A Statistical Analysis of Polyak-Ruppert Averaged Q-learning

Xiang Li
lx10077@pku.edu.cn
Peking University

Wenhao Yang
yangwenhaosms@pku.edu.cn
Peking University

Jiadong Liang
jdliang@pku.edu.cn
Peking University

Zhihua Zhang
zhzhang@math.pku.edu.cn
Peking University

Michael I. Jordan
jordan@cs.berkeley.edu
UC Berkeley

Abstract

We study Q-learning with Polyak-Ruppert averaging in a discounted Markov decision process in synchronous and tabular settings. Under a Lipschitz condition, we establish a functional central limit theorem for the averaged iteration $\bar{\boldsymbol{Q}}_T$ and show that its standardized partial-sum process converges weakly to a rescaled Brownian motion. The functional central limit theorem implies a fully online inference method for reinforcement learning. Furthermore, we show that $\bar{\boldsymbol{Q}}_T$ is the regular asymptotically linear (RAL) estimator for the optimal Q-value function $\boldsymbol{Q}^*$ that has the most efficient influence function. We present a nonasymptotic analysis for the ℓ_∞ error, $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty$, showing that it matches the instance-dependent lower bound for polynomial step sizes. Similar results are provided for entropy-regularized Q-learning without the Lipschitz condition.

1 INTRODUCTION

Q-learning [Watkins, 1989], as a model-free approach seeking the optimal Q-function of a Markov decision process (MDP), is perhaps the most widely deployed algorithm in reinforcement learning (RL) [Sutton and Barto, 2018]. Unlike policy evaluation where the underlying structure is linear in nature and the goal is essentially to solve a linear system, Q-learning is nonlinear, nonsmooth and nonstationary. Theoretical analysis for Q-learning ranges from asymptotic convergence [Jaakkola et al., 1993, Tsitsiklis, 1994, Borkar and Meyn, 2000, Szepesvári et al., 1998] to nonasymptotic rates [Even-Dar et al., 2003, Beck and Srikant, 2012, Chen et al., 2020b, Li et al., 2021a, 2020b]. Variants of Q-learning [Lattimore and Hutter, 2014, Sidford et al., 2018a,b, Wainwright, 2019c] have been proposed that achieve the minimax lower bound of sample complexity established in [Azar et al., 2013].

On the other hand, Q-learning can be viewed through the lens of stochastic approximation (SA) [Konda and Tsitsiklis, 1999], a general iterative framework for solving root-finding problems [Robbins and Monro, 1951]. It is a particular instance of SA that targets the Bellman fixed-point equation, $\mathcal{T}Q^* = Q^*$, where $\mathcal{T}$ is the population Bellman operator (see Eq. (5) for the definition).

The last-iterate behavior of Q-learning has been analyzed thoroughly within the nonlinear SA framework. In particular, on the asymptotic side, the ODE approach [Kushner and Yin, 2003, Abounadi et al., 2002, Borkar, 2009, Gadat et al., 2018, Borkar et al., 2021] establishes a functional central limit theorem (functional CLT), showing that the interpolated process that connects rescaled last iterates converges weakly to the solution of a specific SDE. From the nonasymptotic side, specific nonlinear SA convergence analyses have been tailored for Q-learning, capturing its nonasymptotic convergence rate [Chen et al., 2020b, 2021, Qu and Wierman, 2020].

An important gap in this literature is the behavior of Q-learning under averaging, specifically Polyak-Ruppert averaging [Polyak and Juditsky, 1992]. Polyak-Ruppert averaging provides a general tool for stabilizing and accelerating SA algorithms. It is known to accelerate policy evaluation [Mou et al., 2020a,b] and exhibits superior empirical performance in various RL problems [Lillicrap et al., 2016, Anschel et al., 2017]. However, a theoretical understanding of Q-learning with Polyak-Ruppert averaging is not yet available.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

In this paper, we analyze averaged Q-learning in the setting of a discounted infinite-horizon MDP and in the synchronous setting where a generative model produces independent samples for all state-action pairs in every iteration [Kearns et al., 2002]. We provide both asymptotic and nonasymptotic analyses. On the asymptotic side, we establish an functional CLT for averaged Q-learning, showing that the partial-sum process, $\boldsymbol{\phi}_T(r) := \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr \rfloor}(\boldsymbol{Q}_t - \boldsymbol{Q}^*)$, converges weakly to a rescaled Brownian motion, namely $\mathrm{Var}_{\boldsymbol{Q}}^{1/2}\boldsymbol{B}_D(r)$, where $r \in [0,1]$ is the fraction of data used, $\lfloor \cdot \rfloor$ is the floor function, $\mathrm{Var}_{\boldsymbol{Q}}$ (see Eq. (10)) is the asymptotic variance, and $\boldsymbol{B}_D(\cdot)$ is a standard D-dimensional Brownian motion on $[0,1]$. Such a functional result for partial-sum processes has not been presented previously in the RL literature. This allows us to construct an asymptotically pivotal statistic using information from the whole function $\boldsymbol{\phi}_T(\cdot)$ (see Proposition 3.1). This obviates the need to estimate the asymptotic variance in providing asymptotically valid confidence intervals for $\boldsymbol{Q}^*$, which is required by [Chen et al., 2020a, Zhu et al., 2021, Hao et al., 2021, Shi et al., 2020, Khamaru et al., 2022]. It opens a door to online statistical inference for RL.

As a complementary result, we establish a semiparametric efficiency lower bound for any regular asymptotically linear (RAL) estimator (see Definition 4.2 for details) of the optimal Q-value function $\boldsymbol{Q}^*$. Given the r-th fraction of data, we further show that $\boldsymbol{\phi}_T(r)$ is the most efficient RAL estimator with the smallest asymptotic variance, confirming its optimality in the asymptotic regime.

On the nonasymptotic side, we provide the first finite-sample error analysis of $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty$ in the ℓ_∞-norm for both linearly rescaled and polynomial step sizes. The error is dominated by $\mathcal{O}(\sqrt{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln|\mathcal{S}\times\mathcal{A}|}{T}})$ for polynomial step sizes given a sufficiently large T, which matches the instance-dependent lower bound established by [Khamaru et al., 2021b]. This, together with the worst-case bound $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty = \mathcal{O}((1-\gamma)^{-3})$, implies that averaged Q-learning already achieves the optimal minimax sample complexity $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}\times\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$ established by [Azar et al., 2013]. Those lower bounds have only been shown to hold for a complicated variance-reduced version of Q-learning in this setting [Wainwright, 2019c, Khamaru et al., 2021b].

From a technical perspective, we carefully decompose the partial sum process, $\boldsymbol{\phi}_T(r)$, into several processes, each of which either has a nice structure (e.g., a sum of i.i.d. variables) or vanishes in the ℓ_∞-norm with probability one. In this way, the nonasymptotic analysis reduces to careful examination of these diminishing rates. To underpin the functional CLT, we develop a new lemma that shows that a certain residual error converges to zero in probability (see Lemma D.1). Generalizing an existing result from Lee et al. [2021], Li et al. [2022], this technical lemma may be of independent interest. Finally, while both our asymptotic and nonasymptotic analyses rely on a Lipschitz condition, stated in Assumption 3.2, we find that averaged Q-learning regularized by entropy achieves a similar functional CLT and instance-dependent bound without the Lipschitz assumption.

Paper organization. The remainder of this paper is organized as follows. In Section 2, we introduce our notation and preliminaries on RL. We present the formal functional CLT in Section 3 and the semiparametric efficiency lower bound in Section 4. In Section 5, we show the nonasymptotic convergence bound and contrast it with previous work. We summarize our results and discuss future research directions in Section 7. We provide additional discussion of related work, and all proof details, in the appendix.

2 PRELIMINARIES

Discounted infinite-horizon MDPs. An infinite-horizon MDP is represented by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, R, r)$. Here $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and $\gamma \in (0,1)$ is the discount factor. For simplicity, we define $D = |\mathcal{S}\times\mathcal{A}| = SA$. We use $P\colon \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ to represent the probability transition kernel with $P(s'|s,a)$ the probability of transiting to s' from a given state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$. Let $R\colon \mathcal{S}\times\mathcal{A} \to [0,\infty)$ stand for the random reward, i.e., $R(s,a)$ is the immediate reward collected in state $s \in \mathcal{S}$ when action $a \in \mathcal{A}$ is taken. Unlike previous works [Wainwright, 2019b, Li et al., 2021a] which assume the immediate reward R is deterministic, we consider a general setting where R itself is a random function with $r = \mathbb{E}R$ the expected reward. A policy π maps each $s \in \mathcal{S}$ to a probability over $\mathcal{A}$. In a γ-discounted MDP, a common objective is to maximize the expected long-term reward. For a given policy $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$, the expected long-term reward is measured by the Q-function Q^π defined as follows

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\middle| s_0 = s, a_0 = a\right],$$

and its companion value function is defined via $V^\pi(s) = \sum_{a\in\mathcal{A}}\pi(a|s)Q^\pi(s,a)$. Here $\mathbb{E}_\pi(\cdot)$ is taken with respect to the randomness of the trajectory of the MDP induced by the policy π. The optimal value function V^* and optimal Q-function Q^* are defined as $V^*(s) = \max_\pi V^\pi(s)$ and $Q^*(s,a) = \max_\pi Q^\pi(s,a)$. For simplicity, we employ the vectors $\boldsymbol{V}^\pi, \boldsymbol{V}^* \in \mathbb{R}^S$ and $\boldsymbol{Q}^\pi, \boldsymbol{Q}^*, \boldsymbol{Q}_t, \bar{\boldsymbol{Q}}_t \in \mathbb{R}^D$ to denote evaluations of the functions $V^\pi, V^*, Q^\pi, Q^*, Q_t, \bar{Q}_t$.

A generative model is assumed [cf. Kearns and Singh, 1999, Sidford et al., 2018a, Li et al., 2021a]. In iteration t, we collect independent samples of rewards $r_t(s,a)$ and the next state $s_t(s,a) \sim P(\cdot|s,a)$ for every state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$. We summarize the observations into the reward vector $\boldsymbol{r}_t = (r_t(s,a))_{(s,a)} \in \mathbb{R}^D$ and the empirical transition matrix $\boldsymbol{P}_t = (\boldsymbol{e}_{s_t(s,a)})_{(s,a)} \in \mathbb{R}^{D \times S}$ with each row a one-hot vector. We introduce the transition matrix $\boldsymbol{P} \in \mathbb{R}^{D \times S}$ to represent the probability transition kernel P, whose (s,a)-th row $\boldsymbol{P}_{s,a}$ is a probability vector representing $P(\cdot|s,a)$. The square probability transition matrix $\boldsymbol{P}^\pi \in \mathbb{R}^{D \times D}$ (resp. $\boldsymbol{P}_\pi \in \mathbb{R}^{S \times S}$) induced by the deterministic policy π over the state-action pairs (resp. states) is

$$\boldsymbol{P}^\pi := \boldsymbol{P}\boldsymbol{\Pi}^\pi \quad \text{and} \quad \boldsymbol{P}_\pi := \boldsymbol{\Pi}^\pi \boldsymbol{P}, \tag{1}$$

where $\boldsymbol{\Pi}^\pi \in \mathbb{R}^{S \times D}$ is a projection matrix associated with a given policy π:

$$\boldsymbol{\Pi}^\pi = \operatorname{diag}\{\pi(\cdot|1)^\top, \cdots, \pi(\cdot|S)^\top\}, \tag{2}$$

where $\pi(\cdot|s) \in \mathbb{R}^A$ is the policy vector at state s.

Q-learning. The synchronous Q-learning algorithm maintains a Q-function vector, $\boldsymbol{Q}_t \in \mathbb{R}^D$, for all $t \geq 0$ and updates its entries via the following update rule:

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r}_t + \widehat{\mathcal{T}}_t \boldsymbol{Q}_{t-1}), \tag{3}$$

where $\eta_t \in (0,1]$ is the step size in the t-th iteration and $\widehat{\mathcal{T}}_t : \mathbb{R}^D \to \mathbb{R}^D$ is the empirical Bellman operator constructed by samples collected in the t-th iteration:

$$(\widehat{\mathcal{T}}_t \boldsymbol{Q})(s,a) = r_t(s,a) + \gamma \max_{a' \in \mathcal{A}} \boldsymbol{Q}(s_t, a'), \tag{4}$$

with $r_t(s,a) \sim R(s,a)$ and $s_t = s_t(s,a) \sim P(\cdot|s,a)$ for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. In matrix form, $\widehat{\mathcal{T}}_t \boldsymbol{Q}_{t-1} = \boldsymbol{P}_t \boldsymbol{V}_{t-1}$ where $\boldsymbol{V}_{t-1}(s) = \max_a \boldsymbol{Q}_{t-1}(s,a)$ is the greedy value. Clearly, $\widehat{\mathcal{T}}_t$ is an unbiased estimate of the Bellman operator $\mathcal{T} : \mathbb{R}^D \to \mathbb{R}^D$ given by

$$(\mathcal{T}\boldsymbol{Q})(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} \boldsymbol{Q}(s', a'). \tag{5}$$

The optimal $\boldsymbol{Q}^*$ is the unique fixed point of the Bellman operator, $\mathcal{T}\boldsymbol{Q}^* = \boldsymbol{Q}^*$. Let π_t be the greedy policy w.r.t. Q_t; i.e., $\pi_t(s) \in \arg\max_{a \in \mathcal{A}} Q_t(s,a)$ for $s \in \mathcal{S}$ and π^* the optimal policy.

Averaged Q-learning. Ruppert [1988] and Polyak and Juditsky [1992] showed that averaging the iterates generated by a stochastic approximation (SA) algorithm has favorable asymptotic statistical properties. There is a line of work which has adapted Polyak-Ruppert averaging to the problem of policy evaluation in RL [Bhandari et al., 2018, Khamaru et al., 2021a, Mou et al., 2020a]. Q-learning is different than policy evaluation due to the nonstationarity (i.e., π_t changes over time) and the nonlinearity of $\mathcal{T}$. The averaged Q-learning iterate has the form

$$\bar{Q}_T = \frac{1}{T}\sum_{t=1}^{T} Q_t$$

with $\{Q_t\}_{t \geq 0}$ updated as in Eq. (3) and T is the number of iterates. When we conduct inference, we use the average estimate $\bar{Q}_T$ rather than the last iterative value Q_T given an iteration budget T. The application of Polyak-Ruppert averaging in deep RL has been shown empirically to have benefits in terms of error reduction and stability [Lillicrap et al., 2016, Anschel et al., 2017].

Bellman noise. Let $\boldsymbol{Z}_t \in \mathbb{R}^D$ be the Bellman noise at the t-th iteration, whose (s,a)-th entry is

$$Z_t(s,a) = \widehat{\mathcal{T}}_t(Q^*)(s,a) - \mathcal{T}(Q^*)(s,a). \tag{6}$$

In matrix form, the Bellman noise at iteration t can be equivalently presented as $\boldsymbol{Z}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*$. The Bellman noise $\boldsymbol{Z}_t$ reflects the noise present in the empirical Bellman operator (4) using samples collected at iteration t as an estimate of the population Bellman operator (5).

In our synchronous setting, $\boldsymbol{r}_t$ and $\boldsymbol{P}_t$ are independent of each other and the past history. Therefore, $\{\boldsymbol{Z}_t\}$ is an i.i.d. random vector sequence with coordinates that are mean zero and mutually independent. When it is clear from the context, we drop the dependence on t and use $\boldsymbol{Z}$ to denote an independent copy of $\boldsymbol{Z}_t$. We refer to $\boldsymbol{Z}$ as the Bellman noise (vector). Finally, an important quantity in our analysis is the covariance matrix of $\boldsymbol{Z}$:

$$\operatorname{Var}(\boldsymbol{Z}) = \mathbb{E}_{r_t, s_t} \boldsymbol{Z}\boldsymbol{Z}^\top \in \mathbb{R}^{D \times D}, \tag{7}$$

where the expectation $\mathbb{E}_{r_t,s_t}(\cdot)$ is taken over the randomness of rewards r_t and states s_t. Clearly, $\operatorname{Var}(\boldsymbol{Z})$ is a diagonal matrix with the (s,a)-th diagonal entry given by $\mathbb{E}Z_t^2(s,a)$.

3 FUNCTIONAL CENTRAL LIMIT THEOREM FOR PARTIAL-SUM AVERAGED Q-LEARNING

Our main result is a functional central limit theorem for the partial-sum process of averaged Q-learning. To that end, we make three assumptions. The first is that all random rewards have uniformly bounded fourth moments (Assumption 3.1). Though typical in the SA literature [Borkar, 2009], it is weaker than the uniform boundedness assumption which is often used for nonasymptotic analysis in RL. It is required for a technical reason (that we should ensure a residual error vanishes uniformly in probability, a result which is one of our technical contributions).

The second is a Lipschitz condition (Assumption 3.2) over a specific optimal policy $\pi^* \in \Pi^*$, where Π^* collects all optimal policies. The condition is true when $|\Pi^*| = 1$ (See Lemma B.1 for the reason). Similar assumptions have been adapted for asymptotic analysis for general nonlinear SA [Mokkadem and Pelletier, 2006], and nonasymptotic analysis for both variance reduced Q-learning [Khamaru et al., 2021b] and policy iteration [Puterman and Brumelle, 1979]. The condition implies that when $\boldsymbol{Q}_t \approx \boldsymbol{Q}^*$ the asymptotic behavior of averaged Q-learning is captured by a linear system up to a high-order approximation error. As a result, we can explicitly formulate the asymptotic variance matrix. The approach of approximating a nonlinear SA by a specific linear SA and analyzing the approximation errors is also standard in the SA literature [Polyak and Juditsky, 1992, Mokkadem and Pelletier, 2006, Lee et al., 2021, Li et al., 2022].

The last assumption (Assumption 3.3) requires that the step size decays at a sufficiently slow rate; this is necessary in order to establish asymptotic normality [Polyak and Juditsky, 1992, Su and Zhu, 2018, Chen et al., 2020a, Li et al., 2022]. A typical example satisfying Assumption 3.3 is the polynomial step size, $\eta_t = t^{-\alpha}$ with $\alpha \in (0.5, 1)$.

