end{bmatrix}.$$

Applying the triangle inequality leads to

$$\begin{bmatrix} \|\frac{1}{N_e}\sum_{n=1}^{N_e} Y_n - \Theta_H\|_\infty \\ \|\frac{1}{N_e}\sum_{n=1}^{N_e} Z_n - \Theta_H\|_\infty \end{bmatrix} \leq \frac{1}{N_e}\sum_{n=1}^{N_e} \begin{bmatrix} a_n^y \\ a_n^z \end{bmatrix} + \frac{1}{N_e}\sum_{n=1}^{N_e} \begin{bmatrix} \|P_n^y\|_\infty \\ \|P_n^z\|_\infty \end{bmatrix}. \tag{30}$$

The lemma below establishes an upper bound on $\sum_{n=1}^{N_e}\begin{bmatrix} a_n^y \\ a_n^z \end{bmatrix}$.

Lemma 5. *The sequences $\{a_n^y\}_{n=1}^{N_e}$ and $\{a_n^z\}_{n=1}^{N_e}$ satisfy the following inequality*

$$\sum_{n=1}^{N_e}\begin{bmatrix} a_n^y \\ a_n^z \end{bmatrix} \leq (1-H)^{-1}\left(\begin{bmatrix} a_1^y \\ a_1^z \end{bmatrix} + F\sum_{n=1}^{N_e}\begin{bmatrix} \|P_n^y\|_\infty \\ \|P_n^z\|_\infty \end{bmatrix}\right). \tag{31}$$

By utilizing Lemma 5 and taking into account the fact that $a_1^z = a_1^y = \|\Theta_0 - \Theta^\star\|_\infty$, we find that

$$(1-H)^{-1}\begin{bmatrix} a_1^y \\ a_1^z \end{bmatrix} = \frac{\|\Theta_0 - \Theta^\star\|_\infty}{(1-\gamma)\lambda}\begin{bmatrix} 2-\lambda\gamma \\ (1-\lambda)\gamma+1 \end{bmatrix} \leq \frac{2\|\Theta_0 - \Theta^\star\|_\infty}{(1-\gamma)\lambda}\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Moreover, we have

$$(1-H)^{-1}F\sum_{n=1}^{N_e}\begin{bmatrix} \|P_n^y\|_\infty \\ \|P_n^z\|_\infty \end{bmatrix} = \frac{\gamma}{1-\gamma}\left((1-\lambda)\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \lambda\sum_{n=1}^{N_e}\|P_n^z\|_\infty\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Therefore, according to (31), we can conclude that

$$\frac{1}{N_e}\sum_{n=1}^{N_e}\begin{bmatrix} a_n^y \\ a_n^z \end{bmatrix} \leq \left(\frac{2\|\Theta_0 - \Theta^\star\|_\infty}{(1-\gamma)\lambda N_e} + \frac{\gamma}{1-\gamma}\Big(\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty\Big)\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{32}$$

Finally, by utilizing (30), (32) and the triangle inequality, we find that

$$\begin{aligned}\|\frac{1}{N_e}\sum_{n=1}^{N_e}Y_{n+1} - \Theta^\star\|_\infty &\leq (1-\lambda)\|\frac{1}{N_e}\sum_{n=1}^{N_e}Y_n - \Theta^\star\|_\infty + \lambda\|\frac{1}{N_e}\sum_{n=1}^{N_e}Z_n - \Theta^\star\|_\infty \\ &\leq \sum_{n=1}^{N_e}\left(\frac{1-\lambda}{N_e}(a_n^y + \|P_n^y\|_\infty) + \frac{\lambda}{N_e}(a_n^z + \|P_n^z\|_\infty)\right) \\ &\leq \frac{2\|\Theta_0 - \Theta_H\|_\infty}{(1-\gamma)\lambda N_e} + \frac{1}{1-\gamma}\left(\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty\right),\end{aligned}$$

as claimed.

A.2. **Proof of Lemma 2**

A.2.1. *Proof of Lemma 2.a*

$W_n(x,u)$ can be written as $W_n(x,u) = \bar{W}_n(x,u) + \widehat{W}_n(x,u)$, where $\bar{W}_n := \hat{r}(x,u) - r(x,u)$ and $\widehat{W}_n(x,u) := \gamma\Big(\max_{\bar{u}}\Theta^\star(x_n,\bar{u}) - \mathbb{E}_{\bar{x}}\max_{\bar{u}}\Theta^\star(\bar{x},\bar{u})\Big)$. Since $\bar{W}_n(x,u)$ is σ_r-sub-Gaussian, we have $\log(\mathbb{E}[e^{s\bar{W}_n(x,u)}]) \leq \frac{s^2\sigma_r^2}{2}$. Moreover, $\widehat{W}_n(x,u)$ is bounded in absolute value by $\gamma\|\Theta^\star\|_{\text{span}}$ and has variance $\sigma^2(\Theta^\star)(x,u)$. Hence, it satisfies Bernstein's condition (see, e.g., [BLM13; Wai19a])

$$\log(\mathbb{E}e^{s\widehat{W}(x,u)}) \leq \frac{\frac{1}{2}s^2\sigma^2(\Theta^\star)(x,u)}{1-\frac{1}{3}|s|\gamma\|\Theta^\star\|_{\text{span}}}, \quad \forall|s| \leq \frac{3}{\gamma\|\Theta^\star\|_{\text{span}}}.$$

The processes $\{P_n^z\}_{n=1}^{N_e}$ and $\{P_n^y\}_{n=1}^{N_e}$ can be expressed as $P_n^z = \bar{P}_n^z + \widehat{P}_n^z$ and $P_n^y = \bar{P}_n^y + \widehat{P}_n^y$, where

$$\begin{aligned}\bar{P}_{n+1}^z = (1-\lambda)\bar{P}_n^z + \lambda\bar{W}_n, \quad \widehat{P}_{n+1}^z = (1-\lambda)\widehat{P}_n^z + \lambda\widehat{W}_n \\ \bar{P}_{n+1}^y = (1-\lambda)\bar{P}_n^y + \lambda\bar{P}_n^z, \quad \widehat{P}_{n+1}^y = (1-\lambda)\widehat{P}_n^y + \lambda\widehat{P}_n^z\end{aligned}$$

and $\bar{P}_1^z = \widehat{P}_1^z = \bar{P}_1^y = \widehat{P}_1^y = 0$. We now provide bounds on the moment generating functions of the stochastic processes $\bar{P}_n^z$, $\widehat{P}_n^z$, $\bar{P}_n^y$, and $\widehat{P}_n^y$ through an inductive argument. For $n = 1$ we have $\log(\mathbb{E}e^{s\bar{P}_1^z(x,u)}) = \log(\mathbb{E}e^{s\widehat{P}_1^z(x,u)}) = \log(\mathbb{E}e^{s\bar{P}_1^y(x,u)}) = \log(\mathbb{E}e^{s\widehat{P}_1^y(x,u)}) = 0$. Assume that

$$\log(\mathbb{E}e^{s\bar{P}_n^z(x,u)}) \leq \frac{\lambda\sigma_r^2 s^2}{2}, \quad \log(\mathbb{E}e^{s\widehat{P}_n^z(x,u)}) \leq \frac{\frac{1}{2}s^2\lambda\sigma^2(\Theta^\star)(x,u)}{1-\frac{1}{3}|s|\lambda\gamma\|\Theta^\star\|_{\text{span}}}, \tag{33a}$$

$$\log(\mathbb{E}e^{s\bar{P}_n^y(x,u)}) \leq \frac{\lambda^2\sigma_r^2 s^2}{2}, \quad \log(\mathbb{E}e^{s\widehat{P}_n^y(x,u)}) \leq \frac{\frac{1}{2}s^2\lambda^2\sigma^2(\Theta^\star)(x,u)}{1-\frac{1}{3}|s|\lambda^2\gamma\|\Theta^\star\|_{\text{span}}}, \tag{33b}$$

then we have

$$\begin{aligned}
\log(\mathbb{E}e^{s\bar{P}_{n+1}^z(x,u)}) &\leq (1-\lambda)\frac{\lambda\sigma_r^2 s^2}{2} + \lambda\frac{\lambda\sigma_r^2 s^2}{2} = \frac{\lambda\sigma_r^2 s^2}{2},\\
\log(\mathbb{E}e^{s\bar{P}_{n+1}^y(x,u)}) &\leq (1-\lambda)\frac{\lambda^2\sigma_r^2 s^2}{2} + \lambda\frac{\lambda^2\sigma_r^2 s^2}{2} = \frac{\lambda^2\sigma_r^2 s^2}{2}\\
\log(\mathbb{E}e^{s\widehat{P}_{n+1}^z(x,u)}) &\leq (1-\lambda)\frac{\frac{1}{2}s^2\lambda\sigma^2(\Theta^\star)(x,u)}{1-\frac{1}{3}|s|\lambda\gamma\|\Theta^\star\|_{\text{span}}} + \lambda\frac{\frac{1}{2}s^2\lambda\sigma^2(\Theta^\star)(x,u)}{1-\frac{1}{3}|s|\lambda\gamma\|\Theta^\star\|_{\text{span}}} = \frac{\frac{1}{2}s^2\lambda\sigma^2(\Theta^\star)(x,u)}{1-\frac{1}{3}|s|\lambda\gamma\|\Theta^\star\|_{\text{span}}}\\
\log(\mathbb{E}e^{s\widehat{P}_{n+1}^y(x,u)}) &\leq (1-\lambda)\frac{\frac{1}{2}s^2\lambda^2\sigma^2(\Theta^\star)(x,u)}{1-\frac{1}{3}|s|\lambda^2\gamma\|\Theta^\star\|_{\text{span}}} + \lambda\frac{\frac{1}{2}s^2\lambda^2\sigma^2(\Theta^\star)(x,u)}{1-\frac{1}{3}|s|\lambda^2\gamma\|\Theta^\star\|_{\text{span}}} = \frac{\frac{1}{2}s^2\lambda^2\sigma^2(\Theta^\star)(x,u)}{1-\frac{1}{3}|s|\lambda^2\gamma\|\Theta^\star\|_{\text{span}}}
\end{aligned}$$

Hence, the inequalities (33a) and (33b) hold for all $n = 1, \ldots, N_e$. Moreover, since $e^{s\|\bar{P}_n\|_\infty} \leq \sum_{(x,u)}\left(e^{s\bar{P}_n^z(x,u)} + e^{-s\bar{P}_n^z(x,u)}\right)$, we find that

$$\mathbb{E}e^{s\|\bar{P}_n^z\|_\infty} \leq \sum_{(x,u)}\left(\mathbb{E}e^{s\bar{P}_n^z(x,u)} + \mathbb{E}e^{-s\bar{P}_n^z(x,u)}\right) \leq 2De^{\frac{\lambda\sigma_r^2 s^2}{2}}.$$

Similarly, we have

$$\mathbb{E}e^{s\|\bar{P}_n^y\|_\infty} \leq 2De^{\frac{\lambda\sigma_r^2 s^2}{2}}, \quad \mathbb{E}e^{s\|\widehat{P}_n^z\|_\infty} \leq 2D\frac{\frac{1}{2}s^2\lambda\|\sigma(\Theta^\star)\|_\infty^2}{1-\frac{1}{3}|s|\lambda\gamma\|\Theta^\star\|_{\text{span}}}, \quad \mathbb{E}e^{s\|\widehat{P}_n^y\|_\infty} \leq 2D\frac{\frac{1}{2}s^2\lambda^2\|\sigma(\Theta^\star)\|_\infty^2}{1-\frac{1}{3}|s|\lambda^2\gamma\|\Theta^\star\|_{\text{span}}}$$