Assumption 3.1. *We assume* $\mathbb{E}|R(s,a)|^4 < \infty$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Assumption 3.2. *There exists* $\pi^* \in \Pi^*$ *such that for any Q-function estimator* $\boldsymbol{Q} \in \mathbb{R}^D$, $\|(\boldsymbol{P}^{\pi_{\boldsymbol{Q}}} - \boldsymbol{P}^{\pi^*})(\boldsymbol{Q} - \boldsymbol{Q}^*)\|_\infty \leq L\|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty^2$ *where* $\pi_Q(s) := \arg\max_{a \in \mathcal{A}} Q(s,a)$ *is the greedy policy w.r.t.* Q.

Assumption 3.3. *Assume* (i) $0 \leq \sup_t \eta_t \leq 1, \eta_t \downarrow 0$ *and* $t\eta_t \uparrow \infty$*;* (ii) $\frac{\eta_{t-1} - \eta_t}{\eta_{t-1}} = o(\eta_{t-1})$*;* (iii) $\frac{1}{\sqrt{T}} \sum_{t=0}^{T} \eta_t \to 0$ *for all* $t \geq 1$*;* (iv) $\frac{\sum_{t=0}^{T} \eta_t}{T\eta_T} \leq C$ *for all* $T \geq 1$.

We now present the functional CLT for averaged Q-learning under the same conditions. Define the standardized partial-sum processes associated with $\{\boldsymbol{Q}_t\}_{t \geq 0}$ as follows:

$$\boldsymbol{\phi}_T(r) := \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} (\boldsymbol{Q}_t - \boldsymbol{Q}^*), \tag{8}$$

where $r \in [0,1]$ is the fraction of the data used to compute the partial-sum process and $\lfloor \cdot \rfloor$ returns the largest integer smaller than or equal to the input number.

Theorem 3.1. *Under Assumptions 3.1, 3.2 and 3.3, we have*

$$\boldsymbol{\phi}_T(\cdot) \stackrel{w}{\to} \mathrm{Var}_{\boldsymbol{Q}}^{1/2} \boldsymbol{B}_D(\cdot), \tag{9}$$

where $\mathrm{Var}_{\boldsymbol{Q}} \in \mathbb{R}^{D \times D}$ *is the asymptotic variance*

$$\mathrm{Var}_{\boldsymbol{Q}} = (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1} \mathrm{Var}(\boldsymbol{Z}) (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-\top} \tag{10}$$

and $\boldsymbol{B}_D(\cdot) \in \mathbb{R}^D$ *is a standard Brownian motion on* $[0,1]$.

The conventional CLT asserts that $\boldsymbol{\phi}_T(1) = \sqrt{T}(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*)$ converges in distribution to a rescaled Gaussian random variable $\mathrm{Var}_{\boldsymbol{Q}}^{1/2} \boldsymbol{B}_D(1)$ as $T \to \infty$ (see Appendix B for more details). The functional CLT in Theorem 3.1 extends this convergence to the whole function $\boldsymbol{\phi}_T = \{\boldsymbol{\phi}_T(r)\}_{r \in [0,1]}$ in the sense that any finite-dimensional projections of $\boldsymbol{\phi}_T$ converge in distribution. That is, for any given integer $n \geq 1$ and any $0 \leq t_1 < \cdots < t_n \leq 1$, as $T \to \infty$, $(\boldsymbol{\phi}_T(t_1), \cdots, \boldsymbol{\phi}_T(t_n)) \stackrel{d}{\to} \mathrm{Var}_{\boldsymbol{Q}}^{1/2}(\boldsymbol{B}_D(t_1), \cdots, \boldsymbol{B}_D(t_n))$. The convergence $\stackrel{w}{\to}$ in (9) also corresponds to the weak convergence of measures in the D-dimensional Skorokhod spaces $\mathsf{D}([0,1], \mathbb{R}^D)$ (see Appendix C.1.1 for a short introduction). Here $\mathsf{D}([0,1], \mathbb{R}^D) = \{$right continuous with left limits $\omega(r) \in \mathbb{R}^D, r \in [0,1]\}$. Eq. (9) is equivalent to the convergence of finite-dimensional projections.

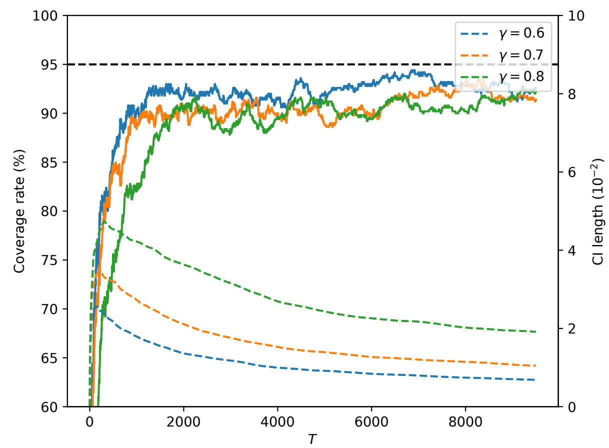

Figure 1: Empirical coverage rates (left) and CI lengths (right) of $\bar{Q}_T(s_0, a_0)$ against the number of iterations T on a specific (s_0, a_0). Both are obtained by averaging over 500 independent Q-learning trajectories. Black dashed line denotes the nominal coverage rate of 95%.

Theorem 3.1 can be viewed as a generalization of Donsker's theorem [Donsker, 1951] to Q-learning iterates. Donsker's theorem shows the partial-sum process of a sequence of independent and identically distributed (i.i.d.) random variables weakly converges to a standard Brownian motion, while subsequent works extend this functional result to weakly dependent stationary sequences [Dudley, 2014]. Since in our case π_t and $\boldsymbol{V}_t$ might depend on history data arbitrarily, $\{\boldsymbol{Q}_t\}_{t \geq 0}$ is neither i.i.d. nor stationary. To prove the functional CLT, we use a particular error decomposition and partial-sum decomposition. We give a proof sketch in Section 3.2.

Comparison with previous (functional) CLTs. Most CLT results consider linear SA which is non-applicable here (see Mou et al. [2020a,b] and references therein). The original result for Polyak-Ruppert averaging [Polyak and Juditsky, 1992, Moulines and Bach, 2011, Durmus et al., 2022] also doesn't apply in our case because it assumes a locally strongly convex Lyapunov function—which is not known to exist for Q-learning. Konda and Tsitsiklis [1999] shows $\frac{\boldsymbol{Q}_T - \boldsymbol{Q}^*}{\sqrt{\eta_T}} \stackrel{d}{\to} \mathcal{N}(\boldsymbol{0}, \mathrm{Var})$ with $\mathrm{Var} = \frac{\lim_T}{\eta_T} \mathbb{E}(\boldsymbol{Q}_T - \boldsymbol{Q}^*)(\boldsymbol{Q}_T - \boldsymbol{Q}^*)^\top$ when we assume the limit involved exists. Mokkadem and Pelletier [2006] shows $\boldsymbol{\phi}_T(1) \stackrel{d}{\to} \mathcal{N}(\boldsymbol{0}, \mathrm{Var}_{\boldsymbol{Q}})$ under a similar Lipschitz condition Assumption 3.2.

To date, formal functional CLT results for SA are mainly based on the ODE approach [Abounadi et al., 2002, Borkar, 2009, Gadat et al., 2018, Borkar et al., 2021]. These works focus on the asymptotic behavior of the interpolated process connecting properly rescaled last iterates. An example interpolated process $\widetilde{\phi}_T(\cdot)$ satisfies $\widetilde{\phi}_T(0) = \frac{Q_T - Q^*}{\sqrt{\eta_T}}$ and $\widetilde{\phi}_T(t_k^T) = \frac{Q_{T+k} - Q^*}{\sqrt{\eta_{T+k}}}$ for a specific sequence $\{t_k^T\}_{k\geq 0}$ depending on the step size and satisfying $t_0^T = 0$ and $\lim_k t_k^T = \infty$. This functional CLT result implies $\widetilde{\phi}_T(\cdot)$ converges weakly to the solution of a specific SDE. Theorem 3.1 is different because it is concerned with the partial-sum process $\phi_T(\cdot)$ and explicitly formulates the asymptotic variance Var_{Q}. Recent work studying statistical inference via SGD variants also provides functional CLTs for a similar partial-sum process [Lee et al., 2021, Li et al., 2022], given the loss function is smooth and strongly convex. However, those results don't apply here since Q-learning doesn't meet the underlying assumptions. Our functional CLT for the partial-sum process of Q-learning is novel.

3.1 Online Statistical Inference

The functional CLT opens a path towards statistical inference in RL. While traditional approaches estimate asymptotic variances in RL by batch-mean estimators [Chen et al., 2020a, Zhu et al., 2021] or bootstrapping [Hao et al., 2021], by contrast, the functional CLT allows us to construct an asymptotically pivotal statistic using the whole function ϕ_T. The inference method, known as random scaling, was originally designed for strongly convex optimization [Lee et al., 2021, Li et al., 2022].

Proposition 3.1. *The continuous mapping theorem together with Theorem 3.1 yields that with probability approaching one, $\int_0^1 \phi_T(r)\phi_T(r)^\top dr$ is invertible and*

$$\begin{aligned}&\phi_T(1)^\top \left(\int_0^1 \bar{\phi}_T(r)\bar{\phi}_T(r)^\top \mathrm{d}r\right)^{-1} \phi_T(1)\\ &\overset{d}{\to} B_D(1)^\top \left(\int_0^1 \bar{B}_D(r)\bar{B}_D(r)^\top \mathrm{d}r\right)^{-1} B_D(1),\end{aligned} \tag{11}$$

where $\bar{\phi}_T(r) := \phi_T(r) - r\cdot\phi_T(1)$ and $\bar{B}_D(r) := B_D(r) - r\cdot B_D(1)$ for simplicity.

The left-hand side of (11) is a pivotal quantity involving samples and the unobservable parameter of interest Q^*. The pivotal quantity can be constructed in a fully online fashion and thus is computationally efficient.[1] The right-hand side of (11) is a known distribution whose quantiles can be computed via simulation [Kiefer et al., 2000, Abadir and Paruolo, 2002]. In this way, we don't need a consistent estimator for the asymptotic variance in order to provide asymptotically valid confidence intervals for Q^*, as are required by previous work [Hao et al., 2021, Shi et al., 2020, Khamaru et al., 2022]. As an illustration, Figure 1 shows the empirical coverage rates and confidence interval (CI) lengths on a random MDP with three values of γ. As T increases, the empirical coverage rates increase rapidly, approaching 95%, and the CI lengths decay. More details are placed in Appendix J.

[1] See Algorithm 1 in [Lee et al., 2021] or Algorithm 2 in [Li et al., 2022] for the online procedure.

3.2 Proof Sketch

In the part, we provide a proof sketch of Theorem 3.1 to highlight our technical contributions. A full proof of Theorem 3.1 is provided in Appendix C.

Step 1: Error decomposition. Let $\Delta_t = Q_t - Q^*$. Recall that the Q-learning update rule is (3). It follows that

$$\begin{aligned}\Delta_t &= (1-\eta_t)\Delta_{t-1} + \eta_t\left[(r_t - r) + \gamma(P_t V_{t-1} - PV^*)\right]\\ &= (1-\eta_t)\Delta_{t-1} + \eta_t\left[Z_t + \gamma P_t(V_{t-1} - V^*)\right],\end{aligned}$$

where $Z_t = (r_t - r) + \gamma(P_t - P)V^*$ is the Bellman noise. Notice that $P_t(V_{t-1} - V^*) = (P_t - P)(V_{t-1} - V^*) + P(V_{t-1} - V^*)$. Using $PV_{t-1} = P^{\pi_{t-1}}Q_{t-1}$ and $PV^* = P^{\pi^*}Q^*$, we further have $P(V_{t-1} - V^*) = P^{\pi_{t-1}}Q_{t-1} - P^{\pi^*}Q^* = (P^{\pi_{t-1}} - P^{\pi^*})Q_{t-1} + P^{\pi^*}\Delta_{t-1}$. Putting the pieces together,

$$\Delta_t = A_t\Delta_{t-1} + \eta_t\left[Z_t + \gamma Z_t' + \gamma Z_t''\right],$$

where $A_t = I - \eta_t G, G = I - \gamma P^{\pi^*}, Z_t' = (P_t - P)(V_{t-1} - V^*)$, and $Z_t'' = (P^{\pi_{t-1}} - P^{\pi^*})Q_{t-1}$. Recursing the last equality gives

$$\Delta_t = \prod_{j=1}^t A_j\Delta_0 + \sum_{j=1}^t\prod_{i=j+1}^t A_i\eta_j\left(Z_j + \gamma Z_t' + \gamma Z_t''\right). \tag{12}$$

In addition, using the general step size in Assumption 3.3, we can show $\frac{1}{\sqrt{T}}\sum_{t=1}^T \mathbb{E}\|\Delta_t\|_\infty^2 \to 0$ (in Theorem E.1).

Step 2: Partial-sum decomposition. For simplicity, for any $T \geq j \geq 0$ we denote

$$A_j^T = \eta_j\sum_{t=j}^T\prod_{i=j+1}^t A_i. \tag{13}$$

Setting $\psi_0(r) := \frac{1}{\eta_0\sqrt{T}}(\boldsymbol{A}_0^{\lfloor Tr\rfloor} - \eta_0 \boldsymbol{I})\boldsymbol{\Delta}_0$ and plugging (12) into $\boldsymbol{\phi}_T(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{\Delta}_t$, yields

$$
\begin{aligned}
&\boldsymbol{\phi}_T(r) = \boldsymbol{\psi}_0(r) + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^{\lfloor Tr\rfloor}\left(\boldsymbol{Z}_j + \gamma\boldsymbol{Z}_j' + \gamma\boldsymbol{Z}_j''\right) \\
&= \boldsymbol{\psi}_0(r) + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{G}^{-1}\boldsymbol{Z}_j + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j \\
&+ \frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^T\boldsymbol{Z}_j' + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}(\boldsymbol{A}_j^{\lfloor Tr\rfloor} - \boldsymbol{A}_j^T)\left[\boldsymbol{Z}_j + \gamma\boldsymbol{Z}_j'\right] \\
&+ \frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^T\boldsymbol{Z}_j'' := \sum_{i=0}^{5}\boldsymbol{\psi}_i(r). \qquad (14)
\end{aligned}
$$

Step 3: Establish the functional CLT. To measure the distance between random functions, we define $\|\boldsymbol{\psi}\|_{\sup} = \sup_{r\in[0,1]}\|\boldsymbol{\psi}(r)\|_\infty$. The standard martingale functional CLT [Hall and Heyde, 2014, Jirak, 2017] implies $\boldsymbol{\psi}_1(\cdot) \overset{w}{\to} \mathrm{Var}_{\boldsymbol{Q}}^{1/2}\boldsymbol{B}_D(\cdot)$. To complete the proof, it suffice to show $\|\boldsymbol{\phi}_T - \boldsymbol{\psi}_1\|_{\sup} = o_{\mathbb{P}}(1)$ which is implied by $\|\boldsymbol{\psi}_i\|_{\sup} = o_{\mathbb{P}}(1)$ for $i = 0, 2, 3, 4, 5$.

By Lemma 1 in [Polyak and Juditsky, 1992], we know $\sup_{T\geq j\geq 0}\|\boldsymbol{A}_j^T\|_\infty \leq C_0$ and $\lim_{T\to\infty}\frac{1}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_2 = 0$. Then it is obvious $\|\boldsymbol{\psi}_0\|_{\sup} = o_{\mathbb{P}}(1)$. Noting that $\boldsymbol{Z}_j, \boldsymbol{Z}_j'$ are martingale differences, we can show $\mathbb{E}\|\boldsymbol{\psi}_i\|_{\sup}^2 = o(1)$ for $i = 2, 3$ by Doob's inequality.

By definition of greedy policies π^* and π_{t-1}, we know $\boldsymbol{P}^{\pi^*}\boldsymbol{Q}_{t-1} \leq \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1}$ and $\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}^* \leq \boldsymbol{P}^{\pi^*}\boldsymbol{Q}^*$, which implies $\|\boldsymbol{Z}_t''\|_\infty = \|(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{Q}_{t-1}\|_\infty \leq \|(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{t-1}\|_\infty \leq L\|\boldsymbol{\Delta}_{t-1}\|_\infty^2$ from Assumption 3.2. Then $\mathbb{E}\|\boldsymbol{\psi}_5\|_{\sup} \leq \frac{LC_0}{\sqrt{T}}\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \to 0$.

The most challenging step is to show $\|\boldsymbol{\psi}_4\|_{\sup} = o_{\mathbb{P}}(1)$. Notice that $\boldsymbol{\psi}_4$ is a weighted sum of martingale differences, $\boldsymbol{Z}_j + \gamma\boldsymbol{Z}_j'$, with the coefficients varying in r such that we can't apply Doob's inequality. To deal with this issue, we relate $\boldsymbol{\psi}_4$ to an autoregressive sequence indexed by $k \in [T]$ and analyze the maximum over k directly. More specifically, we can show

$$
\|\boldsymbol{\psi}_4\|_{\sup} \precsim \sup_{k\in[T]}\left\|\frac{1}{\sqrt{T}\eta_{k+1}}\sum_{j=1}^{k}\prod_{i=j+1}^{k}\boldsymbol{A}_i\eta_j(\boldsymbol{Z}_j + \gamma\boldsymbol{Z}_j')\right\|.
$$

Previous results Lee et al. [2021], Li et al. [2022] do not apply here, since they require $\boldsymbol{G} = \boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*}$ to be positive semidefinite, which isn't our case. Noticing that all eigenvalues of $\boldsymbol{G}$ have nonnegative real parts, we provide a novel analysis of the right-hand side in Lemma D.1, showing it is indeed $o_{\mathbb{P}}(1)$ under Assumption 3.1. This is one of our technical contributions.

Remark 3.1. *If we consider policy evaluation (so that π_t remains unchanged and $\boldsymbol{\psi}_5$ disappears), $\boldsymbol{\psi}_4$ is still present. Showing $\|\boldsymbol{\psi}_4\|_{\sup} = o_{\mathbb{P}}(1)$ is required even for linear SA.*

4 INFORMATION-THEORETIC LOWER BOUND

The standard CLT implies $\bar{\boldsymbol{Q}}_T$ is a $\sqrt{T}$-consistent estimate for $\boldsymbol{Q}^*$. It is of theoretical interest to investigate whether or not $\bar{\boldsymbol{Q}}_T$ is asymptotically efficient. In parametric statistics [Lehmann and Casella, 2006], the Cramer-Rao lower bound assesses the hardness of estimating a target parameter $\beta(\theta)$ in a parametric model $\mathcal{P}_\theta$ indexed by parameter θ. Any unbiased estimator whose variance achieves the Cramer-Rao lower bound is viewed as optimal and efficient. The concept of Cramer-Rao lower bounds can be extended to possibly biased but asymptotically unbiased estimators and also to nonparametric statistical models where the dimension of the parameter θ is infinity [Van der Vaart, 2000, Tsiatis, 2006].