Proof of bound (17a)**:** By employing the Jensen's inequality, we find that

$$\begin{aligned}
\mathbb{E}\|\bar{P}_n^z\|_\infty &\leq \frac{\lambda\sigma^2 s}{2} + \frac{\log(2D)}{s}, \quad \mathbb{E}\|\widehat{P}_n^z\|_\infty \leq \frac{\frac{1}{2}s\lambda\|\sigma(\Theta^\star)\|_\infty^2}{1-\frac{1}{3}s\lambda\gamma\|\Theta^\star\|_{\text{span}}} + \frac{\log(2D)}{s}\\
\mathbb{E}\|\bar{P}_n^y\|_\infty &\leq \frac{\lambda^2\sigma^2 s}{2} + \frac{\log(2D)}{s}, \quad \mathbb{E}\|\widehat{P}_n^y\|_\infty \leq \frac{\frac{1}{2}s\lambda^2\|\sigma(\Theta^\star)\|_\infty^2}{1-\frac{1}{3}s\lambda^2\gamma\|\Theta^\star\|_{\text{span}}} + \frac{\log(2D)}{s}.
\end{aligned}$$

Minimizing the right-hand side of the above inequalities with respect to s results in

$$\begin{aligned}
&\mathbb{E}\|\bar{P}_n^z\|_\infty \leq \sqrt{2\lambda\log(2D)}\sigma_r, \quad \mathbb{E}\|\widehat{P}_n^z\|_\infty \leq \frac{\gamma}{3}\lambda\log(2D)\|\Theta^\star\|_{\text{span}} + \sqrt{2\lambda\log(2D)}\|\sigma(\Theta^\star)\|_\infty\\
&\mathbb{E}\|\bar{P}_n^y\|_\infty \leq \sqrt{2\lambda^2\log(2D)}\sigma_r, \quad E\mathbb{E}\|\widehat{P}_n^y\|_\infty \leq \frac{\gamma}{3}\lambda^2\log(2D)\|\Theta^\star\|_{\text{span}} + \sqrt{2\lambda^2\log(2D)}\|\sigma(\Theta^\star)\|_\infty.
\end{aligned}$$

Moreover, using the triangle inequality, we find that

$$\begin{aligned}
\mathbb{E}\|P_n^z\|_\infty &\leq \mathbb{E}\|\bar{P}_n^z\|_\infty + \mathbb{E}\|\widehat{P}_n^z\|_\infty = \frac{\gamma}{3}\lambda\log(2D)\|\Theta^\star\|_{\text{span}} + \sqrt{2\lambda\log(2D)}\big(\|\sigma(\Theta^\star)\|_\infty + \sigma_r\big),\\
\mathbb{E}\|P_n^y\|_\infty &\leq \mathbb{E}\|\bar{P}_n^y\|_\infty + \mathbb{E}\|\widehat{P}_n^y\|_\infty = \frac{\gamma}{3}\lambda^2\log(2D)\|\Theta^\star\|_{\text{span}} + \sqrt{2\lambda^2\log(2D)}\big(\|\sigma(\Theta^\star)\|_\infty + \sigma_r\big).
\end{aligned}$$

Finally, we have

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^z\|_\infty \leq \gamma \underbrace{\Big((1-\lambda)\frac{\lambda^2}{3}+\frac{\lambda^2}{3}\Big)}_{\leq\frac{2}{3}\lambda^2}\log(2D)\|\Theta^\star\|_{\text{span}}$$

$$+\underbrace{\Big((1-\lambda)\lambda+\lambda\sqrt{\lambda}\Big)}_{\leq 2\lambda}\sqrt{2\log(2D)}\big(\|\sigma(\Theta^\star)\|_\infty+\sigma_r\big),$$

which establishes the claim.

Proof of bound (17b)**:** Emplying the exponential Chebyshev's inequality leads to

$$\|\bar{P}_n^z\|_\infty \leq \sqrt{2\lambda\log(\frac{2D}{\delta})}\sigma_r,\ \ \|\widehat{P}_n^z\|_\infty \leq \frac{\gamma}{3}\lambda\log(\frac{2D}{\delta})\|\Theta^\star\|_{\text{span}} + \sqrt{2\lambda\log(\frac{2D}{\delta})}\|\sigma(\Theta^\star)\|_\infty$$

$$\|\bar{P}_n^y\|_\infty \leq \sqrt{2\lambda^2\log(\frac{2D}{\delta})}\sigma_r,\ \ \|\widehat{P}_n^y\|_\infty \leq \frac{\gamma}{3}\lambda^2\log(\frac{2D}{\delta})\|\Theta^\star\|_{\text{span}} + \sqrt{2\lambda^2\log(\frac{2D}{\delta})}\|\sigma(\Theta^\star)\|_\infty.$$

with probability at least $1-\delta$. By applying the union bound, we obtain

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty \leq \frac{2\gamma}{3}\lambda^2\log(\frac{8DN_e}{\delta})\|\Theta^\star\|_{\text{span}} + 2\lambda\sqrt{2\log(\frac{8DN_e}{\delta})}\big(\|\sigma(\Theta^\star)\|_\infty+\sigma_r\big)$$

With probability at least $1-\delta$, as claimed.

A.2.2. *Proof of Lemma 2.b*

The random matrix W_n can be written as $W_n = \widehat{W}_n + W^\dagger + W^\circ$, where $\widehat{W}_n := \widehat{\mathcal{T}}_n(\Theta^\star) - \widehat{\mathcal{T}}_n(\Theta_m) + \mathcal{T}(\Theta_m) - \mathcal{T}(\Theta^\star)$, $W^\dagger := \widetilde{\mathcal{T}}(\Theta_m) - \widetilde{\mathcal{T}}(\Theta^\star) - \mathcal{T}(\Theta_m) + \mathcal{T}(\Theta^\star)$, and $W^\circ := \widetilde{\mathcal{T}}(\Theta^\star) - \mathcal{T}(\Theta^\star)$. Consequently, the stochastic processes P_n^z and P_n^y can be expressed as $P_n^z = \widehat{P}_n^z + \bar{P}_n^z$ and $P_n^y = \widehat{P}_n^y + \bar{P}_n^z$, where

$$\widehat{P}_{n+1}^y = (1-\lambda)\widehat{P}_n^y + \lambda\widehat{P}_n^z,\ \ \bar{P}_{n+1}^y = (1-\lambda)\bar{P}_n^z + \lambda\bar{P}_n^z$$

$$\widehat{P}_{n+1}^z = (1-\lambda)\widehat{P}_n^z + \lambda\widehat{W}_n,\ \ \bar{P}_{n+1}^z = (1-\lambda)\bar{P}_n^z + \lambda(W^\circ + W^\dagger)$$

with $\widehat{P}_1^y = \widehat{P}_1^z = \bar{P}_1^y = \bar{P}_1^z = 0$. Since W° and $W^\dagger$ are independent of n, it follows that

$$\bar{P}_n^z = \frac{1-(1-\lambda)^{n-1}}{1-(1-\lambda)}\lambda(W^\circ + W^\dagger) \to \|\bar{P}_n^z\|_\infty \leq \|W^\circ\|_\infty + \|W^\dagger\|_\infty.$$

Also, we have $\bar{P}_n^y = \sum_{i=1}^{n-1}(1-\lambda)^{n-1-i}\lambda\bar{P}_i^z$. As a result, by using the triangle inequality, we have

$$\|\bar{P}_n^y\|_\infty \leq \sum_{i=1}^{n-1}(1-\lambda)^{n-1-i}\lambda\|\bar{P}_i^z\|_\infty \leq \frac{1-(1-\lambda)^{n-1}}{1-(1-\lambda)}\lambda(\|W^\circ\|_\infty + \|W^\dagger\|_\infty) \leq \|W^\circ\|_\infty + \|W^\dagger\|_\infty.$$

Consequently, we find that

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty \leq \frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|\widehat{P}_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|\widehat{P}_n^z\|_\infty + \|W^\circ\|_\infty + \|W^\dagger\|_\infty. \tag{34}$$

Next, we bound each term on the right-hand side of the above inequality.

Upper bound on $\|\widehat{P}_n^y\|_\infty$ and $\|\widehat{P}_n^z\|_\infty$: The random variable $\widehat{W}(x,u)$ is bounded in absolute value by $2\gamma\|\Theta_m - \Theta^\star\|_\infty$ and its variance is at most $\gamma^2\|\Theta_m - \Theta^\star\|_\infty^2$. As a result, it satisfies Bernstein's condition

$$\log(\mathbb{E}e^{s\widehat{W}(x,u)}) \leq \frac{\frac{1}{2}s^2\gamma^2\|\Theta_m - \Theta^\star\|_\infty^2}{1-\frac{2}{3}|s|\gamma\|\Theta_m - \Theta^\star\|_\infty}, \quad \forall |s| \leq \frac{3}{2\gamma\|\Theta_m - \Theta^\star\|_\infty}.$$

By using an inductive argument similar to the proof of (17a), we find that

$$\log(\mathbb{E}e^{s\widehat{P}_n^z(x,u)}) \leq \frac{\frac{1}{2}s^2\lambda^2\gamma^2\|\Theta_m - \Theta^\star\|_\infty^2}{1-\frac{2}{3}|s|\lambda\gamma\|\Theta_m - \Theta^\star\|_\infty}, \quad \log(\mathbb{E}e^{s\widehat{P}_n^y(x,u)}) \leq \frac{\frac{1}{2}s^2\lambda^2\gamma^2\|\Theta_m - \Theta^\star\|_\infty^2}{1-\frac{2}{3}|s|\lambda^2\gamma\|\Theta_m - \Theta^\star\|_\infty}$$

for $n = 1, ..., N_e$. Again, using the same lines of arguments as in the proof of Lemma 2.a, we have

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|\widehat{P}_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|\widehat{P}_n^z\|_\infty \leq 2\gamma C(\lambda^2, 1)\|\Theta_m - \Theta^\star\|_\infty,$$
$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|\widehat{P}_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|\widehat{P}_n^z\|_\infty \leq 2\gamma C(\lambda^2, \frac{\delta}{N_e})\|\Theta_m - \Theta^\star\|_\infty$$

with probability at least $1 - 2\delta$.

Upper bound on $\|W^\circ\|_\infty$: The random variable $W^\circ(x, u)$ can be written as

$$W^\circ(x,u) = \underbrace{\sum_{i=1}^{N_e}\frac{1}{N_{\mathcal{T}}}\Big(\hat{r}_i(x,u) - r(x,u)\Big)}_{:=\bar{W}^\circ} + \underbrace{\sum_{i=1}^{N_e}\frac{\gamma}{N_{\mathcal{T}}}\Big(\max_{\bar{u}}\Theta^\star(x_i,\bar{u}) - \mathbb{E}_{\bar{x}}\max_{\bar{u}}\Theta^\star(\bar{x},\bar{u})\Big)}_{:=\widehat{W}^\circ}$$

$\bar{W}^\circ(x, u)$ is the sum of N_e i.i.d σ_r-sub Gaussian random variable and $\widehat{W}^\circ(x, u)$ is the sum of $N_{\mathcal{T}}$ independent and identically distributed random variables, where each of these variables has variance $\sigma^2(\Theta^\star)(x, u) \leq \gamma^2\|\Theta^\star\|_\infty^2$ and is bounded in absolute value by $\|\Theta^\star\|_{\text{span}} \leq 2\gamma\|\Theta^\star\|_\infty$. Hence, we have $\log\left(\mathbb{E}e^{\bar{W}^\circ(x,u)}\right) \leq \frac{\sigma_r^2 s^2}{2N_{\mathcal{T}}}$, and $\log\left(\mathbb{E}e^{\widehat{W}^\circ(x,u)}\right) \leq \frac{\frac{1}{2N_{\mathcal{T}}}s^2\gamma^2\|\Theta^\star\|_\infty^2}{1-\frac{2}{3N_{\mathcal{T}}}|s|\gamma\|\Theta^\star\|_\infty}$. Moreover, following the same lines of arguments as in the proof of Lemma 2.a, we find that

$$\mathbb{E}\|\bar{W}^\circ\|_\infty \leq \sqrt{\frac{2\log(2D)}{N_{\mathcal{T}}}}\sigma_r, \quad \mathbb{E}\|\widehat{W}^\circ\|_\infty \leq \gamma C(\frac{1}{N_{\mathcal{T}}}, 1)\|\Theta^\star\|_\infty,$$
$$\|\bar{W}^\circ\|_\infty \leq \sqrt{\frac{2\log(\frac{2D}{\delta})}{N_{\mathcal{T}}}}\sigma_r, \quad \|\widehat{W}^\circ\|_\infty \leq \gamma C(\frac{1}{N_{\mathcal{T}}}, \delta)\|\Theta^\star\|_\infty$$