The semiparametric model. In our case, the transition kernel $\{P(\cdot|s,a)\}_{s,a}$ is specified by D parametric distributions on $\mathcal{D}$, while the random reward $\{R(s,a)\}_{s,a}$ is fully nonparametric because the $R(s,a)$ are not assumed to come from finite-dimensional models. Hence, to derive an extended Cramer-Rao lower bound for Q^* estimation, we need to enter the world of semiparametric statistics. In particular, our MDP model $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, R, r)$ has parameter $\theta = (P, R)$. Our parameter of interest is $\beta(\theta) = \boldsymbol{Q}^*$. At iteration t, we observe the random rewards and empirical transitions for each (s,a) and concatenate them into $\boldsymbol{r}_t \in \mathbb{R}^D$ and $\boldsymbol{P}_t \in \mathbb{R}^{D\times S}$. The distribution of $\boldsymbol{P}_t$ is determined by its expectation $\boldsymbol{P} = \mathbb{E}\boldsymbol{P}_t$, which belongs to

$$
\begin{aligned}
\mathcal{P}_P := \Big\{\boldsymbol{P} \in \mathbb{R}^{D\times S} : P(s'|s,a) \geq 0, \forall(s,a,s') \\
\text{and} \sum_{s'\in\mathcal{S}} P(s'|s,a) = 1, \forall(s,a)\Big\}, \qquad (15)
\end{aligned}
$$

while R is nonparametric and belongs to

$$
\mathcal{P}_R = \{\{R(s,a)\}_{s,a} : \mathbb{E}R(s,a) = r(s,a), \forall(s,a)\}.
$$

According to the generative model, the $\boldsymbol{r}_t$ and $\boldsymbol{P}_t$ are mutually independent and also independent of the historical data. Let $\mathcal{D} = \{(\boldsymbol{r}_t, \boldsymbol{P}_t)\}_{t\in[T]}$ contain the T samples generated as described above.

Semiparametric efficiency lower bound. Tsiatis [2006] has argued that regular asymptotically linear (RAL) estimators provide a good tradeoff between expressivity and tractability. In RL, RAL estimators are widely considered in off-policy evaluation problems [Kallus and Uehara, 2020].

Definition 4.1 (Regular estimator)**.** *Denote the distribution of $\boldsymbol{r}_t$ and $\boldsymbol{P}_t$ by $\mathcal{L}(\boldsymbol{r})$ and $\mathcal{L}(\boldsymbol{P})$.*[2] *For any given T, let*

[2]Given a probability space $(\Omega, P, \mathcal{F})$, $\mathcal{L}(X)$ is the law of the random variable X in this probability space. Since $\boldsymbol{r}_t$ are i.i.d., they share the same distribution $\mathcal{L}(\boldsymbol{r})$ and similarly for $\mathcal{L}(\boldsymbol{P})$.

$\mathcal{L}_T(\boldsymbol{r})$ and $\mathcal{L}_T(\boldsymbol{P})$ be the perturbed distributions of $\mathcal{L}(\boldsymbol{r})$ and $\mathcal{L}(\boldsymbol{P})$ which are consistent in the sense that they converge[3] *to $\mathcal{L}(\boldsymbol{r})$ and $\mathcal{L}(\boldsymbol{P})$ when T goes infinity. Let $\widehat{\boldsymbol{Q}}_T$ be any estimator of $\boldsymbol{Q}^*$ computed from $\mathcal{D}$. Let $\boldsymbol{Q}_T^*$ be the true optimal Q-value function when rewards and transition probabilities are generated i.i.d. from $\mathcal{L}_T(\boldsymbol{r})$ and $\mathcal{L}_T(\boldsymbol{P})$. We say $\widehat{\boldsymbol{Q}}_T$ is a* regular estimator *of $\boldsymbol{Q}^*$ if $\sqrt{T}(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}_T^*)$ weakly converges to a random variable that depends only on $\mathcal{L}(\boldsymbol{r})$ and $\mathcal{L}(\boldsymbol{P})$, when samples are distributed according to the probability measure $(\mathcal{L}_T(\boldsymbol{r}), \mathcal{L}_T(\boldsymbol{P}))$.*

Remark 4.1. *Informally speaking, an estimator is regular if its limiting distribution is unaffected by local changes in the data-generating process. The assumption of regularity excludes super-efficient estimators, whose asymptotic variance can be smaller than the Cramer-Rao lower bound for some parameter values, but which perform poorly in the neighborhood of points of super-efficiency. We refer interested readers to Section 3.1 in [Tsiatis, 2006] for a detailed exposition.*

Definition 4.2 (Regular asymptotically linear). *Let $\widehat{\boldsymbol{Q}}_T \in \mathbb{R}^D$ be a measurable random function of $\mathcal{D} = \{(\boldsymbol{r}_t, \boldsymbol{P}_t)\}_{t\in[T]}$. We say that $\widehat{\boldsymbol{Q}}_T$ is* regular asymptotically linear *(RAL) for $\boldsymbol{Q}^*$ if it is regular and asymptotically linear with a measurable random function $\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t) \in \mathbb{R}^D$ such that*

$$\sqrt{T}(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t) + o_{\mathbb{P}}(1).$$

Here $\boldsymbol{\phi}(\cdot,\cdot)$ is referred to as an influence function, *and it satisfies $\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t) = \mathbf{0}$ and $\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)^\top$.*

Theorem 4.1. *Given the dataset $\mathcal{D} = \{(\boldsymbol{r}_t, \boldsymbol{P}_t)\}_{t\in[T]}$, for any RAL estimator $\widehat{\boldsymbol{Q}}_T$ of $\boldsymbol{Q}^*$ computed from $\mathcal{D} = \{(\boldsymbol{r}_t, \boldsymbol{P}_t)\}_{t\in[T]}$, its variance satisfies*

$$\lim_{T\to\infty} T\mathbb{E}(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}^*)(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}^*)^\top \succeq \mathrm{Var}_{\boldsymbol{Q}},$$

where $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite and $\mathrm{Var}_{\boldsymbol{Q}}$ is given in (10).

By Definition 4.2, any influence function determines an asymptotic linear estimator for $\boldsymbol{Q}^*$. The semiparametric efficiency bound in Theorem 4.1 gives us a concrete target in the construction of the influence function. If we can find an influence function that achieves the bound, we know that it is the most efficient among all RAL estimators. Fortunately, Theorem 4.2 implies that $\bar{\boldsymbol{Q}}_T$ is the most efficient estimator among all RAL estimators with the efficient influence function $(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-1}\boldsymbol{Z}_t$. It also implies that for any fixed $r \in [0,1]$, $\boldsymbol{\phi}_T(r) = \sqrt{r}\cdot\sqrt{\lfloor Tr \rfloor}(\bar{\boldsymbol{Q}}_{\lfloor Tr \rfloor} - \boldsymbol{Q}^*)$ has the optimal asymptotic variance (scaled by a factor $\sqrt{r}$). Proofs are provided in Appendix G.

[3] $\mathcal{L}_T(\boldsymbol{r})$ and $\mathcal{L}_T(\boldsymbol{P})$ are differentiable in quadratic mean at $\mathcal{L}(\boldsymbol{r})$ and $\mathcal{L}(\boldsymbol{P})$. See Chapter 25.3 in Van der Vaart [2000].

Theorem 4.2. *Under Assumptions 3.1, 3.2 and 3.3, the averaged Q-learning iterate $\bar{\boldsymbol{Q}}_T$ is a RAL estimator for $\boldsymbol{Q}^*$. In particular, we have the following decomposition*

$$\sqrt{T}\,(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-1}\boldsymbol{Z}_t + o_{\mathbb{P}}(1),$$

where $\boldsymbol{Z}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*$ is the Bellman noise at iteration t.

5 INSTANCE-DEPENDENT NONASYMPTOTIC CONVERGENCE

In the section, we explore the nonasymptotic behavior of averaged Q-learning, i.e., we study the dependence of $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty$ on finite T and $(1-\gamma)^{-1}$.

Theorem 5.1. *Let Assumptions 3.2 hold and $0 \le R(s,a) \le 1$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.*[4] *When D is larger than a universal constant,*

- *If $\eta_t = t^{-\alpha}$ with $\alpha \in (0.5, 1)$ for $t \ge 1$ and $\eta_0 = 1$, it follows that for all $T \ge 1$, $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty =$*

$$\mathcal{O}\left(\sqrt{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}} + \frac{\sqrt{\ln D}}{(1-\gamma)^3}\frac{1}{T^{1-\frac{\alpha}{2}}}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^{3+\frac{2}{1-\alpha}}}\frac{1}{T} + \frac{\gamma L}{(1-\gamma)^{4+\frac{1}{1-\alpha}}}\frac{1}{T^\alpha}\right).$$

- *If $\eta_t = \frac{1}{1+(1-\gamma)t}$, it follows that for all $T \ge 1$, $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty =$*

$$\mathcal{O}\left(\sqrt{\frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}}\sqrt{\frac{\ln D}{T}}\right) + \widetilde{\mathcal{O}}\left(\frac{L}{(1-\gamma)^6}\frac{1}{T}\right).$$

Here $\widetilde{\mathcal{O}}(\cdot)$ hides polynomial dependence on α, L and logarithmic factors (i.e., $\ln D$ and $\ln T$).

Instance-dependent behavior. For the polynomial step size, Theorem 5.1 shows that the instance-dependent term $\mathcal{O}(\sqrt{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}})$ dominates the ℓ_∞ error, which matches the instance-dependent lower bound established by Khamaru et al. [2021b] given a sufficiently large T. To the best of our knowledge, this is the first finite-sample analysis of averaged Q-learning in the ℓ_∞-norm showing instance-dependent optimality. However, for the linearly

[4] To simplify the parameter dependence, we assume rewards are uniformly bounded as in previous work [Wainwright, 2019c, Khamaru et al., 2021b, Li et al., 2021a] . Note that, thanks to the error decomposition in (14), it is possible to provide a nonasymptotic analysis assuming rewards have finite second moments. The consequence is that the dependence on d and δ would change from $\log D, \log\frac{1}{\delta}$ to D and $\frac{1}{\delta}$.

rescaled step size, we see that $\mathcal{O}\left(\sqrt{\frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}}\sqrt{\frac{\ln D}{T}}\right)$ is the dominant factor, which is larger because we have

$$\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty \overset{(a)}{\le} \|(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}\|_\infty^2\|\mathrm{Var}(\boldsymbol{Z})\|_\infty \overset{(b)}{\le} \frac{1}{(1-\gamma)^2}\|\mathrm{Var}(\boldsymbol{Z})\|_\infty,$$

where (a) uses $\|\mathrm{diag}(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^\top)\|_\infty \le \|\boldsymbol{V}\|_\infty\|\boldsymbol{A}\|_\infty^2$ for any diagonal matrix $\boldsymbol{V}$ (see Lemma F.2) and (b) uses $\|(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}\|_\infty \le (1-\gamma)^{-1}$. Hence, the linearly rescaled step size doesn't match the instance-dependent lower bound. It might be true because the linearly rescaled step size doesn't satisfy Assumption 3.3, implying that (22) does not necessarily hold for it.

Comparison with variance-reduced Q-learning. Under the same assumptions, Khamaru et al. [2021b] analyzed a variance-reduced variant of Q-learning that also achieves instance-dependent optimality with the following guarantee:

$$\mathbb{E}\|\widehat{\boldsymbol{Q}}_T-\boldsymbol{Q}^*\|_\infty = \mathcal{O}\left(\sqrt{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}}\right)+\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^2}\frac{1}{T}\right),$$

which has a better nonleading term than averaged Q-learning. This might somewhat explain the finding of Khamaru et al. [2021a] that averaging can be sub-optimal in the nonasymptotic regime with limited samples. However, the dominant terms are equal, implying that averaging is still powerful and efficient in the asymptotic regime. Instance-dependent convergence with a variance structure in the dominant term has also been found for other settings; please see Appendix A.

Worst-case behavior. The instance-dependent bound provides more information about the convergence rate. Previous works [Azar et al., 2013, Li et al., 2020a] imply the worst-case bound $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty = \mathcal{O}((1-\gamma)^{-3})$. Such a dependence on $(1-\gamma)^{-1}$ is tight, because Khamaru et al. [2021b] constructs a family of MDPs parameterized by $\lambda \ge 0$ where $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty = \Theta((1-\gamma)^{-3+\lambda})$. When plugging in the worst-case bound, we find that for polynomial step sizes and for sufficiently small ε, averaged Q-learning already achieves the optimal minimax sample complexity $\widetilde{\mathcal{O}}\left(\frac{D}{(1-\gamma)^3\varepsilon^2}\right)$ established by Azar et al. [2013]. Wainwright [2019c] uses a variance-reduced variant of Q-learning to achieve the optimality, but the algorithm requires an additional collection of i.i.d. samples at each outer loop to obtain an Monte Carlo approximation of the population Bellman operator (5). Our results show that a simple average is sufficient to guarantee optimality. Moreover, the computation of $\bar{\boldsymbol{Q}}_T$ is fully online with no additional samples needed.

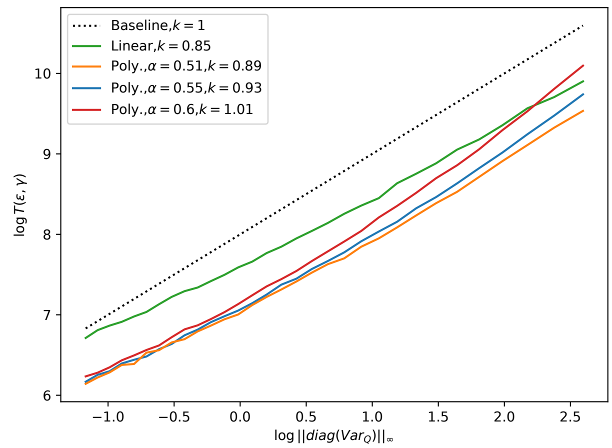

Figure 2: Log-log plots of the sample complexity $T(\varepsilon,\gamma)$ versus the asymptotic variance $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty$.

Confirming the theoretical predictions. We provide numerical experiments to illustrate instance-adaptivity as well as the worst-case behavior delineated in Theorem 5.1. We focus on the sample complexity $T(\varepsilon,\gamma) = \inf\{T : \mathbb{E}\|\bar{\boldsymbol{Q}}_T-\boldsymbol{Q}^*\|_\infty \le \varepsilon\}$ for $\varepsilon = 10^{-4}$. We conduct 10^3 independent trials in a random MDP to compute $T(\varepsilon,\gamma)$ for different values of $\gamma \in \Gamma$ and two step sizes. We plot the least-squares fits, $\{(\log\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty, \log T(\varepsilon,\gamma))\}_{\gamma\in\Gamma}$, and provide the slopes k of these lines in the legend. Further details are provided in Appendix J. At a high level, we see that averaged Q-learning produces sample complexity that is well predicted by our theory—all the slopes are no larger than the theoretical limit k predicted by our theory.

Proof Sketch. The proof idea of Theorem 5.1 is based on that of Theorem 3.1. Notice that $\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^* = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\Delta}_t = \frac{1}{\sqrt{T}}\boldsymbol{\phi}_T(1)$. From (14), we know that[5]

$$\mathbb{E}\|\bar{\boldsymbol{Q}}_T-\boldsymbol{Q}^*\|_\infty = \mathbb{E}\left\|\frac{\boldsymbol{\phi}_T(1)}{\sqrt{T}}\right\|_\infty \le \frac{\sum_{i\ne 4}\mathbb{E}\|\boldsymbol{\psi}_i(1)\|_\infty}{\sqrt{T}}.$$

Bounding the term of $i=0$ is easy since it's deterministic. Because $\boldsymbol{\psi}_i(1)(i=1,2,3)$ is a weighted sum of martingale differences, we use the variance-aware multi-dimensional Freedman's inequality (in Lemma H.1) to analyze its expectation under ℓ_∞-norm. The instance-dependent dominant term comes from the variance term for $\mathbb{E}\|\boldsymbol{\psi}_1(1)\|_\infty$. Analyzing the variance of $\boldsymbol{\psi}_2(1)$ is quite challenging since it relies on $\frac{1}{T}\sum_{j=1}^T\|\boldsymbol{A}_j^T-\boldsymbol{G}^{-1}\|_\infty^2$ with $\boldsymbol{A}_j^T$ defined in (13). We then bound that quantity in terms of $\alpha, 1-\gamma$ and T in Lemma C.4. Finally, due to $\|\boldsymbol{\psi}_5(1)\|_\infty \le L\|\boldsymbol{\Delta}_t\|_\infty^2$, bounding $\mathbb{E}\|\boldsymbol{\psi}_5(1)\|_\infty$ is reduced to bound $\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2$ for all $t \ge 0$, which can be given by a similar argument from Wainwright [2019b]. Putting all pieces together completes the proof; the detailed proof is in Appendix F.

[5] Since $r=1$, $\boldsymbol{\psi}_4$ doesn't appear in the decomposition.

6 RELAXATION OF THE LIPSCHITZ CONDITION

Both our asymptotic and nonasymptotic analysis rely on the Lipschitz condition in Assumption 3.2. That condition is essentially equivalent to assuming a unique optimal policy. It turns out that, once regularized by entropy, the (regularized) optimal policy is naturally unique. In the following, we show that entropy-regularized Q-learning enjoys a similar functional CLT and instance-dependent bounds without Assumption 3.2.