with probability at least $1 - \delta$. By applying the triangle inequality and union bound, we have

$$\mathbb{E}\|W^\circ\|_\infty \leq E\|\bar{W}^\circ\|_\infty + \mathbb{E}\|\widehat{W}^\circ\|_\infty \leq \gamma C(\frac{1}{N_{\mathcal{T}}}, 1)\|\Theta^\star\|_\infty + \sqrt{\frac{2\log(2D)}{N_{\mathcal{T}}}}\sigma_r,$$
$$\|W^\circ\|_\infty \leq \|\bar{W}^\circ\|_\infty + \|\widehat{W}^\circ\|_\infty \leq \gamma C(\frac{1}{N_{\mathcal{T}}}, \delta)\|\Theta^\star\|_\infty + \sqrt{\frac{2\log(\frac{2D}{\delta})}{N_{\mathcal{T}}}}\sigma_r$$

with probability at least $1 - 2\delta$.

Upper bound on $\|W^\dagger\|_\infty$: The random variable $W^\dagger(x, u)$ is the sum of $N_{\mathcal{T}}$ i.i.d random variables. Each of these term is bounded in absolute value by $2\gamma\|\Theta_m - \Theta^\star\|_\infty$, and has a variance at most $\gamma^2\|\Theta_m - \Theta^\star\|_\infty^2$. Therefore, we have $\mathbb{E}\|W^\dagger\|_\infty \leq \gamma C(\frac{1}{N_{\mathcal{T}}}, 1)\|\Theta_m - \Theta^\star\|_\infty$, and $\|W^\dagger\|_\infty \leq \gamma C(\frac{1}{N_{\mathcal{T}}}, \delta)\|\Theta_m - \Theta^\star\|_\infty$, with probability at least $1 - \delta$.

Putting together the pieces: By substituting the above inequalities into (34) and applying the union bound, we obtain

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\mathbb{E}\|P_n^z\|_\infty \le \big(2\gamma C(\lambda^2,1) + \gamma C(\frac{1}{N_{\mathcal{T}}},1)\big)\|\Theta_m - \Theta^\star\|_\infty + \gamma C(\frac{1}{N_{\mathcal{T}}},1)\|\Theta^\star\|_\infty + \sqrt{\frac{2\log(2D)}{N_{\mathcal{T}}}}\sigma_r$$

$$\frac{1-\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^y\|_\infty + \frac{\lambda}{N_e}\sum_{n=1}^{N_e}\|P_n^z\|_\infty \le \big(2\gamma C(\lambda^2,\frac{\delta}{5N_eM}) + \gamma C(\frac{1}{N_{\mathcal{T}}},\frac{\delta}{5M})\big)\|\Theta_m - \Theta^\star\|_\infty + \gamma C(\frac{1}{N_{\mathcal{T}}},\frac{\delta}{5M})\|\Theta^\star\|_\infty + \sqrt{\frac{2\log(\frac{10MD}{\delta})}{N_{\mathcal{T}}}}\sigma_r$$

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

A.2.3. *Proof of Lemma 2.c*

We have $W_n = \widehat{H}_n(\widehat{\Theta}) - H(\widehat{\Theta}) = \widehat{\mathcal{T}}_n(\Theta^\star) - \widehat{\mathcal{T}}_n(\Theta_m) + \mathcal{T}(\Theta_m) - \mathcal{T}(\Theta^\star)$. Note that W_n is identical to the term $\widehat{W_n}$ used in the proof of Lemma2.b. Thus, the proof follows by employing the same argument as presented in the proof of Lemma 2.b.

A.3. **Proof of Lemma 3**

At first, consider the derivative of the expression $N-\alpha\log(\beta N)$ with respect to N, which is equal to $1-\frac{\alpha}{N}$. This derivative is non-negative for $N \ge \max\{\alpha,\ 2\alpha\log(\alpha\beta)\}$. Consequently, if the inequality $N \ge \alpha\log(\beta N)$ holds for $N = \max\{\alpha,\ 2\alpha\log(\alpha\beta)\}$, it also holds for $N > \max\{\alpha,\ 2\alpha\log(\alpha\beta)\}$. Therefore, our objective is to establish that $N \ge \alpha\log(\beta N)$ when $N = \max\{\alpha,\ 2\alpha\log(\alpha\beta)\}$. To this end, first assume that $\alpha \ge\ 2\alpha\log(\alpha\beta)$. In this case, we have $N = \max\{\alpha,\ 2\alpha\log(\alpha\beta)\} = \alpha$. Substituting $N=\alpha$ into $N \ge \alpha\log(\beta N)$ gives $\alpha \ge \alpha\log(\alpha\beta)$, , which is trivial since we have assumed $\alpha \ge 2\alpha\log(\alpha\beta)$. Now, consider the scenario where $2\alpha\log(\alpha\beta) \ge \alpha$. In this case, we find that $N = \max\{\alpha,\ 2\alpha\log(\alpha\beta)\} = 2\alpha\log(\alpha\beta)$. Substituting $N = 2\alpha\log(\alpha\beta)$ into $N \ge \alpha\log(\beta N)$ leads to

$$2\alpha\log(\alpha\beta) \ge \alpha\log(2\beta\alpha\log(\alpha\beta)) \ \leftrightarrow\ (\alpha\beta)^2 \ge 2\beta\alpha\log(\alpha\beta) \ \leftrightarrow\ \alpha\beta \ge 2\log(\alpha\beta).$$

The last inequality holds for all $\alpha,\beta > 0$ (noting that $x \ge 2\log(x)$ for all $x > 0$).