Entropy-regularized Q-learning uses the following matrix-form update rule,

$$\widetilde{\boldsymbol{Q}}_t = (1-\eta_t)\widetilde{\boldsymbol{Q}}_{t-1} + \eta_t \widetilde{\mathcal{T}}_t \widetilde{\boldsymbol{Q}}_{t-1}, \tag{16}$$

where

$$\widetilde{\mathcal{T}}_t(\boldsymbol{Q})(s,a) = \boldsymbol{r}_t(s,a) + \gamma(\boldsymbol{P}_t \mathcal{L}_\lambda \boldsymbol{Q})(s_t) \tag{17}$$

is a soft version of the empirical Bellman operator $\widehat{\mathcal{T}}$. The nonlinear operator $\mathcal{L}_\lambda(\cdot): \mathbb{R}^D \to \mathbb{R}^S$ is a soft version of a hard max, with regularization coefficient λ. It is defined by

$$(\mathcal{L}_\lambda \boldsymbol{Q})(s) := \max_{\pi \in \Pi} \mathbb{E}_{a\sim\pi(\cdot|s)} \left[Q(s,a) - \lambda \log \pi(a|s) \right].$$

Let $\boldsymbol{Q}_\lambda^*$ denote the unique fixed point of the regularized Bellman equation $\boldsymbol{Q}_\lambda^* = \boldsymbol{r} + \gamma \boldsymbol{P}\mathcal{L}_\lambda \boldsymbol{Q}_\lambda^*$ and let π_λ^* be the unique optimal policy.

Theorem 6.1. *Define $\{\widetilde{\boldsymbol{Q}}_t\}_{t\geq 0}$ in (16). The corresponding partial-sum process is $\widetilde{\boldsymbol{\phi}}_T(r) := \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr \rfloor}(\widetilde{\boldsymbol{Q}}_t - \boldsymbol{Q}_\lambda^*)$. Under Assumptions 3.1 and 3.3,*

$$\widetilde{\boldsymbol{\phi}}_T(\cdot) \overset{w}{\to} \widetilde{\mathrm{Var}}_{\boldsymbol{Q}}^{1/2} \boldsymbol{B}_D(\cdot),$$

where $\widetilde{\mathrm{Var}}_{\boldsymbol{Q}}$ is the asymptotic matrix defined by

$$\widetilde{\mathrm{Var}}_{\boldsymbol{Q}} := (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi_\lambda^*})^{-1} \mathrm{Var}(\widetilde{\boldsymbol{Z}})(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi_\lambda^*})^{-\top}.$$

with $\widetilde{\boldsymbol{Z}} \overset{d.}{=} \widetilde{\boldsymbol{Z}}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\mathcal{L}_\lambda \boldsymbol{Q}_\lambda^*$ the regularized Bellman noise.

Theorem 6.2. *Under Assumptions 3.1 and 3.3, when the two step sizes are considered, $\mathbb{E}\|\frac{1}{T}\sum_{t=1}^T \widetilde{\boldsymbol{Q}}_t - \boldsymbol{Q}_\lambda^*\|_\infty$ has similar bounds as in Theorem 5.1 except that we replace $\mathrm{Var}_{\boldsymbol{Q}}, L$ with $\widetilde{\mathrm{Var}}_{\boldsymbol{Q}}$ and $\frac{1}{\lambda}$.*

We note that the two theorems in this section can be proved via an almost identical argument as Theorem 3.1 and 5.1, since Assumption 3.2 is naturally satisfied with $L = \frac{1}{\lambda}$ for entropy-regularized Q-learning (see Appendix I). Actually, our proof is applicable to a class of nonlinear SAs.[6] Second, due to the bias introduced by entropy, the instance-dependent factor changes from $\mathrm{Var}_{\boldsymbol{Q}}$ to $\widetilde{\mathrm{Var}}_{\boldsymbol{Q}}$ and $\frac{1}{T}\sum_{t=1}^T \widetilde{\boldsymbol{Q}}_t$ converges to $\boldsymbol{Q}_\lambda^*$ instead of $\boldsymbol{Q}^*$ in expectation. Finally, note that these results provide a new argument for the benefits of entropy regularization; it smooths the Bellman operator and weakens the assumptions required for asymptotic analysis. It is supplementary to previous efforts that shows entropy regularization aids exploration [Fox et al., 2016], encourages robust optimal policies [Eysenbach and Levine, 2021], induces a smoother landscape [Ahmed et al., 2019], and hastens the convergence of RL algorithms [Cen et al., 2022].

[6] More specifically, our method can analyze $\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r}_t + \gamma \boldsymbol{P}_t \mathcal{L}\boldsymbol{Q}_{t-1})$ where $\mathcal{L}$ is a smooth nonlinear non-expansive operator.

7 DISCUSSION

We have studied the asymptotic and nonasymptotic convergence of averaged Q-learning, establishing its statistical efficiency. We first established a functional central limit theorem, showing that the standardized partial-sum process converges weakly to a rescaled Brownian motion, a result which can serve as an underpinning for the development of statistical inference methods for RL. We then established a semiparametric efficiency lower bound for $\boldsymbol{Q}^*$ estimation, showing that the averaged iterate $\bar{\boldsymbol{Q}}_T$ is the most efficient RAL estimator in the sense of having the smallest asymptotic variance. Finally, we presented the first finite-sample error analysis of $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty$ in the ℓ_∞-norm for both linearly rescaled and polynomial step sizes. We showed that averaged Q-learning achieves the same instance-dependent optimality and worst-case optimality as previous variance-reduced algorithms [Khamaru et al., 2021b, Wainwright, 2019c] under a Lipschitz condition.

Some open problems remain. On the one hand, with the Lipschitz condition, it's unclear whether averaged Q-learning with linearly rescaled step sizes can match the instance-dependent lower bound. Additionally, we suspect that the dependence on $(1-\gamma)^{-1}$ of the nonleading terms in Theorem 5.1 is loose and speculate it can be improved by finer analysis. On the other hand, without the Lipschitz condition, it is not clear whether averaged Q-learning still achieves the optimal instance-dependent bound. Finally, previous analysis [Kozuno et al., 2022] shows the last-iterate entropy-regularized Q-learning is minimax optimal. It is also unknown whether the averaged iterates of entropy-regularized Q-learning achieve the optimal instance-dependent bound.

Acknowledgement

The authors would like to express their gratitude to Prof. Csaba Szepesvári for his valuable suggestion regarding the relaxation of the Lipschitz condition through entropy regularization. Xiang Li and Zhihua Zhang have been supported by the National Key Research and Development Project of China (No. 2022YFA1004002) and the National Natural Science Foundation of China (No. 12271011).

References

Karim M Abadir and Paolo Paruolo. Simple robust testing of regression hypotheses: A comment. *Econometrica*, 70 (5):2097–2099, 2002.

Jinane Abounadi, Dimitri P Bertsekas, and Vivek Borkar. Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms. *SIAM Journal on Control and Optimization*, 41(1):1–22, 2002.

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

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pages 151–160. PMLR, 2019.

Oron Anschel, Nir Baram, and Nahum Shimkin. Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In *International Conference on Machine Learning*, pages 176–185. PMLR, 2017.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Necdet Batir. Inequalities for the gamma function. *Archiv der Mathematik*, 91(6):554–563, 2008.

Carolyn L Beck and Rayadurgam Srikant. Error bounds for constant step-size Q-learning. *Systems and Control Letters*, 61(12):1203–1208, 2012.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, pages 1691–1692. PMLR, 2018.

Vivek Borkar, Shuhang Chen, Adithya Devraj, Ioannis Kontoyiannis, and Sean Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *arXiv preprint arXiv:2110.14427*, 2021.

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48. Springer, 2009.

Vivek S Borkar and Sean P Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38 (2):447–469, 2000.

Donald L Burkholder. Sharp inequalities for martingales and stochastic integrals. *Astérisque*, 157(158):75–94, 1988.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *Annals of Statistics*, 48(1):251–273, 2020a.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *arXiv preprint arXiv:1905.11425*, 2019.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv e-prints*, pages arXiv–2002, 2020b.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567*, 2021.

Monroe David Donsker. An invariance principle for certain probability limit theorems. *American Mathematical Society*, pages 1–10, 1951.

Richard M Dudley. *Uniform Central Limit Theorems*, volume 142. Cambridge University Press, 2014.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for Polyak-Ruppert averaged iterates of linear stochastic approximation. *arXiv preprint arXiv:2207.04475*, 2022.

Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for Q-learning. *Journal of Nachine Learning Research*, 5(1), 2003.

Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. In *International Conference on Learning Representations*, 2021.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Conference on Uncertainty in Artificial Intelligence*, pages 202–211, 2016.

David A Freedman. On tail probabilities for martingales. *Annals of Probability*, pages 100–118, 1975.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12 (1):461–529, 2018.

Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Peter Hall and Christopher C Heyde. *Martingale Limit Theory and its Application*. Academic Press, 2014.

Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvári, and Mengdi Wang. Bootstrapping statistical inference for off-policy evaluation. *arXiv preprint arXiv:2102.03607*, 2021.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*, 1993.

Jean Jacod and Albert N Shiryaev. Skorokhod topology and convergence of processes. In *Limit Theorems for Stochastic Processes*, pages 324–388. Springer, 2003.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

Moritz Jirak. On weak invariance principles for partial sums. *Journal of Theoretical Probability*, 30(3):703–728, 2017.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Mach. Learn. Res.*, 21(167):1–63, 2020.

Michael Kearns and Satinder Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in Neural Information Processing Systems*, pages 996–1002, 1999.

Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, 49 (2):193–208, 2002.

Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. Is temporal difference learning optimal? An instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4): 1013–1040, 2021a.

Koulik Khamaru, Eric Xia, Martin J Wainwright, and Michael I Jordan. Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning. *arXiv preprint arXiv:2106.14352*, 2021b.

Koulik Khamaru, Eric Xia, Martin J Wainwright, and Michael I Jordan. Instance-dependent confidence and early stopping for reinforcement learning. *arXiv preprint arXiv:2201.08536*, 2022.

Nicholas M Kiefer, Timothy J Vogelsang, and Helle Bunzel. Simple robust testing of regression hypotheses. *Econometrica*, 68(3):695–714, 2000.

Vijaymohan Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in Neural Information Processing Systems, 12*, 1999.

Tadashi Kozuno, Wenhao Yang, Nino Vieillard, Toshinori Kitamura, Yunhao Tang, Jincheng Mei, Pierre Ménard, Mohammad Gheshlaghi Azar, Michal Valko, Rémi Munos, et al. KL-entropy-regularized RL with a generative model is minimax optimal. *arXiv preprint arXiv:2205.14211*, 2022.

Harold Kushner and G George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media, 2003.

Tor Lattimore and Marcus Hutter. Near-optimal PAC bounds for discounted MDPs. *Theoretical Computer Science*, 558:125–143, 2014.

Donghwan Lee and Niao He. Target-based temporal-difference learning. In *International Conference on Machine Learning*, pages 3713–3722. PMLR, 2019a.

Donghwan Lee and Niao He. A unified switching system perspective and ODE analysis of Q-learning algorithms. *arXiv preprint arXiv:1912.02270*, 2019b.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. *arXiv preprint arXiv:2106.03156*, 2021.

Erich L Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33: 12861–12872, 2020a.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *arXiv preprint arXiv:2006.03041*, 2020b.

Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is Q-learning minimax optimal? A tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021a.

Tianjiao Li, Guanghui Lan, and Ashwin Pananjady. Accelerated and instance-optimal policy evaluation with linear function approximation. *arXiv preprint arXiv:2112.13109*, 2021b.

Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via local SGD. *Proceedings of Thirty Fifth Conference on Learning Theory*, 178:1613–1661, 2022.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning of Representations*, 2016.

Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using V-learning. *Journal of the American Statistical Association*, 2019.

Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *International Conference on Machine Learning*, pages 664–671, 2008.

Abdelkader Mokkadem and Mariane Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16(3):1671–1702, 2006.

Terrence Joseph Moore Jr. *A theory of Cramér-Rao bounds for constrained parametric models*. University of Maryland, College Park, 2010.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pages 2947–2997. PMLR, 2020a.

Wenlong Mou, Ashwin Pananjady, and Martin J Wainwright. Optimal oracle inequalities for solving projected fixed-point equations. *arXiv preprint arXiv:2012.05299*, 2020b.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24:451–459, 2011.

Whitney K Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

Ashwin Pananjady and Martin J Wainwright. Instance-dependent ℓ_∞ bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585, 2020.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Martin L Puterman and Shelby L Brumelle. On the convergence of policy iteration in stationary dynamic programming. *Mathematics of Operations Research*, 4(1):60–69, 1979.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Conference on Learning Theory*, pages 3185–3205. PMLR, 2020.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Devavrat Shah and Qiaomin Xie. Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 3115–3125, 2018.

Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*, 2020.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5192–5202, 2018a.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM, 2018b.

Weijie J Su and Yuancheng Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Csaba Szepesvári et al. The asymptotic convergence-rate of Q-learning. *Advances in Neural Information Processing Systems*, pages 1064–1070, 1998.

Anastasios A Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

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

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

Aad W Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Karel Vermeulen. *Semiparametric Efficiency*. Gent Universiteit, 2011.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019a.

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

Martin J Wainwright. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019c.

Christopher Watkins. *Learning from delayed rewards*. PhD thesis, 1989.

Wenhao Yang, Xiang Li, and Zhihua Zhang. A regularized approach to sparse optimal policy in reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5938–5948, 2019.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR, 2020.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in Neural Information Processing Systems*, 34, 2021.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, pages 1–30, 2021.

A RELATED WORK

Due to the rapidly growing literature on Q-learning, we review only the theoretical results that are most relevant to our work. Interested readers can check references therein for more information.

Asymptotic normality in RL. Establishing asymptotic normality of an estimator permits statistical inference and the quantification of uncertainty. Existing work on statistical inference for Q-learning has focused mainly on the off-policy evaluation (OPE) problem, where one aims to estimate the value function of a given policy using pre-collected data. In this setting, a parametric Cramer–Rao lower bound has been established by Jiang and Li [2016], and asymptotic efficiency has been established for certain estimators using linear approximation [Uehara et al., 2020, Hao et al., 2021, Yin and Wang, 2020, Mou et al., 2020a] or bootstrapping [Hao et al., 2021]. Further inferential work includesthe asymptotic analysis of multi-stage algorithms [Luckett et al., 2019, Shi et al., 2020], asymptotic behavior of robust estimators [Yang et al., 2021], and work by Kallus and Uehara [2020] on a semiparametric doubly robust estimator.

In contradistinction to existing work, we establish a functional central limit theorem that captures the weak convergence of the whole trajectory rather than its endpoint. Such functional results have not been presented previously in the RL literature. Furthermore, we supplement these upper bounds with a semiparametric efficiency lower bound which additionally considers the randomness of rewards. We also show that averaged Q-learning is the most efficient RAL estimator vis-a-vis this lower bound.

Sample complexity for Q-learning. For the goal of obtaining an ε-accurate estimate of the optimal Q-function in a γ-discounted MDP in the presence of a generative model, model-based Q-value-iteration has been shown to achieve optimal minimax sample complexity $\widetilde{\mathcal{O}}\left(\frac{D}{\varepsilon^2(1-\gamma)^3}\right)$ [Azar et al., 2013, Agarwal et al., 2020, Li et al., 2020a]. In the model-free context, Wainwright [2019b] showed empirically that classical Q-learning suffers from at least worst-case fourth-order scaling in $(1-\gamma)^{-1}$ in sample complexity. A complexity bound of $\widetilde{\mathcal{O}}\left(\frac{D}{\varepsilon^2(1-\gamma)^5}\right)$ has been provided [Wainwright, 2019b, Chen et al., 2020b]; this is far from the optimal though better than previous efforts [Even-Dar et al., 2003, Beck and Srikant, 2012]. Li et al. [2021a] gave a sophisticated analysis showing the complexity of Q-learning is $\widetilde{\mathcal{O}}\left(\frac{D}{\varepsilon^2(1-\gamma)^4}\right)$ and provided a matching lower bound to confirm its sharpness. Wainwright [2019c], Khamaru et al. [2021b] introduced a variance-reduced variant of Q-learning [Gower et al., 2020] that achieves the optimal sample complexity and instance complexity. Our results show that a simple average over all history Q_t is sufficient to guarantee the same optimality. The averaged method is fully online without requiring additional samples and storage space.

Instance-dependent convergence in RL. Recent years have witnessed new instance-specific bounds, where an instance-dependent functional of a variance structure appears as the dominant term on stochastic errors. Unlike global minimax bounds which are worst-case in nature, instance-specific bounds help identify the difficulty of estimation case by case. Such bounds have been established for policy evaluation in the tabular setting [Pananjady and Wainwright, 2020, Khamaru et al., 2021a, Li et al., 2020a] or with linear function approximation [Li et al., 2021b] and for optimal value function estimation [Yin and Wang, 2021]. The most related work to ours is by Khamaru et al. [2021b], who show that a variance-reduced variant of Q-learning achieves the instance-dependent optimality after identifying an instance-dependent lower bound for Q^* estimation. By contrast, our result shows that a simple average is sufficient to yield optimality.

Nonlinear stochastic approximation. Q-learning has also been studied through the lens of nonlinear stochastic approximation. From this general point of view, asymptotic convergence has been provided [Tsitsiklis, 1994, Borkar and Meyn, 2000]. On the nonasymptotic side, Q-learning is studied either in the synchronous setting [Shah and Xie, 2018, Wainwright, 2019b, Chen et al., 2020b] or the asynchronous setting where only one sample from current state-action pair is available at a time [Qu and Wierman, 2020, Li et al., 2020b, Chen et al., 2021]. The sample complexities obtained therein are far from optimal. Others consider Q-learning with linear function approximation in the ℓ_2-norm [Melo et al., 2008, Chen et al., 2019]. Asymptotic convergence of averaged Q-learning has been studied by Lee and He [2019a,b] via the ODE (ordinary differential equation) approach. Our results are complementary to these results, including asymptotic statistical properties and finite-sample analysis in the ℓ_∞-norm. Though peculiar to averaged Q-learning, we believe our analysis can be extended to nonlinear SA problems.

B CENTRAL LIMIT THEOREM FOR AVERAGED Q-LEARNING

For completeness, we present a CLT for the averaged Q-learning sequence $\bar{\boldsymbol{Q}}_T := \frac{1}{T}\sum_{t=1}^T \boldsymbol{Q}_t$ in this part. This result can be derived not only from our Theorem 3.1 but also from CLT for non-linear SA, e.g., Mokkadem and Pelletier [2006].

Theorem B.1 (Asymptotic normality for Q^*)**.** *Under Assumptions 3.1, 3.2 and 3.3, we have*

$$\sqrt{T}(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \mathrm{Var}_{\boldsymbol{Q}}),$$

where the asymptotic variance is given by

$$\mathrm{Var}_{\boldsymbol{Q}} = (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\mathrm{Var}(\boldsymbol{Z})(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-\top} \in \mathbb{R}^{D\times D}. \tag{18}$$

Here $\mathrm{Var}(\boldsymbol{Z})$ *is the covariance matrix of the Bellman noise* $\boldsymbol{Z}$ *defined in* (7).

Asymptotic variance. Theorem B.1 implies that the average of the sequence $(\boldsymbol{Q}_t)$ has an asymptotic normal distribution with $\mathrm{Var}_{\boldsymbol{Q}}$ the asymptotic variance. $\mathrm{Var}_{\boldsymbol{Q}}$ includes $\mathrm{Var}(\boldsymbol{Z})$, the covariance matrix of Bellman noise $\boldsymbol{Z}$, multiplied with a pre-factor $(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}$. By a von Neumann expansion, $(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}$ is equivalent to $\sum_{t=0}^{\infty}(\gamma\boldsymbol{P}^{\pi^*})^t$. As argued by Khamaru et al. [2021b], the sum of the powers of $\gamma\boldsymbol{P}^{\pi^*}$ accounts for the compounded effect of an initial perturbation when following the MDP induced by π^*. The Bellman noise $\boldsymbol{Z}$ reflects the noise present in the empirical Bellman operator (4) as an estimate of the population Bellman operator (5). Note that this implies $\|(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-1}\| \le \sum_{t=0}^{\infty}\gamma^t\|(\boldsymbol{P}^{\pi^*})^t\|_\infty = (1-\gamma)^{-1}$. $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty$ coincides with the instance-dependent functional proposed by Khamaru et al. [2021b] that controls the difficulty of estimating Q^* in the ℓ_∞-norm.

Asymptotic normality for V^* estimation. If the optimal policy is unique, we can obtain a similar result for the optimal value function V^*, making use of the asymptotic normality of $\bar{\boldsymbol{Q}}_T$. We define an estimator $\bar{\boldsymbol{V}}_T \in \mathbb{R}^S$ greedily from $\bar{\boldsymbol{Q}}_T \in \mathbb{R}^D$: the s-th entry of $\bar{\boldsymbol{V}}_T$ is $\bar{V}_T(s) \in \arg\max_{a\in\mathcal{A}} \bar{Q}_T(s,a)$. As a corollary of Theorem B.1, $\bar{\boldsymbol{V}}_T$ enjoys a similar asymptotic normality with the asymptotic variance defined by $\mathrm{Var}_{\boldsymbol{V}}$. One can check that

$$\mathrm{Var}_{\boldsymbol{V}} = \boldsymbol{\Pi}^{\pi^*}\mathrm{Var}_{\boldsymbol{Q}}(\boldsymbol{\Pi}^{\pi^*})^\top, \tag{19}$$

where $\boldsymbol{\Pi}^{\pi^*} \in \{0,1\}^{S\times D}$ is the projection matrix associated with the deterministic optimal policy π^* (see (2)). Hence, $\mathrm{Var}_{\boldsymbol{V}}$ is formed by selecting entries from $\mathrm{Var}_{\boldsymbol{Q}}$. In particular, $\mathrm{Var}_{\boldsymbol{V}}(s,s') = \mathrm{Var}_{\boldsymbol{Q}}((s,\pi^*(s)),(s',\pi^*(s')))$ for any $s,s'\in\mathcal{S}$. The proof is deferred to Appendix B.2.

Lemma B.1. *If* π^* *is unique, then we have a positive optimality gap* $\mathrm{gap} := \min_s \min_{a\neq\pi^*(s)} |V^*(s) - Q^*(s,a)| > 0$ *where* $\pi^*(s)$ *is the unique action satisfying* $V^*(s) = Q^*(s,a^*(s))$*. For any Q-function estimator* $\boldsymbol{Q}\in\mathbb{R}^D$*, it follows that* $\{\pi_{\boldsymbol{Q}} \neq \pi^*\} \subseteq \{\|\boldsymbol{Q}-\boldsymbol{Q}^*\|_\infty \ge \frac{\mathrm{gap}}{2}\}$ *and*

$$\|(\boldsymbol{P}^{\pi_{\boldsymbol{Q}}} - \boldsymbol{P}^{\pi^*})(\boldsymbol{Q}-\boldsymbol{Q}^*)\|_\infty \le L\|\boldsymbol{Q}-\boldsymbol{Q}^*\|_\infty^2 \text{ with } L = \frac{4}{\mathrm{gap}}, \tag{20}$$

where π_Q *is the greedy policy with respective to* Q *defined by* $\pi_Q(s) := \arg\max_{a\in\mathcal{A}} Q(s,a)$*. If* $\arg\max_{a\in\mathcal{A}} Q(s,a)$ *has more than one element, we break the tie by randomness.*

Corollary B.1 (Asymptotic normality for V^*)**.** *Let* $\bar{\boldsymbol{V}}_T \in \mathbb{R}^S$ *be the greedy value function computed from* $\bar{\boldsymbol{Q}}_T \in \mathbb{R}^D$*, i.e.,* $\bar{V}_T(s) \in \arg\max_{a\in\mathcal{A}} \bar{Q}_T(s,a)$*. Under Assumptions 3.1 and 3.3, if we assume the optimal policy* π^* *is unique, then*

$$\sqrt{T}(\bar{\boldsymbol{V}}_T - \boldsymbol{V}^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \mathrm{Var}_{\boldsymbol{V}}),$$

where the asymptotic variance is

$$\mathrm{Var}_{\boldsymbol{V}} = (\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi^*})^{-1}\mathrm{Var}(\boldsymbol{\Pi}^{\pi^*}\boldsymbol{Z})(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi^*})^{-\top} \in \mathbb{R}^{S\times S}, \tag{21}$$

and $\mathrm{Var}(\boldsymbol{\Pi}^{\pi^*}\boldsymbol{Z})$ *is the covariance matrix of the projected Bellman noise* $\boldsymbol{\Pi}^{\pi^*}\boldsymbol{Z}$.

Insights on sample efficiency. The asymptotic results shed light on the sample efficiency of averaged Q-learning. Under ideal conditions, we have

$$\sqrt{T}\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty \to \mathbb{E}\|\mathcal{Z}\|_\infty \approx \sqrt{\ln D}\sqrt{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty} \text{ where } \mathcal{Z}\sim\mathcal{N}(\boldsymbol{0},\mathrm{Var}_{\boldsymbol{Q}}). \tag{22}$$

In this case, roughly speaking, to obtain an ε-accurate estimator of the optimal Q-value function $\boldsymbol{Q}^*$ (i.e., $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty \le \varepsilon$), we require approximately $T = \mathcal{O}\left(\frac{\ln D}{\varepsilon^2}\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty\right)$ iterations or equivalently $DT = \mathcal{O}\left(\frac{D\ln D}{\varepsilon^2}\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty\right)$ samples. This explains why Khamaru et al. [2021b] regarded $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty$ as the difficulty indicator because it affects the sample complexity directly.

B.1 Proof of Theorem B.1

Proof of Theorem B.1. One can prove Theorem B.1 by applying continuous mapping theorem to Theorem 3.1 with the functional $f : \mathsf{D}([0,1],\mathbb{R}^D) \to \mathbb{R}^D$, $f(w) = w(1)$. Once we can prove f is a continuous functional in $(\mathsf{D}([0,1],\mathbb{R}^D), d_0)$, an application of (27) would conclude the proof. Recalling the metric (25) defined on $\mathsf{D}([0,1],\mathbb{R}^D)$, we have for any $w_1, w_2 \in \mathsf{D}([0,1],\mathbb{R}^D)$,

$$\begin{aligned}\|f(w_1) - f(w_2)\|_\infty = \|w_1(1) - w_2(1)\|_\infty &\le \inf_{\lambda\in\Lambda} \sup_{t\in[0,1]} \|w_1(\lambda(t)) - w_2(t)\|_\infty \\ &\le \inf_{\lambda\in\Lambda} \left\{ \sup_{0\le s<t\le 1} \left| \ln \frac{\lambda(t)-\lambda(s)}{t-s} \right| + \sup_{t\in[0,1]} \|w_1(\lambda(t)) - w_2(t)\|_\infty \right\} = d_0(w_1, w_2).\end{aligned}$$

We even show that f is 1-Lipschitz continuous in $(\mathsf{D}([0,1],\mathbb{R}^D), d_0)$ and thus complete the proof. □

B.2 Proof of Corollary B.1

Proof of Corollary B.1. We first prove

$$\mathrm{Var}_{\boldsymbol{V}} = \boldsymbol{\Pi}^{\pi^*} \mathrm{Var}_{\boldsymbol{Q}} (\boldsymbol{\Pi}^{\pi^*})^\top. \tag{23}$$

Recall the definition

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{Q}} &= (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1} \mathrm{Var}(\boldsymbol{Z}) (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-\top} \in \mathbb{R}^{D\times D} \\ \mathrm{Var}_{\boldsymbol{V}} &= (\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi^*})^{-1} \mathrm{Var}(\boldsymbol{\Pi}^{\pi^*}\boldsymbol{Z}) (\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi^*})^{-\top} \in \mathbb{R}^{S\times S}.\end{aligned}$$

For one thing, we have $\mathrm{Var}(\boldsymbol{\Pi}^{\pi^*}\boldsymbol{Z}) = \boldsymbol{\Pi}^{\pi^*}\mathrm{Var}(\boldsymbol{Z})(\boldsymbol{\Pi}^{\pi^*})^\top$. For another thing, we have $\boldsymbol{\Pi}^{\pi^*}(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-1} = (\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi^*})^{-1}\boldsymbol{\Pi}^{\pi^*}$. This is because

$$(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi^*})\boldsymbol{\Pi}^{\pi^*} = \boldsymbol{\Pi}^{\pi^*} - \gamma\boldsymbol{\Pi}^{\pi^*}\boldsymbol{P}\boldsymbol{\Pi}^{\pi^*} = \boldsymbol{\Pi}^{\pi^*}(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*}).$$

Putting these together, (23) follows from direct verification.

We then prove the asymptotic normality of $\bar{\boldsymbol{V}}_T$. Let $\bar{\pi}_t$ is the greedy policy with respect to $\bar{\boldsymbol{Q}}_t$, i.e., $\bar{\pi}_t(s) \in \mathrm{argmax}_{a\in\mathcal{A}} \bar{Q}_t(s,a)$. From the definition of our estimator,

$$\bar{\boldsymbol{V}}_T = \boldsymbol{\Pi}^{\bar{\pi}_T}\bar{\boldsymbol{Q}}_T \quad \text{and} \quad \boldsymbol{V}^* = \boldsymbol{\Pi}^{\pi^*}\boldsymbol{Q}^*$$

which implies

$$\bar{\boldsymbol{V}}_T - \boldsymbol{V}^* = \left(\boldsymbol{\Pi}^{\bar{\pi}_T}\bar{\boldsymbol{Q}}_T - \boldsymbol{\Pi}^{\pi^*}\bar{\boldsymbol{Q}}_T\right) + \left(\boldsymbol{\Pi}^{\pi^*}\bar{\boldsymbol{Q}}_T - \boldsymbol{\Pi}^{\pi^*}\boldsymbol{Q}^*\right).$$

On the other hand, it is easy to see that

$$\sqrt{T}\left(\boldsymbol{\Pi}^{\pi^*}\bar{\boldsymbol{Q}}_T - \boldsymbol{\Pi}^{\pi^*}\boldsymbol{Q}^*\right) \overset{d.}{\to} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Pi}^{\pi^*}\mathrm{Var}_{\boldsymbol{Q}}(\boldsymbol{\Pi}^{\pi^*})^\top) = \mathcal{N}(\boldsymbol{0}, \mathrm{Var}_{\boldsymbol{V}}).$$

If we can prove

$$\sqrt{T}\left(\boldsymbol{\Pi}^{\bar{\pi}_T}\bar{\boldsymbol{Q}}_T - \boldsymbol{\Pi}^{\pi^*}\bar{\boldsymbol{Q}}_T\right) = o_{\mathbb{P}}(1), \tag{24}$$

then the conclusion follows from Slutsky's theorem. We have that

$$\begin{aligned}\sqrt{T}\mathbb{E}\|\boldsymbol{\Pi}^{\bar{\pi}_T}\bar{\boldsymbol{Q}}_T - \boldsymbol{\Pi}^{\pi^*}\bar{\boldsymbol{Q}}_T\|_\infty &\le \sqrt{T}\mathbb{E}\|\boldsymbol{\Pi}^{\bar{\pi}_T} - \boldsymbol{\Pi}^{\pi^*}\|_\infty\|\bar{\boldsymbol{Q}}_T\|_\infty \\ &\overset{(a)}{\le} \frac{\sqrt{T}}{1-\gamma}\mathbb{E}\|\boldsymbol{\Pi}^{\bar{\pi}_T} - \boldsymbol{\Pi}^{\pi^*}\|_\infty \\ &\overset{(b)}{=} \frac{2\sqrt{T}}{1-\gamma}\mathbb{P}\left(\bar{\pi}_T \ne \pi^*\right) \\ &\overset{(c)}{\le} \frac{2\sqrt{T}}{1-\gamma}\mathbb{P}\left(\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty \ge \frac{\mathrm{gap}}{2}\right) \\ &\le \frac{2\sqrt{T}}{1-\gamma}\frac{4}{\mathrm{gap}^2}\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty^2 \\ &\overset{(d)}{\le} \frac{1}{1-\gamma}\frac{8}{\mathrm{gap}^2}\frac{1}{\sqrt{T}}\sum_{t=1}^T \mathbb{E}\|\boldsymbol{Q}_t - \boldsymbol{Q}^*\|_\infty^2,\end{aligned}$$

where (a) uses $\|\bar{\boldsymbol{Q}}_T\|_\infty \le (1-\gamma)^{-1}$, (b) uses the fact that both $\bar{\pi}_T$ and π^* are deterministic policies and thus $\|\boldsymbol{\Pi}^{\bar{\pi}_T} - \boldsymbol{\Pi}^{\pi^*}\|_\infty = 2 \cdot 1_{\{\bar{\pi}_T \neq \pi^*\}}$, (c) uses the fact $\{\bar{\pi}_t \neq \pi^*\} \subseteq \{\|\bar{\boldsymbol{Q}}_t - \boldsymbol{Q}^*\|_\infty \ge \frac{\text{gap}}{2}\}$ which we derived in Lemma B.1, and finally (d) follows from Jensen's inequality.

From Theorem E.1, we know $\frac{1}{\sqrt{T}}\sum_{t=1}^T \mathbb{E}\|\boldsymbol{Q}_t - \boldsymbol{Q}^*\|_\infty^2 \to 0$ as $T \to \infty$. Therefore, we have that $\sqrt{T}\mathbb{E}\|\boldsymbol{\Pi}^{\bar{\pi}_T}\bar{\boldsymbol{Q}}_T - \boldsymbol{\Pi}^{\pi^*}\bar{\boldsymbol{Q}}_T\|_\infty = o(1)$ which implies (24) is true. □

B.3 Proof of Lemma B.1

Proof of Lemma B.1. Recall that $\text{gap} = \min_s \min_{a \neq \pi^*(s)} |Q^*(s, \pi^*(s)) - Q^*(s,a)|$. If $\text{gap} = 0$, by definition, there must exist some $s_0 \in \mathcal{S}$ and $a_0 \in \mathcal{A}$ such that $V^*(s_0) = Q^*(s_0, a_0)$ and $a_0 \neq \pi^*(s_0)$, which is contradictory with the uniqueness of π^*. Hence, a unique π^* implies a positive gap.

For any $\boldsymbol{Q}$ satisfying $\|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty < \frac{\text{gap}}{2}$, we must have $\|\boldsymbol{Q}(s,\cdot) - \boldsymbol{Q}^*(s,\cdot)\|_\infty < \frac{\text{gap}}{2}$ for any $s \in \mathcal{S}$. In this case, it must be true that $\pi_Q(s) = \pi^*(s)$ for all $s \in \mathcal{S}$. Otherwise, there exists some $s \in \mathcal{S}$ such that $\pi_Q(s) \neq \pi^*(s)$. We then have

$$Q(s, \pi_Q(s)) < Q^*(s, \pi_Q(s)) + \frac{\text{gap}}{2} \overset{(a)}{\le} Q^*(s, \pi^*(s)) - \frac{\text{gap}}{2} < Q(s, \pi^*(s)),$$

where (a) follows from the definition of the optimality gap. The result $Q(s,\pi_Q(s)) < Q(s, \pi^*(s))$ contradicts with the fact that $\pi_Q(s)$ is the greedy policy with respect to Q at state s, which implies $Q(s, \pi^*(s)) \le Q(s, \pi_Q(s))$. This implies that the event $\{\pi_{\boldsymbol{Q}} \neq \pi^*\} \subseteq \{\|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty \ge \frac{\text{gap}}{2}\}$ and thus $1_{\{\pi_Q \neq \pi^*\}} \le 1_{\{\|\boldsymbol{Q}-\boldsymbol{Q}^*\|_\infty \ge \frac{\text{gap}}{2}\}}$. Hence,

$$\begin{aligned}
\|(\boldsymbol{P}^{\pi_{\boldsymbol{Q}}} - \boldsymbol{P}^{\pi^*})(\boldsymbol{Q} - \boldsymbol{Q}^*)\|_\infty &\le \|\boldsymbol{P}^{\pi_{\boldsymbol{Q}}} - \boldsymbol{P}^{\pi^*}\|_\infty \|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty \\
&\le \|\boldsymbol{P}\|_\infty \|\boldsymbol{\Pi}^{\pi_{\boldsymbol{Q}}} - \boldsymbol{\Pi}^{\pi^*}\|_\infty \|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty \\
&= 1 \cdot 2 \cdot 1_{\{\pi_{\boldsymbol{Q}} \neq \pi^*\}} \cdot \|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty \\
&\le 2 \cdot 1_{\{\|\boldsymbol{Q}-\boldsymbol{Q}^*\|_\infty \ge \frac{\text{gap}}{2}\}} \|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty \\
&\le \frac{4}{\text{gap}} \|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty^2,
\end{aligned}$$

where the last line uses $1_{\{\|\boldsymbol{Q}-\boldsymbol{Q}^*\|_\infty \ge \frac{\text{gap}}{2}\}} \le \frac{2}{\text{gap}}\|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty$. □

B.4 Proof of Proposition 3.1

Proof of Proposition 3.1. Let $g : \mathsf{D}([0,1], \mathbb{R}^D) \to \mathbb{R}$ be a functional defined as

$$g(w) = w(1)^\top \left(\int_0^1 w(r) w(r)^\top dr \right)^{-1} w(1) \text{ for any } w \in \mathsf{D}([0,1], \mathbb{R}^D).$$

Here the domain of g is

$$\text{dom}(g) = \left\{ w \in \mathsf{D}([0,1], \mathbb{R}^D), \int_0^1 w(r) w(r)^\top dr \text{ is invertible} \right\}.$$

Once we prove g is continuous in $(\text{dom}(g), d_0)$, the continuous mapping theorem together with Theorem 3.1 would complete the proof for Proposition 3.1.