A.4. **Proof of Lemma 4**

We prove Lemma 4 via induction. The base case ($n=1$) is trivial. Now, assuming that (29) holds for iteration n, we will show that it also holds for iteration $n+1$. Based on the definitions of the recursions $\{Y_n\}_{n=1}^{N_e}$ and $\{Z_n\}_{n=1}^{N_e}$ in Lemma 1, we have

$$\begin{bmatrix} Y_{n+1}-\Theta_H \\ Z_{n+1}-\Theta_H \end{bmatrix} = \begin{bmatrix} (1-\lambda)(Y_n-\Theta_H)+\lambda(Z_n-\Theta_H) \\ (1-\lambda)(Z_n-\Theta_H)+\lambda\left(\widehat{\mathcal{T}}_n(Y_{n+1})-\widehat{\mathcal{T}}_n(\Theta_H)+W_n\right) \end{bmatrix}. \tag{35}$$

Since $\widehat{\mathcal{T}}_n$ is $\gamma-$contractive we have $-\gamma\|Y_{n+1}-\Theta_H\|_\infty\mathbb{1} \le -\|\widehat{\mathcal{T}}_n(Y_{n+1})-\widehat{\mathcal{T}}_n(\Theta_H)\|_\infty\mathbb{1} \le \widehat{\mathcal{T}}_n(Y_{n+1}) - \widehat{\mathcal{T}}_n(\Theta_H) \le \|\widehat{\mathcal{T}}_n(Y_{n+1})-\widehat{\mathcal{T}}_n(\Theta_H)\|_\infty\mathbb{1} \le \gamma\|Y_{n+1}-\Theta_H\|_\infty\mathbb{1}$, which leads to

$$-\gamma\Big((1-\lambda)\|Y_n-\Theta_H\|_\infty+\lambda\|Z_n-\Theta_H\|_\infty\Big)\mathbb{1} \le \widehat{\mathcal{T}}_n(Y_{n+1})-\widehat{\mathcal{T}}_n(\Theta_H) \le \gamma\Big((1-\lambda)\|Y_n-\Theta_H\|_\infty+\lambda\|Z_n-\Theta_H\|_\infty\Big)\mathbb{1}. \tag{36}$$

Also, from induction we have

$$\left[\begin{array}{c} P_n^y - a_n^y\mathbb{1} \\ P_n^z - a_n^z\mathbb{1} \end{array}\right] \leq \left[\begin{array}{c} Y_n - \Theta_H \\ Z_n - \Theta_H \end{array}\right] \leq \left[\begin{array}{c} a_n^y\mathbb{1} + P_n^y \\ a_n^z\mathbb{1} + P_n^z \end{array}\right], \tag{37}$$

which implies

$$\left[\begin{array}{c} \|Y_n - \Theta_H\|_\infty \\ \|Z_n - \Theta_H\|_\infty \end{array}\right] \leq \left[\begin{array}{c} a_n^y \\ a_n^z \end{array}\right] + \left[\begin{array}{c} \|P_n^y\|_\infty \\ \|P_n^z\|_\infty \end{array}\right]. \tag{38}$$

Substituting the above inequality into (36) gives

$$\begin{aligned} -\gamma\Big((1-\lambda)(a_n^y + \|P_n^y\|_\infty) + \lambda(a_n^z + \|P_n^z\|_\infty)\Big)\mathbb{1} &\leq \widehat{\mathcal{T}}_n(Y_{n+1}) - \widehat{\mathcal{T}}_n(\Theta^\star) \\ &\leq \gamma\,\Big((1-\lambda)(a_n^y + \|P_n^y\|_\infty) + \lambda(a_n^z + \|P_n^z\|_\infty)\Big)\mathbb{1}. \end{aligned} \tag{39}$$

By substituting (37), (38) and (39) into (35) we find that

$$\left[\begin{array}{c} \underbrace{-\Big((1-\lambda)a_n^y + \lambda a_n^z\Big)}_{-a_{n+1}^y}\mathbb{1} + \underbrace{(1-\lambda)P_n^y + \lambda P_n^z}_{P_{n+1}^y} \\ \underbrace{-\Big((1-\lambda+\lambda^2\gamma)a_n^z + \gamma(1-\lambda)\lambda a_n^y + \gamma\lambda\big((1-\lambda)\|P_n^y\|_\infty + \lambda\|P_n^z\|_\infty\big)\Big)}_{-a_{n+1}^z}\mathbb{1} + \underbrace{(1-\lambda)P_n^z + \lambda W_n}_{P_{n+1}^z} \end{array}\right]$$

$$\leq \left[\begin{array}{c} Y_{n+1} - \Theta_H \\ Z_{n+1} - \Theta_H \end{array}\right] \leq$$

$$\left[\begin{array}{c} \underbrace{\Big((1-\lambda)a_n^y + \lambda a_n^z\Big)}_{a_{n+1}^y}\mathbb{1} + \underbrace{(1-\lambda)P_n^y + \lambda P_n^z}_{P_{n+1}^y} \\ \underbrace{\Big((1-\lambda+\lambda^2\gamma)a_n^z + \gamma(1-\lambda)\lambda a_n^y + \gamma\lambda\big((1-\lambda)\|P_n^y\|_\infty + \lambda\|P_n^z\|_\infty\big)\Big)}_{a_{n+1}^z}\mathbb{1} + \underbrace{(1-\lambda)P_n^z + \lambda W_n}_{P_{n+1}^z} \end{array}\right],$$

which completes the proof.

A.5. **Proof of Lemma 5**

According to Lemma 4, we have

$$\left[\begin{array}{c} a_{n+1}^y \\ a_{n+1}^z \end{array}\right] = H\left[\begin{array}{c} a_n^y \\ a_n^z \end{array}\right] + F\left[\begin{array}{c} \|P_n^y\|_\infty \\ \|P_n^z\|_\infty \end{array}\right].$$