In Appendix B.1, we have shown $f : \mathsf{D}([0,1], \mathbb{R}^D) \to \mathbb{R}^D$, $f(w) = w(1)$ is 1-Lipschitz continuous in $(\mathsf{D}([0,1], \mathbb{R}^D), d_0)$. Let $h : \mathsf{D}([0,1], \mathbb{R}^D) \to \mathbb{R}^{D \times D}$ be defined by $h(w) = \int_0^1 w(r) w(r)^\top dr$. Hence, once we prove h is continuous in $(\mathsf{D}([0,1], \mathbb{R}^D), d_0)$, it follows that $g = f^\top h^{-1} f$ is also continuous in $(\text{dom}(g), d_0)$. To that end, we only show each entry of h is continuous in w. This is true because of each entry of h is in form of integration which is a continuous functional on the Skorohod space $\mathsf{D}([0,1], \mathbb{R})$.

Finally, by Theorem 3.1 and definition of weak convergence, we know that as T goes to infinity,

$$\mathbb{P}\left(\boldsymbol{\phi}_T \notin \text{dom}(g)\right) \to \mathbb{P}\left(\boldsymbol{B}_D \notin \text{dom}(g)\right) = 0.$$

Hence, with probability approaching to one, $\int_0^1 \boldsymbol{\phi}_T(r) \boldsymbol{\phi}_T(r)^\top dr$ is invertible and thus $g(\boldsymbol{\phi}_T)$ is well defined. □

C PROOF OF THEOREM 3.1

C.1 Preliminaries and High-level Idea

In this section, we provide a self-contained proof of our functional central limit theorem (FCLT). Let $\boldsymbol{\Delta}_t = \boldsymbol{Q}_t - \boldsymbol{Q}^*$ be the error vector at iteration t. The application of Polyak-Ruppert average [Polyak and Juditsky, 1992] gives an estimator for $\boldsymbol{Q}^*$: $\bar{\boldsymbol{Q}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Q}_t$. Then its partial sum of the first r-fraction ($r \in [0,1]$) is $\frac{1}{T}\sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{Q}_t$. The associated standardized partial-sum process is defined by

$$\boldsymbol{\phi}_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{\Delta}_t = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} (\boldsymbol{Q}_t - \boldsymbol{Q}^*).$$

Here $\boldsymbol{\phi}_T(\cdot)$ should be viewed as a D-dimensional random function. For simplicity, we also use $\boldsymbol{\phi}_T = \{\boldsymbol{\phi}_T(r)\}_{r\in[0,1]}$ to denote the whole function.

C.1.1 Weak convergence of measures in Polish spaces

We will introduce some basic knowledge of weak convergence in metric spaces. See Chapter VI in [Jacod and Shiryaev, 2003] for a detailed introduction.

A Polish space is a topological space that is separable, complete, and metrizable. Let $\mathsf{D}([0,1],\mathbb{R}^d) = \{\text{càdlàg function } \omega(r) \in \mathbb{R}^d, r \in [0,1]\}$ collect all d-dimensional functions which are right continuous with left limits. Define $\mathcal{D}([0,1],\mathbb{R}^d)$ as the σ-field generated by all maps $X \mapsto X(r)$ for $r \in [0,1]$. The J_1 Skorokhod topology equips $\mathsf{D}([0,1],\mathbb{R}^d)$ with a metric d_0 such that $(\mathsf{D}([0,1],\mathbb{R}^d), d_0)$ is a Polish space and $\mathcal{D}([0,1],\mathbb{R}^d)$ is its Borel σ-field (the σ-field generated by all open subsets). In particular, for any $w_1, w_2 \in \mathsf{D}([0,1],\mathbb{R}^d)$,

$$d_0(w_1,w_2) = \inf_{\lambda\in\Lambda}\left\{ \sup_{0\le s<t\le 1}\left| \ln \frac{\lambda(t)-\lambda(s)}{t-s} \right| + \sup_{t\in[0,1]} \|w_1(\lambda(t)) - w_2(t)\|_\infty \right\}, \tag{25}$$

where Λ denotes the class of strictly increasing continuous mappings $\lambda : [0,1] \to [0,1]$ with $\lambda(0) = 0$ and $\lambda(1) = 1$.

An important subset of $\mathsf{D}([0,1],\mathbb{R}^d)$ is $\mathsf{C}([0,1],\mathbb{R}^d) = \{\text{continuous } \omega(r) \in \mathbb{R}^d, r \in [0,1]\}$, which collects all d-dimensional continuous functions defined on $[0,1]$. The uniform topology equips $\mathsf{C}([0,1],\mathbb{R}^d)$ with the uniform norm

$$\|\omega\|_{\sup} := \sup_{r\in[0,1]} \|\omega(r)\|_\infty. \tag{26}$$

The resulting $(\mathsf{C}([0,1],\mathbb{R}^d), \|\cdot\|_{\sup})$ is a Polish space. Additionally, we have $d_0(w_1,w_2) \le \|w_1 - w_2\|_{\sup}$ for any $w_1, w_2 \in \mathsf{D}([0,1],\mathbb{R}^d)$. The J_1 Skorokhod topology is weaker than the uniform topology. However, if $X \in \mathsf{D}([0,1],\mathbb{R}^d)$ is a continuous function, a sequence $\{X_t\}_{t\ge0} \subseteq \mathsf{D}([0,1],\mathbb{R}^d)$ converges to X for the Skorokhod topology if and only if it converges to X under the uniform norm $\|\cdot\|_{\sup}$. Hence, the Skorokhod topology relativized to $\mathsf{C}([0,1],\mathbb{R}^d)$ coincides with the uniform topology there.

Any random element $X_t \in \mathsf{D}([0,1],\mathbb{R}^d)$ introduces a probability measure on $\mathsf{D}([0,1],\mathbb{R}^d)$ denoted by $\mathcal{L}(X_t)$ such that $(\mathsf{D}([0,1],\mathbb{R}^d), \mathcal{D}([0,1],\mathbb{R}^d), \mathcal{L}(X_t))$ becomes a probability space. We say a sequence of random elements $\{X_t\}_{t\ge0} \subseteq \mathsf{D}([0,1],\mathbb{R}^d)$ weakly converges to X, if for any bounded continuous function $f : \mathsf{D}([0,1],\mathbb{R}^d) \to \mathbb{R}$, we have

$$\mathbb{E}f(X_T) \to \mathbb{E}f(X) \text{ as } T \text{ goes to infinity}. \tag{27}$$

The condition is equivalent to that any finite-dimensional projections of $\boldsymbol{\phi}_T$ converge in distribution. We denote the weak convergence by $X_T \overset{w}{\to} X$.

Theorem C.1 (Slutsky's theorem on Polish spaces)**.** *Suppose $\mathcal{S}$ is a Polish space with metric d and $\{(X_t, Y_t)\}_{t\ge0}$ are random elements of $\mathcal{S}\times\mathcal{S}$. Suppose $X_T \overset{w}{\to} X$ and $d(X_T, Y_T) \overset{w}{\to} 0$, then $Y_T \overset{w}{\to} X$.*

By Slutsky's theorem in Theorem C.1, if $\|Y_T\|_{\sup} \overset{w}{\to} 0$ and $X_T \overset{w}{\to} X$, then $X_T + Y_T \overset{w}{\to} X$. A sufficient condition to $\|Y_T\|_{\sup} \overset{w}{\to} 0$ is $\mathbb{E}\|Y_T\|_{\sup} \to 0$ by Markov's inequality.

Proposition C.1. *For two random sequences $\{X_t\}_{t\ge0}, \{Y_t\}_{t\ge0} \subseteq \mathsf{D}([0,1],\mathbb{R}^d)$ satisfying $\mathbb{E}\|Y_T\|_{\sup} \to 0$ and $X_T \overset{w}{\to} X$, we have $X_T + Y_T \overset{w}{\to} X$.*

C.1.2 Proof Idea

In the following, we will show under the three assumptions in the main text, we can establish

$$\phi_T \xrightarrow{w} \mathrm{Var}_{\boldsymbol{Q}}^{1/2} \boldsymbol{B}_D,$$

where $\boldsymbol{B}_D \in \mathsf{C}([0,1], \mathbb{R}^D)$ is the standard D-dimensional Brownian motion on $[0,1]$. That is the associated measure of ϕ_T weakly converges to the measure introduced by $\mathrm{Var}_{\boldsymbol{Q}}^{1/2} \boldsymbol{B}_D$ on $\mathsf{D}([0,1], \mathbb{R}^D)$.

To proceed the proof, we will use two auxiliary sequences $\{\boldsymbol{\Delta}_t^1\}_{t\geq 0}$ and $\{\boldsymbol{\Delta}_t^2\}_{t\geq 0}$ defined in Lemma C.1. The proof of Lemma C.1 can be found in Appendix C.4.1.

Lemma C.1. *Denote* $\boldsymbol{G} = \boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*}$, $\boldsymbol{A}_t = \boldsymbol{I} - \eta_t \boldsymbol{G}$ *and* $\boldsymbol{W}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}$ *for short. The auxiliary sequences* $\{\boldsymbol{\Delta}_t^1\}_{t\geq 0}$ *and* $\{\boldsymbol{\Delta}_t^2\}_{t\geq 0}$ *are defined iteratively:* $\boldsymbol{\Delta}_0^1 = \boldsymbol{\Delta}_0^2 = \boldsymbol{\Delta}_0$ *and for* $t \geq 1$

$$\boldsymbol{\Delta}_t^1 = \boldsymbol{A}_t \boldsymbol{\Delta}_{t-1}^1 + \eta_t \left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*}) \boldsymbol{\Delta}_{t-1} \right] \tag{28}$$

$$\boldsymbol{\Delta}_t^2 = \boldsymbol{A}_t \boldsymbol{\Delta}_{t-1}^2 + \eta_t \boldsymbol{W}_t \ . \tag{29}$$

As long as $\sup_t \eta_t \leq 1$*, it follows that all* $t \geq 0$*,*

$$\boldsymbol{\Delta}_t^2 \leq \boldsymbol{\Delta}_t \leq \boldsymbol{\Delta}_t^1. \tag{30}$$

The two sequences form a sandwich bound for $\boldsymbol{\Delta}_t$, producing $\boldsymbol{\Delta}_t^2 \leq \boldsymbol{\Delta}_t \leq \boldsymbol{\Delta}_t^1$ coordinate-wise. We similarly define the error vectors of their first r-fraction partial sums as

$$\phi_T^1(r) := \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{\Delta}_t^1 \text{ and } \phi_T^2(r) := \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{\Delta}_t^2.$$

Then, it is valid that $\phi_T^1, \phi_T^2 \in \mathsf{D}([0,1], \mathbb{R}^D)$ and for any $r \in [0,1]$,

$$\phi_T^2(r) \leq \phi_T(r) \leq \phi_T^1(r). \tag{31}$$

In the following subsections, we will show that under Assumption 3.1, 3.2 and 3.3, we can find a random function $\mathcal{Z} \in \mathsf{D}([0,1], \mathbb{R}^D)$ which satisfies

$$\mathcal{Z} \xrightarrow{w} \mathrm{Var}_{\boldsymbol{Q}}^{1/2} \boldsymbol{B}_D. \tag{32}$$

Furthermore, ϕ_T^1 and ϕ_T^2 weakly converge to $\mathcal{Z}$ such that

$$\lim_{T\to\infty} \mathbb{E}\|\phi_T^k - \mathcal{Z}\|_{\sup} = 0 \text{ for } k = 1, 2. \tag{33}$$

By the sandwich inequality (31), we have

$$\mathbb{E}\|\phi_T - \mathcal{Z}\|_{\sup} \leq \mathbb{E}\|\phi_T^1 - \mathcal{Z}\|_{\sup} + \mathbb{E}\|\phi_T^2 - \mathcal{Z}\|_{\sup} \to 0$$

as T goes to infinity. Proposition C.1 implies ϕ_T weakly converges to a rescaled Brownian motion $\mathrm{Var}_{\boldsymbol{Q}}^{1/2} \boldsymbol{B}_D$, by which we complete the proof.

C.2 Functional CLT for ϕ_T^1

We first establish the FLCT of $\phi_T^1(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{\Delta}_t^1$, i.e., $\lim_{T\to\infty} \mathbb{E}\|\phi_T^1 - \mathcal{Z}\|_{\sup} = 0$ for some $\mathcal{L}(\mathcal{Z}) = \mathcal{L}(\mathrm{Var}_{\boldsymbol{Q}}^{1/2} \boldsymbol{B}_D)$. The FCLT of $\phi_T^2(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{\Delta}_t^2$ can be validated in an almost identical way. We start by rewriting (28) as

$$\boldsymbol{\Delta}_t^1 = \boldsymbol{A}_t \boldsymbol{\Delta}_{t-1}^1 + \eta_t \left(\boldsymbol{Z}_t + \gamma \boldsymbol{D}_{t-1}^1 \right), \tag{34}$$

where $\boldsymbol{A}_t = \boldsymbol{I} - \eta_t(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})$, $\boldsymbol{Z}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*$, and

$$\boldsymbol{D}_{t-1}^1 = (\boldsymbol{P}_t - \boldsymbol{P})(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*) + (\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*}) \boldsymbol{\Delta}_{t-1}. \tag{35}$$

We comment that $\{\boldsymbol{Z}_t\}_{t\geq 0}$ collects the i.i.d. noise inherent in the empirical Bellman operator and $\{\boldsymbol{D}_{t-1}^1\}_{t\geq 1}$ captures the closeness between the current Q-function estimator $\boldsymbol{Q}_{t-1}$ and the optimal $\boldsymbol{Q}^*$. Recurring (34) gives

$$\boldsymbol{\Delta}_t^1 = \prod_{j=1}^t \boldsymbol{A}_j \boldsymbol{\Delta}_0 + \sum_{j=1}^t \prod_{i=j+1}^t \boldsymbol{A}_i \eta_j \left(\boldsymbol{Z}_j + \gamma \boldsymbol{D}_{j-1}^1\right).$$

Here we use the convention that $\prod_{i=t+1}^t \boldsymbol{A}_i = \boldsymbol{I}$ for any $t \geq 0$. For any $r \in [0,1]$, summing the last equality over $t = 1, \cdots, \lfloor Tr \rfloor$ and scaling it properly, we have

$$\begin{aligned}
\boldsymbol{\phi}_T^1(r) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{\Delta}_t^1 = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \prod_{j=1}^t \boldsymbol{A}_j \boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \sum_{j=1}^t \prod_{i=j+1}^t \boldsymbol{A}_i \eta_j \left(\boldsymbol{Z}_j + \gamma \boldsymbol{D}_{j-1}^1\right) \\
&= \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \prod_{j=1}^t \boldsymbol{A}_j \boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} \sum_{t=j}^{\lfloor Tr \rfloor} \prod_{i=j+1}^t \boldsymbol{A}_i \eta_j \left(\boldsymbol{Z}_j + \gamma \boldsymbol{D}_{j-1}^1\right) \\
&= \frac{1}{\eta_0 \sqrt{T}} (\boldsymbol{A}_0^{\lfloor Tr \rfloor} - \eta_0 \boldsymbol{I}) \boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} \boldsymbol{A}_j^{\lfloor Tr \rfloor} \left(\boldsymbol{Z}_j + \gamma \boldsymbol{D}_{j-1}^1\right),
\end{aligned} \tag{36}$$

where the last line uses the following notation:

$$\boldsymbol{A}_j^T = \eta_j \sum_{t=j}^T \prod_{i=j+1}^t \boldsymbol{A}_i \text{ for any } T \geq j \geq 0. \tag{37}$$

Define $\boldsymbol{G} = \boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*}$ with $\gamma \in [0,1)$, then $\boldsymbol{A}_i = \boldsymbol{I} - \eta_i \boldsymbol{G}$. Typically speaking, $\boldsymbol{A}_j^T$ approximates $\boldsymbol{G}$ uniformly well (see Lemma C.4). By the observation, we further expand (36) and decompose $\boldsymbol{\phi}_T^1(r)$ into six terms $\{\boldsymbol{\psi}_i\}_{i=0}^5$ which will be analyzed respectively in the following:

$$\begin{aligned}
\boldsymbol{\phi}_T^1(r) &= \frac{1}{\eta_0 \sqrt{T}} (\boldsymbol{A}_0^{\lfloor Tr \rfloor} - \eta_0 \boldsymbol{I}) \boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}} \sum_{j=1}^{[Tr]} \boldsymbol{A}_j^{\lfloor Tr \rfloor} \left(\boldsymbol{Z}_j + \gamma \boldsymbol{D}_{j-1}^1\right) \\
&= \frac{1}{\eta_0 \sqrt{T}} (\boldsymbol{A}_0^{\lfloor Tr \rfloor} - \eta_0 \boldsymbol{I}) \boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} \boldsymbol{G}^{-1} \boldsymbol{Z}_j + \frac{1}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} (\boldsymbol{A}_j^{\lfloor Tr \rfloor} - \boldsymbol{G}^{-1}) \boldsymbol{Z}_j \\
&\quad + \frac{\gamma}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} \boldsymbol{A}_j^{\lfloor Tr \rfloor} (\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*) + \frac{\gamma}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} \boldsymbol{A}_j^{\lfloor Tr \rfloor} (\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*}) \boldsymbol{\Delta}_{j-1} \\
&= \frac{1}{\eta_0 \sqrt{T}} (\boldsymbol{A}_0^{\lfloor Tr \rfloor} - \eta_0 \boldsymbol{I}) \boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} \boldsymbol{G}^{-1} \boldsymbol{Z}_j + \frac{1}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} (\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}) \boldsymbol{Z}_j \\
&\quad + \frac{\gamma}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} \boldsymbol{A}_j^T (\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*) + \frac{1}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} (\boldsymbol{A}_j^{\lfloor Tr \rfloor} - \boldsymbol{A}_j^T) \left[\boldsymbol{Z}_j + \gamma (\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)\right] \\
&\quad + \frac{\gamma}{\sqrt{T}} \sum_{j=1}^{\lfloor Tr \rfloor} \boldsymbol{A}_j^{\lfloor Tr \rfloor} (\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*}) \boldsymbol{\Delta}_{j-1} \\
&:= \boldsymbol{\psi}_0(r) + \boldsymbol{\psi}_1(r) + \boldsymbol{\psi}_2(r) + \boldsymbol{\psi}_3(r) + \boldsymbol{\psi}_4(r) + \boldsymbol{\psi}_5(r).
\end{aligned} \tag{38}$$

Readers should keep in mind that all $\boldsymbol{\psi}_i$'s depend on T, a dependence which we omit for simplicity. In the following, we will show (32) is true by setting $\mathcal{Z} = \boldsymbol{\psi}_1$. In order to establish (33), we will show that $\mathbb{E}\|\boldsymbol{\psi}_i\|_{\sup} = o(1)$ for $i = 0, 2, 3, 4, 5$. In this way, based on (38), we have

$$\mathbb{E}\|\boldsymbol{\phi}_T^1 - \boldsymbol{\psi}_1\|_{\sup} \leq \sum_{i=0,2,3,4,5} \mathbb{E}\|\boldsymbol{\psi}_i\|_{\sup} = o(1) \text{ as } T \to \infty,$$

and validate (33). To that end, we first study the properties of $\{\boldsymbol{A}_j^T\}_{0\leq j\leq T}$ since it appears in many $\boldsymbol{\psi}_i$'s.