Using the above inequality, we find that

$$\begin{aligned} \sum_{n=1}^{N_e}\left[\begin{array}{c} a_n^y \\ a_n^z \end{array}\right] = \left[\begin{array}{c} a_1^y \\ a_1^z \end{array}\right] + \sum_{n=1}^{N_e-1}\left[\begin{array}{c} a_{n+1}^y \\ a_{n+1}^z \end{array}\right] &= \left[\begin{array}{c} a_1^y \\ a_1^z \end{array}\right] + H\sum_{n=1}^{N_e}\left[\begin{array}{c} a_n^y \\ a_n^z \end{array}\right] + F\sum_{n=1}^{N_e}\left[\begin{array}{c} \|P_n^y\|_\infty \\ \|P_n^z\|_\infty \end{array}\right] \\ &\quad - H\left[\begin{array}{c} a_{N_e}^y \\ a_{N_e}^z \end{array}\right] - F\left[\begin{array}{c} \|P_{N_e}^y\|_\infty \\ \|P_{N_e}^z\|_\infty \end{array}\right], \end{aligned}$$

which implies

$$\sum_{n=1}^{N_e}\left[\begin{array}{c} a_n^y \\ a_n^z \end{array}\right] = (1-H)^{-1}\left(\left[\begin{array}{c} a_1^y \\ a_1^z \end{array}\right] + \sum_{n=1}^{N_e}\left[\begin{array}{c} \|P_n^y\|_\infty \\ \|P_n^z\|_\infty \end{array}\right] - H\left[\begin{array}{c} a_{N_e}^y \\ a_{N_e}^z \end{array}\right] - F\left[\begin{array}{c} \|P_{N_e}^y\|_\infty \\ \|P_{N_e}^z\|_\infty \end{array}\right]\right). \tag{40}$$

Note that all the entries of the state matrix H and the input matrix F are between zero and one. Moreover, we have

$$(1-H)^{-1} = \frac{1}{\det(1-H)} \begin{bmatrix} 1-H_{22} & H_{12} \\ H_{21} & 1-H_{11} \end{bmatrix},$$

where $\det(1-H) = \lambda^2(1-\gamma) > 0$. Therefore, $(1-H)^{-1}$ Also has positive entries. Hence,

$$(1-H)^{-1}\left(H\begin{bmatrix} a^y_{N_e} \\ a^z_{N_e} \end{bmatrix} + F\begin{bmatrix} \|P^y_{N_e}\|_\infty \\ \|P^z_{N_e}\|_\infty \end{bmatrix}\right) \geq 0.$$

By adding the above expression to the right-hand side of (40), we obtain

$$\sum_{n=1}^{N_e} \begin{bmatrix} a^y_n \\ a^z_n \end{bmatrix} \leq (1-H)^{-1}\left(\begin{bmatrix} a^y_1 \\ a^z_1 \end{bmatrix} + \sum_{n=1}^{N_e} \begin{bmatrix} \|P^y_n\|_\infty \\ \|P^z_n\|_\infty \end{bmatrix}\right),$$

as claimed.

References

[AKY20] A. Agarwal, S. Kakade, and L. F. Yang. "Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal". In: *Proceedings of Thirty Third Conference on Learning Theory.* Vol. 125. 2020, pp. 67–83.

[All18] Z. Allen-Zhu. "Katyusha: The First Direct Acceleration of Stochastic Gradient Methods". In: *Journal of Machine Learning Research* 18 (2018), pp. 1–51.

[AMK13] M. G. Azar, R. Munos, and H. J. Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model". In: *Machine Learning* 91 (2013), pp. 325–346.

[AMT95] T. W. Archibald, K. I. M. McKinnon, and L. C. Thomas. "On the Generation of Markov Decision Processes". In: *Journal of the Operational Research Society* 43 (1995), pp. 125–143.

[AOM17] M. G. Azar, I. Osband, and R. Munos. "Minimax Regret Bounds for Reinforcement Learning". In: *Proceedings of the 34th International Conference on Machine Learning.* Vol. 70. 2017.

[Bel57] R. Bellman. "A Markovian Decision Process". In: *Journal of Mathematics and Mechanics* 6 (1957), pp. 679–684.

[Ber17] D. P. Bertsekas. *Dynamic Programming and Optimal Control.* 4th. Athena Scientific optimization and computation series. Athena Scientific, 2017.

[Bha+09] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. "Natural Actor-Critic Algorithms". In: *Automatica* 45 (2009), pp. 2471–2482.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, Oxford, 2013.

[BR24] J. Bhandari and D. Russo. "Model-Based Reinforcement Learning for Offline Zero-Sum Markov Games". In: *Operations research* (2024).

[BT96] D. P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming.* 1st. AthenaScientific, 1996.

[Cen+22] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. "Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization". In: *Operations Research* 70 (2022), pp. 2563–2578.

[Gha+11] M. Ghavamzadeh, H. Kappen, M. Gheshlaghi Azar, and R. Munos. “Speedy Q-Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 24. 2011.

[Jin+18] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. “Is Q-Learning Provably Efficient?” In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[JZ13] R. Johnson and T. Zhang. “Accelerating Stochastic Gradient Descent using Predictive Variance Reduction”. In: *Advances in Neural Information Processing Systems*. Vol. 26. 2013.

[Kak03] S. Kakade. “On the sample complexity of reinforcement learning”. PhD thesis. University of London, 2003.

[Kal+25] K. C. Kalagarla, D. Kartik, D. Shen, R. Jain, A. Nayyar, and P. Nuzzo. “Optimal Control of Logically Constrained Partially Observable and Multiagent Markov Decision Processes”. In: *IEEE Transactions on Automatic Control* 70.1 (2025), pp. 263–277.

[Kau+23] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza. “Champion-level drone racing using deep reinforcement learning”. In: *Nature* 620 (2023), pp. 982–987.

[Kha+21] K. Khamaru, A. Pananjady, F. Ruan, M. J. Wainwright, and M. I. Jordan. “Is Temporal Difference Learning Optimal? An Instance-Dependent Analysis”. In: *SIAM Journal on Mathematics of Data Science* 3 (2021), pp. 1013–1040.

[Kha+24] K. Khamaru, E. Xia, M. J. Wainwright, and M. I. Jordan. “Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning”. In: *IEEE Transactions on Information Theory* (2024).

[Kir+22] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. A. Sallab, S. Yogamani, and P. Pérez. “Deep Reinforcement Learning for Autonomous Driving: A Survey”. In: *IEEE Transactions on Intelligent Transportation Systems* 23 (2022), pp. 4909–4926.

[KMN02] M. Kearns, Y. Mansour, and A. Y. Ng. “A sparse sampling algorithm for near-optimal planning in large Markov decision processes”. In: *Machine Learning* 49 (2002), pp. 193–208.

[KYV22] T. Kozuno, W. Yang, and N. Vieillard et al. “KL-Entropy-Regularized RL with a Generative Model is Minimax Optimal”. In: *arXiv preprint arXiv:2205.14211* (2022).

[Lev+16] S. Levine, C. Finn, T. Darrell, and P. Abbeel. “End-to-end training of deep visuomotor policies”. In: *Journal of Machine Learning Research* 17 (2016), pp. 1–40.

[LH14] T. Lattimore and M. Hutter. “Near-optimal PAC bounds for discounted MDPs”. In: *Theoretical Computer Science* 558 (2014), pp. 125–143.

[LH19] D. Lee and J. Hu. “Primal-Dual Q-Learning Framework for LQR Design”. In: *IEEE Transactions on Automatic Control* 64.9 (2019), pp. 3756–3763.

[Li+21] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. “Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction”. In: *IEEE Transactions on Information Theory* 61 (2021), pp. 448–473.

[Li+23a] G. Li, C. Cai, Y. Chen, Y. Wei, and Y. Chi. “Is Q-Learning Minimax Optimal? A Tight Sample Complexity Analysis”. In: *Operations Research* (2023), pp. 1–15.

[Li+23b] X. Li, W. Yang, J. Liang, Z. Zhang, and M. I. Jordan. “A Statistical Analysis of Polyak-Ruppert Averaged Q-Learning”. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Vol. 206. Proceedings of Machine Learning Research. PMLR, 2023, pp. 2207–2261.

[Li+24] G. Li, Y. Wei, Y. Chi, and Y. Chen. “Breaking the Sample Size Barrier in Model-Based ReinforcementLearning with a Generative Model”. In: *Operations Research* 72 (2024), pp. 203–221.

[LS18] C. Lakshminarayanan and C. Szepesvari. “Linear Stochastic Approximation: How Far Does Constant Step-Size and Iterate Averaging Go?” In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics.* Vol. 84. PMLR, 2018, pp. 1347–1355.

[MBM16] V. Mnih, A. P. Badia, and M. Mirza et al. “Asynchronous Methods for Deep Reinforcement Learning”. In: *Proceedings of The 33rd International Conference on Machine Learning.* Vol. 48. 2016, pp. 1928–1937.

[Mey24] S. Meyn. “The Projected Bellman Equation in Reinforcement Learning”. In: *IEEE Transactions on Automatic Control* (2024).

[MKS13] V. Mnih, K. Kavukcuoglu, and D. Silver et al. “Playing Atari with Deep Reinforcement Learning”. In: *arXiv preprint arXiv:1312.5602* (2013).

[MKS15] V. Mnih, K. Kavukcuoglu, and D. Silver et al. “Human-level control through deep reinforcement learning”. In: *Nature* 518 (2015), pp. 529–533.

[Mou+23] W. Mou, K. Khamaru, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. “Optimal variance-reduced stochastic approximation in Banach spaces”. In: *arXiv preprint arXiv:2201.08518* (2023).

[MPW24] W. Mou, A. Pananjady, and M. J. Wainwright. “Optimal Oracle Inequalities for Projected Fixed-Point Equations, with Applications to Policy Evaluation”. In: *Mathematics of Operations Research* 48 (2024).

[PJ92] B. T. Polyak and A. B. Juditsky. “Acceleration of stochastic approximation by averaging”. In: *SIAM Journal on Control and Optimization* 30 (1992), pp. 838–855.

[Put14] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[PW21] A. Pananjady and M. J. Wainwright. “Instance-Dependent ℓ_∞-Bounds for Policy Evaluation in Tabular Reinforcement Learning”. In: *IEEE Transactions on Information Theory* 67 (2021), pp. 566–585.

[QW20] G. Qu and A. Wierman. “Finite-Time Analysis of Asynchronous Stochastic Approximation and Q-Learning”. In: *Proceedings of Thirty Third Conference on Learning Theory.* 2020.

[RSB12] N. L. Roux, M. Schmidt, and F. Bach. “A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets”. In: *Advances in Neural Information Processing Systems.* Vol. 25. 2012.

[SB18] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* 2nd. Cambridge, MA: MIT Press, 2018.

[Sch+15] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. “Trust Region Policy Optimization”. In: *Proceedings of the 32nd International Conference on Machine Learning.* Vol. 37. 2015, pp. 1889–1897.

[SHM16] D. Silver, A. Huang, and C. J. Maddison et al. “Mastering the game of Go with deep neural networks and tree search”. In: *Nature* 529 (2016), pp. 484–489.

[Sid+18] A. Sidford, M. Wang, X. Wu, and Y. Ye. “Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model”. In: *Advances in Neural Information Processing Systems.* Vol. 31. 2018.

[Sze09] C. Szepesvari. *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers, 2009.

[TV97] J. Tsitsiklis and B. Van Roy. "An analysis of temporal-difference learning with function approximation". In: *IEEE Transactions on Automatic Control* 42.5 (1997), pp. 674–690.

[VBC19] O. Vinyals, I. Babuschkin, and W. M. Czarnecki et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575 (2019), pp. 350–354.

[Wai19a] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019.

[Wai19b] M. J. Wainwright. "Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning". In: *arXiv preprint arXiv:1905.06265* (2019).

[Wai19c] M. J. Wainwright. "Variance-reduced Q-learning is minimax optimal". In: *arXiv preprint arXiv:1906.04697* (2019).

[Wat89] C. Watkins. "Learning from Delayed Rewards". PhD thesis. Kings College, Cambridge, England, 1989.

[WKR13] W. Wiesemann, D. Kuhn, and B. Rustem. "Robust Markov Decision Processes". In: *Mathematics of Operations Research* 38 (2013), pp. 153–183.

[YGL23] F. A. Yaghmaie, F. Gustafsson, and L. Ljung. "Linear Quadratic Control Using Model-Free Reinforcement Learning". In: *IEEE Transactions on Automatic Control* 68.2 (2023), pp. 737–752.

[YW19] L. Yang and M. Wang. "Sample-Optimal Parametric Q-Learning Using Linearly Additive Features". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6995–7004.