C.2.1 Properties of $\{\boldsymbol{A}_j^T\}_{0\le j\le T}$

First, prior work [Polyak and Juditsky, 1992] considers a general step size $\{\eta_t\}_{t\ge 0}$ satisfying Assumption 3.3 and establishes the following lemma.

Lemma C.2 (Lemma 1 in [Polyak and Juditsky, 1992])**.** *For $\{\eta_t\}_{t\ge 0}$ satisfying Assumption 3.3,*

- *Uniform boundedness: $\|\boldsymbol{A}_j^T\|_\infty \le C_0$ uniformly for all $T \ge j \ge 0$ for some constant $C_0 \ge 1$;*
- *Uniform approximation: $\lim_{T\to\infty} \frac{1}{T}\sum_{j=1}^T \|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_2 = 0$.*

Lemma C.2 shows that when the step size η_t decreases at a slow rate, $\boldsymbol{A}_j^T$ is uniformly bouned (that is $\sup_{T\ge j\ge 1}\|\boldsymbol{A}_j^T\|_\infty < \infty$) and is a good surrogate of $\boldsymbol{G}^{-1} := (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}$ in the asymptotic sense: $\lim_{T\to\infty}\frac{1}{T}\sum_{j=1}^T \|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_2 = 0$.[7] It is sufficient to derive our asymptotic result. However, on purpose of non-asymptotic analysis, we should provide a non-asymptotic counterpart capturing the specific decaying rate in the ℓ_∞-norm. Therefore, we consider two specific step sizes, namely (S1) the linear rescaled step size and (S2) polynomial step size. Define $\widetilde{\eta}_t = (1-\gamma)\eta_t$ as the rescaled step size for simplicity, we have

(S1) linear rescaled step size that uses $\eta_t = \frac{1}{1+(1-\gamma)t}$ (equivalently $\widetilde{\eta}_t = \frac{1-\gamma}{1+(1-\gamma)t}$);

(S2) polynomial step size that uses $\eta_t = t^{-\alpha}$ with $\alpha \in (0,1)$ for $t \ge 1$ and $\eta_0 = 1$.

The first is uniform boundedness whose proof is provided in Appendix C.4.2.

Lemma C.3 (Uniform boundedness)**.** *There exists some $c > 0$ such that*

$$\|\boldsymbol{A}_j^T\|_\infty \le C_0 := \begin{cases} \frac{\ln(1+(1-\gamma)T)}{1-\gamma} & \text{(S1)} \\ \frac{c2^{\frac{1}{1-\alpha}}}{\sqrt{1-\alpha}}\frac{1}{(1-\gamma)^{\frac{1}{1-\alpha}}} & \text{(S2)} \end{cases} \quad \textit{for any } T \ge j \ge 1.$$

The second is the uniform approximation. The proof is deferred in Appendix C.4.3. We observe that as T grows, $\frac{1}{T}\sum_{j=1}^T \|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\boldsymbol{I}\|_\infty^2$ vanishes under (S2), but is only guaranteed to be bounded for (S1). This is not contradictory with Lemma C.2 since (S1) doesn't satisfy Assumption 3.3.

Lemma C.4 (Uniform approximation)**.** *There exists some constant $c > 0$ such that*

$$\sqrt{\frac{1}{T}\sum_{j=1}^T \|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_\infty^2} \le \begin{cases} \frac{5}{1-\gamma} & \text{(S1)} \\ \frac{c\alpha 2^{\frac{1}{1-\alpha}}}{\sqrt{T}}\left[\frac{1}{(1-\alpha)^{\frac{3}{2}}}\frac{1}{(1-\gamma)^{1+\frac{1}{1-\alpha}}} + \frac{1}{(1-\gamma)^2}\sqrt{\sum_{j=1}^T \frac{1}{j^{2(1-\alpha)}}}\right] + \frac{1}{(1-\gamma)}\sqrt{\frac{1}{T\widetilde{\eta}_T}}. & \text{(S2)} \end{cases}$$

C.2.2 Establishing the Functional CLT

Uniform negligibility of ψ_0. It is clear that ψ_0 is a deterministic function. Using the uniform boundedness of $\boldsymbol{A}_j^T (T \ge j \ge 0)$ in Lemma C.2, we have

$$\begin{aligned} \|\boldsymbol{\psi}_0\|_{\sup} = \sup_{r\in[0,1]} \|\boldsymbol{\psi}_0(r)\|_\infty &= \frac{1}{\eta_0\sqrt{T}} \sup_{r\in[0,1]} \|(\boldsymbol{A}_0^{\lfloor Tr\rfloor} - \eta_0 \boldsymbol{I})\boldsymbol{\Delta}_0\|_\infty \\ &\le \frac{1}{\eta_0\sqrt{T}}\left(\sup_{0\le t\le T}\|\boldsymbol{A}_0^t\|_\infty + \eta_0\right)\|\boldsymbol{\Delta}_0\|_\infty \\ &\le \frac{1}{\eta_0\sqrt{T}}\frac{2C_0}{1-\gamma} \to 0 \text{ as } T \to \infty, \end{aligned}$$

where we use $\eta_0 \le 1 \le C_0$ and $\|\boldsymbol{\Delta}_0\|_\infty \le \frac{1}{1-\gamma}$.

[7] The original Lemma 1 in [Polyak and Juditsky, 1992] uses the ℓ_2-norm and spectral norm. Due to the equivalence between these norms, we formulate our Lemma C.2.

Partial-sum asymptotic behavior of ψ_1. Recall that $\boldsymbol{Z}_j = (\boldsymbol{r}_j - \boldsymbol{r}) + \gamma(\boldsymbol{P}_j - \boldsymbol{P})\boldsymbol{V}^*$ is the noise inherent in the empirical Bellman operator at iteration j. Since at each iteration the simulator generates rewards $\boldsymbol{r}_j$ and produces the empirical transition $\boldsymbol{P}_j$ in an i.i.d. fashion, $\mathcal{T}_1(r)$ is the scaled partial sum of $\lfloor Tr \rfloor$ independent copies of the random vector $\boldsymbol{Z}_j$ which has zero mean and finite variance denoted by $\mathrm{Var}(\boldsymbol{Z}_j) = \mathrm{Var}(\boldsymbol{r}_j + \gamma \boldsymbol{P}_j \boldsymbol{V}^*) = \mathbb{E}\boldsymbol{Z}_j\boldsymbol{Z}_j^\top$. Additionally, it is clear that $\|\boldsymbol{Z}_j\|_\infty \le (1-\gamma)^{-1}$ is uniformly bounded and thus its moments of any order is uniformly bounded. By Theorem 4.2 in [Hall and Heyde, 2014] (or Theorem 2.2 in [Jirak, 2017]), we establish the following FCLT for the partial sums of independent random vectors.

Lemma C.5. *For any* $r \in [0,1]$,

$$\boldsymbol{\psi}_1(\cdot) = \frac{1}{\sqrt{T}} \sum_{j=1}^{\lfloor T\cdot \rfloor} \boldsymbol{G}^{-1}\boldsymbol{Z}_j \overset{w}{\to} \mathrm{Var}_{\boldsymbol{Q}}^{1/2} \boldsymbol{B}_D(\cdot),$$

where $\boldsymbol{B}_D$ *is the* D*-dimensional standard Brownian motion and the variance matrix* $\mathrm{Var}_{\boldsymbol{Q}}$ *is*

$$\mathrm{Var}_{\boldsymbol{Q}} = \boldsymbol{G}^{-1}\mathrm{Var}(\boldsymbol{Z}_j)\boldsymbol{G}^{-\top} = (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\mathrm{Var}(\boldsymbol{Z}_j)(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-\top}.$$

Uniform negligibility of ψ_2. Recall that $\psi_2(r) = \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr \rfloor}(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j$. If we define $\boldsymbol{X}_t = \frac{1}{\sqrt{T}}\sum_{j=1}^{t}(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j$, then $\psi_2(r) = \boldsymbol{X}_{\lfloor Tr \rfloor}$. Let $\mathcal{F}_t = \sigma(\{\boldsymbol{r}_j, \boldsymbol{P}_j\}_{0 \le j \le t})$ be the σ-field generated by all randomness before and including iteration t. Then $\{\boldsymbol{X}_t, \mathcal{F}_t\}$ is a martingale since $\mathbb{E}[\boldsymbol{X}_t|\mathcal{F}_{t-1}] = \boldsymbol{X}_{t-1}$. As a result $\{\|\boldsymbol{X}_t\|_2, \mathcal{F}_t\}$ is a submartingale since by conditional Jensen's inequality, we have $\mathbb{E}[\|\boldsymbol{X}_t\|_2|\mathcal{F}_{t-1}] \ge \|\mathbb{E}[\boldsymbol{X}_t|\mathcal{F}_{t-1}]\|_2 = \|\boldsymbol{X}_{t-1}\|_2$. By Doob's maximum inequality for submartingales (which we use to derive the following $(*)$ inequality),

$$\begin{aligned}
\mathbb{E} \sup_{r\in[0,1]} \|\boldsymbol{\psi}_2(r)\|_2^2 &= \mathbb{E} \sup_{0 \le t \le T} \|\boldsymbol{X}_t\|_2^2 \overset{(*)}{\le} 4\mathbb{E}\|\boldsymbol{X}_T\|_2^2 \\
&= 4\mathbb{E}\|\mathcal{T}_2(1)\|_2^2 = 4\mathbb{E}\left\| \frac{1}{\sqrt{T}} \sum_{j=1}^{T} (\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j \right\|_2^2 \\
&= \frac{4}{T}\sum_{j=1}^{T} \mathbb{E}\|(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j\|_2^2 \le \frac{4}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_2^2\mathbb{E}\|\boldsymbol{Z}_j\|_2^2 \\
&\le c_1 \cdot \frac{1}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_2.
\end{aligned}$$

Here, we change to the ℓ_2-norm since it will facilitate the analysis. The last inequality follows by using a finite c_1 satisfying $\mathbb{E}\|\boldsymbol{Z}_j\|_2^2 \sup_{T \ge j \ge 1}\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_2 \le c_1$. Indeed, we can set $c_1 = (\frac{1}{1-\gamma} + \sup_{T\ge j}\|\boldsymbol{A}_j^T\|_2)\mathrm{tr}(\mathrm{Var}_{\boldsymbol{Q}})$ thanks to Lemma C.2. In addition, Lemma C.2 implies $\frac{1}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_2 \to 0$ as T goes to infinity. As a result, $\mathbb{E}\|\boldsymbol{\psi}_2\|_{\sup} = \mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_2(r)\|_\infty \le \mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_2(r)\|_2 \le \sqrt{\mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_2(r)\|_2^2} = o(1)$.

Uniform negligibility of ψ_3. Recall that $\psi_3(r) = \frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr \rfloor}\boldsymbol{A}_j^T(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)$. By a similar argument in the analysis of $\boldsymbol{\psi}_2$, we have $\mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_3(r)\|_2^2 \le 4\mathbb{E}\|\boldsymbol{\psi}_3(1)\|_2^2$ by Doob's maximum inequality. Therefore,

$$\begin{aligned}
\mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_3(r)\|_2^2 \le 4\mathbb{E}\|\boldsymbol{\psi}_3(1)\|_2^2 &\overset{(a)}{=} \frac{4}{T}\sum_{j=1}^{T}\mathbb{E}\|\boldsymbol{A}_j^T(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)\|_2^2 \\
&\le \frac{4}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T\|_2^2\mathbb{E}\|\boldsymbol{P}_j - \boldsymbol{P}\|_2^2\|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_2^2 \\
&\overset{(b)}{\le} c_2 \cdot \frac{1}{T}\sum_{j=1}^{T}\mathbb{E}\|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_2^2,
\end{aligned}$$

where (a) follows since all cross terms have zero mean due to $\mathbb{E}[(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)|\mathcal{F}_{j-1}] = 0$, and (b) follows by setting $c_2 = 16D(\sup_{T\ge j}\|\boldsymbol{A}_j^T\|_2)^2$ because of the uniform boundedness of $\|\boldsymbol{A}_j^T\|_\infty$ from Lemma C.2 and $\|\boldsymbol{P}_j - \boldsymbol{P}\|_2^2 \le D\|\boldsymbol{P}_j - \boldsymbol{P}\|_\infty^2 = 4D$. By Theorem E.4, we know $\frac{1}{T}\sum_{j=1}^{T}\mathbb{E}\|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_2^2 \to 0$ under the general step size when $T \to \infty$. As a result, $\mathbb{E}\|\boldsymbol{\psi}_3(r)\|_{\sup} = \mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_3(r)\|_\infty \le \mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_3(r)\|_2 \le \sqrt{\mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_3(r)\|_2^2} = o(1)$.

Uniform negligibility of ψ_4. Recall that $\psi_4(r) = \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr \rfloor}(\boldsymbol{A}_j^{\lfloor Tr \rfloor} - \boldsymbol{A}_j^T)\boldsymbol{\varepsilon}_j$ where $\boldsymbol{\varepsilon}_j = \boldsymbol{Z}_j + \gamma(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)$. It is clear that we have $\sup_{j\geq 0}\mathbb{E}\|\boldsymbol{Q}_t - \boldsymbol{Q}^*\|^4 < \infty$ as a result of $\sup_{j\geq 0}\mathbb{E}\mathbb{E}\|\boldsymbol{Q}_j\_\infty^4 < \infty$ in Lemma C.6. Notice that the coefficient $\boldsymbol{A}_j^{\lfloor Tr \rfloor} - \boldsymbol{A}_j^T$ changes as r varies. The analysis of ψ_4 should be more careful and subtle.

Lemma C.6 (Moment bounds)**.** *Under Assumption 3.1, it follows that*

$$\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{Q}_t\|_\infty^4 < \infty.$$

Proof of Lemma C.6. By Lemma E.2, $\|\boldsymbol{\Delta}_t\|_\infty \leq a_t + b_t + \|\boldsymbol{N}_t\|_\infty$. It implies that $\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^4 \leq 3^3\mathbb{E}\left(a_t^4 + b_t^4 + \|\boldsymbol{N}_t\|_\infty^4\right)$. Notice that

$$\begin{aligned} a_t &= (1-\eta_t(1-\gamma))a_{t-1} \\ b_t &= (1-\eta_t(1-\gamma))b_{t-1} + \eta_t\gamma\|\boldsymbol{N}_{t-1}\|_\infty \\ \boldsymbol{N}_t &= (1-\eta_t)\boldsymbol{N}_{t-1} + \eta_t\boldsymbol{Z}_t. \end{aligned}$$

First, it is easy to find that $\sup_{t\geq 0} a_t < \infty$ since it is deterministic and decays exponentially fast. Second, we have $\sup_{t\geq 0}\|\boldsymbol{N}_t\|_\infty^4 < \infty$. This is because we have $\mathbb{E}\|\boldsymbol{N}_t\|_\infty^4 \leq (1-\eta_t)\mathbb{E}\|\boldsymbol{N}_{t-1}\|_\infty^4 + \eta_t\mathbb{E}\|\boldsymbol{Z}_t\|_\infty^4$ from Jensen's inequality. It is easy to show $\sup_{t\geq 0}\|\boldsymbol{N}_t\|_\infty^4 < \sup_{t\geq 0}\mathbb{E}\|\boldsymbol{Z}_t\|_\infty^4 < \infty$ by this inequality and induction. Finally, iterating the expression of b_t, we have $b_T = \gamma\sum_{t=1}^T\prod_{j=t+1}^T(1-(1-\gamma)\eta_j)\eta_t\|\boldsymbol{N}_{t-1}\|_\infty = \frac{\gamma}{1-\gamma}\sum_{t=1}^T\widetilde{\eta}_{(t,T)}\|\boldsymbol{N}_{t-1}\|_\infty$ with $\widetilde{\eta}_{(t,T)}$ a probability defined on $[T]$ in (59). The last equation implies b_T is a probability weighted sum of $\boldsymbol{N}_t (t \in [T])$. Hence, by Jensen's inequality, we know $\sup_{t\geq 0}\mathbb{E}b_t^4 < \sup_{t\geq 0}\mathbb{E}\|\boldsymbol{Z}_t\|_\infty^4 < \infty$. □

Recall $\mathcal{F}_t = \sigma(\{\boldsymbol{r}_j, \boldsymbol{P}_j\}_{0\leq j\leq t})$ is the σ-field generated by all randomness before and including iteration t. $\{\boldsymbol{\varepsilon}_t, \mathcal{F}_t\}$ is a martingale difference since $\mathbb{E}[\boldsymbol{\varepsilon}_t|\mathcal{F}_{t-1}] = \boldsymbol{0}$. Furthermore, $\boldsymbol{\varepsilon}_t$ has finite moments of any order since it is almost surely bounded $\|\boldsymbol{\varepsilon}_t\|_\infty = \mathcal{O}((1-\gamma)^{-1})$. On the other hand, by definition (37), it follows that for any $0 \leq k \leq T$,

$$\begin{aligned} \sum_{j=1}^k(\boldsymbol{A}_j^T - \boldsymbol{A}_j^k)\boldsymbol{\varepsilon}_j &= \sum_{j=1}^k\sum_{t=k+1}^T\left(\prod_{i=j+1}^t\boldsymbol{A}_i\right)\eta_j\boldsymbol{\varepsilon}_j = \sum_{t=k+1}^T\sum_{j=1}^k\left(\prod_{i=j+1}^t\boldsymbol{A}_i\right)\eta_j\boldsymbol{\varepsilon}_j \\ &= \sum_{t=k+1}^T\left(\prod_{i=k+1}^t\boldsymbol{A}_i\right)\sum_{j=1}^k\left(\prod_{i=j+1}^k\boldsymbol{A}_i\right)\eta_j\boldsymbol{\varepsilon}_j \\ &= \frac{1}{\eta_{k+1}}\boldsymbol{A}_{k+1}^T\boldsymbol{A}_{k+1}\sum_{j=1}^k\left(\prod_{i=j+1}^k\boldsymbol{A}_i\right)\eta_j\boldsymbol{\varepsilon}_j \end{aligned}$$

On one hand, $\|\boldsymbol{A}_{k+1}^T\boldsymbol{A}_{k+1}\|_2 \leq c_3$ is uniformly bounded with $c_3 = (\sup_{T\geq j}\|\boldsymbol{A}_j^T\|_2)(1+\|\boldsymbol{G}\|_2)$ for any $T \geq k+1$ from Lemma C.2. On the other hand, we define an auxiliary sequence $\{\boldsymbol{Y}_k\}_{k\geq 1}$ as following: $\boldsymbol{Y}_1 = \boldsymbol{0}$ and $\boldsymbol{Y}_{k+1} = \boldsymbol{A}_k\boldsymbol{Y}_k + \eta_k\boldsymbol{\varepsilon}_k$ for any $k \geq 1$. One can check that $\boldsymbol{Y}_{k+1} = \sum_{j=1}^k\left(\prod_{i=j+1}^k\boldsymbol{A}_i\right)\eta_j\boldsymbol{\varepsilon}_j$ where we use the convention $\prod_{i=k+1}^k\boldsymbol{A}_i = \boldsymbol{I}$ for any $k \geq 0$. These results imply we can apply Lemma D.1. Putting these pieces together, we have that

$$\begin{aligned} \|\boldsymbol{\psi}_4\|_{\sup} \leq \sup_{r\in[0,1]}\|\boldsymbol{\psi}_4(r)\|_2 &\leq c_3\sup_{0\leq k\leq T}\left\|\frac{1}{\sqrt{T}\eta_{k+1}}\sum_{j=1}^k\left(\prod_{i=j+1}^k\boldsymbol{A}_i\right)\eta_j\boldsymbol{\varepsilon}_j\right\|_2 \\ &= c_3\sup_{0\leq k\leq T}\frac{1}{\sqrt{T}}\frac{\|\boldsymbol{Y}_{k+1}\|_2}{\eta_{k+1}} \overset{(*)}{=} o_{\mathbb{P}}(1), \end{aligned}$$

where $(*)$ follows from Lemma D.1.

Uniform negligibility of ψ_5. In the following, we will prove $\|\boldsymbol{\psi}_5\|_{\sup} = o_{\mathbb{P}}(1)$ by showing $\mathbb{E}\|\boldsymbol{\psi}_5\|_{\sup} = o(1)$. It is worth mentioning that $\boldsymbol{\psi}_5$ arises purely due to the non-stationary nature of Q-learning. If we consider a stationary update process, e.g., policy evaluation [Mou et al., 2020a,b, Khamaru et al., 2021b], π_t would remain the same all the time and $\boldsymbol{\psi}_5$ would disappear in the case. Notice that $\boldsymbol{\psi}_5(r) = \frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr \rfloor}\boldsymbol{A}_j^{\lfloor Tr \rfloor}(\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{j-1}$ is a sum of correlated random variables (which are even not mean-zero). We need a high-order residual condition Assumption 3.2 to bound $\mathbb{E}\|\boldsymbol{\psi}_5\|_{\sup}$. With such a

Lipschitz condition, Lemma C.7 shows $\mathbb{E}\|\psi_5\|_{\sup}$ is dominated by $\frac{1}{\sqrt{T}}\sum_{j=1}^T \mathbb{E}\left\|\Delta_{j-1}\right\|_\infty^2$, which is $o(1)$ for the general step size as suggested by Theorem E.1. The proof of Lemma C.7 is in Appendix C.4.4.

Lemma C.7. *It follows that*

$$\mathbb{E}\|\psi_5\|_{\sup} = \mathbb{E}\sup_{r\in[0,1]}\|\psi_5(r)\|_\infty \le \gamma L C_0 \cdot \frac{1}{\sqrt{T}}\sum_{j=1}^T \mathbb{E}\left\|\Delta_{j-1}\right\|_\infty^2.$$

Putting the pieces together. From (36), $\phi_T^1 = \sum_{i=0}^5 \psi_i$. We have shown $\psi_1 \overset{w}{\to} \mathrm{Var}_{Q}^{1/2}B_D$ in the sense of $(\mathsf{D}([0,1],\mathbb{R}^D),d_0)$ and $\|\psi_i\|_{\sup} = o_{\mathbb{P}}(1)$ for $i\neq 1$. Using $\|\phi_T^1-\psi_1\|_{\sup}\le\sum_{i\neq 1}\|\psi_i\|_{\sup}$, we know that $\|\phi_T^1-\psi_1\|_{\sup} = o_{\mathbb{P}}(1)$. Proposition C.1 implies $\phi_T^1 \overset{w}{\to} \mathrm{Var}_{Q}^{1/2}B_D$. We then establish the FCLT for $\phi_T^1(r)$.

C.3 Functional CLT for ϕ_T^2

We can repeat the above analysis for ϕ_T^2. We rewrite (29) as

$$\Delta_t^2 = A_t\Delta_{t-1}^2 + \eta_t\left(Z_t+\gamma D_{t-1}^2\right), \tag{39}$$

where $A_t = I - \eta_t(I-\gamma P^{\pi^*})$ and $Z_t = (r_t - r) + \gamma(P_t - P)V^*$ are the same as those defined in (34) except that D_{t-1}^1 (defined in (35)) is replaced by

$$D_{t-1}^2 = (P_t - P)(V_{t-1}-V^*). \tag{40}$$

Since D_{t-1}^2 is much simpler than D_{t-1}^1, the analysis for $\phi_T^2(r)$ should be easier than $\phi_T^1(r)$. Using the notation A_j^T (see(37)), we decompose $\phi_T^2(r)$ into five terms:

$$\begin{aligned}
\phi_T^2(r) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\Delta_t^2 = \frac{1}{\eta_0\sqrt{T}}(A_0^{\lfloor Tr\rfloor}-\eta_0 I)\Delta_0 + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}A_j^{\lfloor Tr\rfloor}\left(Z_j+\gamma D_{j-1}^2\right)\\
&= \frac{1}{\eta_0\sqrt{T}}(A_0^{\lfloor Tr\rfloor}-\eta_0 I)\Delta_0 + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}G^{-1}Z_j + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}(A_j^T - G^{-1})Z_j\\
&\quad + \frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}A_j^T(P_j-P)(V_{j-1}-V^*)\\
&\quad + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}(A_j^{\lfloor Tr\rfloor}-A_j^T)\left[Z_j+\gamma(P_j-P)(V_{j-1}-V^*)\right]\\
&:= \psi_0(r)+\psi_1(r)+\psi_2(r)+\psi_3(r)+\psi_4(r).
\end{aligned} \tag{41}$$

Here $\{\psi_i\}_{i=0}^4$ are exactly the same as those in (38). Our previous analysis provides us a low-hanging fruit result: $\psi_1 \overset{w}{\to} \mathrm{Var}_{Q}^{1/2}B_D$ in the sense of $(\mathsf{D}([0,1],\mathbb{R}^D),d_0)$ and $\|\psi_i\|_{\sup} = o_{\mathbb{P}}(1)$ for $i\neq 1$. Then we know that $\|\phi_T^2 - \mathcal{T}_1\|_{\sup} = o(1)$ and $\phi_T^2 \overset{w}{\to} \mathrm{Var}_{Q}^{1/2}B_D$ due to Proposition C.1. We thus establish the FCLT for ϕ_T^2.

C.4 Proofs of Lemmas

C.4.1 Proof of Lemma C.1

Proof of Lemma C.1. We use mathematical induction to prove the statement. When $t=0$, the inequality (30) holds by initialization. Assume (30) holds at $t-1$, i.e., $\Delta_{t-1}^2 \le \Delta_{t-1}\le\Delta_{t-1}^1$. Let us analyze the case of t. By the Q-learning update rule, it follows that

$$\begin{aligned}
\Delta_t &= (1-\eta_t)\Delta_{t-1} + \eta_t\left[(r_t - r) + \gamma(P_tV_{t-1} - PV^*)\right]\\
&\overset{(a)}{=} (1-\eta_t)\Delta_{t-1} + \eta_t\left[W_t + \gamma(PV_{t-1}-PV^*)\right]\\
&\overset{(b)}{=} (1-\eta_t)\Delta_{t-1} + \eta_t\left[W_t + \gamma(P^{\pi_{t-1}}Q_{t-1} - P^{\pi^*}Q^*)\right]\\
&\overset{(c)}{=} A_t\Delta_{t-1} + \eta_t\left[W_t + \gamma(P^{\pi_{t-1}} - P^{\pi^*})Q_{t-1}\right],
\end{aligned} \tag{42}$$

where (a) uses $\boldsymbol{W}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}$; (b) uses $\boldsymbol{P}\boldsymbol{V}_{t-1} = \boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1}$ and $\boldsymbol{P}\boldsymbol{V}^* = \boldsymbol{P}^{\pi^*}\boldsymbol{Q}^*$, and (c) follows by arrangement and the shorthand $\boldsymbol{A}_t = \boldsymbol{I} - \eta_t(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})$. Since all the entries of $\boldsymbol{A}_t = \boldsymbol{I} - \eta_t(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})$ are non-negative (which results from the assumption $\sup_t \eta_t \le 1$), then $\boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^2 \le \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1} \le \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^1$.

For one hand, based on (42), we have

$$\begin{aligned}
\boldsymbol{\Delta}_t^2 &= \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^2 + \eta_t\boldsymbol{W}_t \le \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1} + \eta_t\boldsymbol{W}_t \\
&\le \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1} + \eta_t\left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{Q}_{t-1}\right] = \boldsymbol{\Delta}_t,
\end{aligned}$$

where the last inequality uses $\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} \ge \boldsymbol{P}^{\pi^*}\boldsymbol{Q}_{t-1}$ which results from the fact π_{t-1} is the greedy policy with respect to $\boldsymbol{Q}_{t-1}$. For the other hand, it follows that

$$\begin{aligned}
\boldsymbol{\Delta}_t &= \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1} + \eta_t\left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{Q}_{t-1}\right] \\
&\le \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^1 + \eta_t\left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{Q}_{t-1}\right] \\
&= \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^1 + \eta_t\left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{t-1} + \gamma(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{Q}^*\right] \\
&\le \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^1 + \eta_t\left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{t-1}\right] = \boldsymbol{\Delta}_t^1,
\end{aligned}$$

where the last inequality uses $\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}^* \le \boldsymbol{P}^{\pi^*}\boldsymbol{Q}^*$ which results from the fact π^* is the greedy policy with respect to $\boldsymbol{Q}^*$. Hence, we have proved $\boldsymbol{\Delta}_t^2 \le \boldsymbol{\Delta}_t \le \boldsymbol{\Delta}_t^1$ holds at iteration t. □

C.4.2 Proof of Lemma C.3

Proof of Lemma C.3. By the definition of (37), we have $\|\boldsymbol{A}_j^T\|_\infty \le \eta_j \sum_{t=j}^T \prod_{i=j+1}^t (1 - \widetilde{\eta}_i)$. Plugging the specific form of $\{\eta_t\}$, we have for (S1)

$$\begin{aligned}
\|\boldsymbol{A}_j^T\|_\infty &\le \eta_j \sum_{t=j}^T \prod_{i=j+1}^t \frac{1 + (1-\gamma)(i-1)}{1 + (1-\gamma)i} = \eta_j \sum_{t=j}^T \frac{1 + (1-\gamma)j}{1 + (1-\gamma)t} \\
&\le \frac{1}{1-\gamma} \ln \frac{1 + (1-\gamma)T}{1 + (1-\gamma)(j-1)} \le \frac{\ln(1 + (1-\gamma)T)}{1-\gamma}
\end{aligned} \tag{43}$$

and for (S2)

$$\begin{aligned}
\|\boldsymbol{A}_j^T\|_\infty &= \eta_j \sum_{t=j}^T \prod_{i=j+1}^t (1 - (1-\gamma)i^{-\alpha}) \le \eta_j \sum_{t=j}^T \exp\left(-(1-\gamma)\sum_{i=j+1}^t i^{-\alpha}\right) \\
&\overset{(a)}{\le} e\eta_j \sum_{t=j+1}^{T+1} \exp\left(-\frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - j^{1-\alpha}\right)\right) \\
&\le e\eta_j \int_j^\infty \exp\left(-\frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - j^{1-\alpha}\right)\right) dt \\
&\overset{(b)}{\le} \frac{e\eta_j}{1-\gamma} \int_0^\infty \left(\frac{1-\alpha}{1-\gamma}y + j^{1-\alpha}\right)^{\frac{\alpha}{1-\alpha}} \exp(-y)\, dy \\
&\overset{(c)}{\le} \frac{e\eta_j}{1-\gamma} \max\left\{2^{\frac{\alpha}{1-\alpha}}, 1\right\} \int_0^\infty \left[\left(\frac{1-\alpha}{1-\gamma}y\right)^{\frac{\alpha}{1-\alpha}} + j^\alpha\right] \exp(-y)\, dy \\
&= \frac{e}{(1-\gamma)j^\alpha} \max\left\{2^{\frac{\alpha}{1-\alpha}}, 1\right\} \left[\left(\frac{1-\alpha}{1-\gamma}\right)^{\frac{\alpha}{1-\alpha}} \Gamma\left(\frac{1}{1-\alpha}\right) + j^\alpha\right] \\
&\overset{(d)}{\le} e \max\left\{2^{\frac{\alpha}{1-\alpha}}, 1\right\} \left[\frac{\sqrt{2\pi e}}{\sqrt{1-\alpha}(1-\gamma)^{\frac{1}{1-\alpha}}} + \frac{1}{1-\gamma}\right] \\
&\le \frac{c2^{\frac{1}{1-\alpha}}}{\sqrt{1-\alpha}} \frac{1}{(1-\gamma)^{\frac{1}{1-\alpha}}},
\end{aligned}$$

where (a) uses $\sum_{i=j}^{t} i^{-\alpha} \geq \frac{1}{1-\alpha}((t+1)^{1-\alpha} - j^{1-\alpha})$ and $\exp((1-\gamma)j^{-\alpha}) \leq e$, (b) uses the change of variable $y = \frac{1-\gamma}{1-\alpha}(t^{1-\alpha} - j^{1-\alpha})$, (c) uses $(a+b)^p \leq \max\{2^{p-1}, 1\}(a^p + b^p)$ for any $p > 0$, and (d) uses $(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right) \leq \frac{\sqrt{2\pi}e^{1/2}}{\sqrt{1-\alpha}}$ from (65) and $\max\left\{2^{\frac{\alpha}{1-\alpha}}, 1\right\} \leq 2^{\frac{1}{1-\alpha}}$. □

C.4.3 Proof of Lemma C.4

Proof of Lemma C.4. For (S1), we have

$$\begin{aligned}
\frac{1}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_\infty^2 &\leq \frac{2}{T}\sum_{j=1}^{T}(\|\boldsymbol{A}_j^T\|_\infty^2 + \|\boldsymbol{G}^{-1}\|_\infty^2) \\
&\leq 2 + \frac{8}{(1-\gamma)^2}\frac{1}{T}\sum_{j=1}^{T}\ln^2\frac{1+(1-\gamma)T}{1+(1-\gamma)(j-1)} \\
&\leq 2 + \frac{8}{(1-\gamma)^2}\left[\frac{\ln^2(1+(1-\gamma)T)}{T} + \frac{1}{T}\sum_{j=1}^{T-1}\ln^2\frac{T}{j}\right] \\
&\overset{(a)}{\leq} 2 + \frac{7}{1-\gamma} + \frac{16}{(1-\gamma)^2} \leq \frac{25}{(1-\gamma)^2},
\end{aligned}$$

where (a) uses $\ln^2(1+x)/x \leq \frac{7}{8}$ for all $x \geq 0$ and $\int_0^1 \ln^2 x dx = \Gamma(3) = 2\Gamma(1) = 2$.

For (S2), based on (37) and $\boldsymbol{G} = \eta_j^{-1}(\boldsymbol{I} - (\boldsymbol{I} - \eta_j\boldsymbol{G}))$, we have

$$\begin{aligned}
\boldsymbol{A}_j^T - \boldsymbol{G}^{-1} = (\boldsymbol{A}_j^T\boldsymbol{G} - \boldsymbol{I})\boldsymbol{G}^{-1} &= \sum_{t=j}^{T}\left(\prod_{i=j+1}^{t}(\boldsymbol{I} - \eta_i\boldsymbol{G}) - \prod_{i=j}^{t}(\boldsymbol{I} - \eta_i\boldsymbol{G})\right)\boldsymbol{G}^{-1} - \boldsymbol{G}^{-1} \\
&= \sum_{t=j+1}^{T}\left(\prod_{i=j+1}^{t}(\boldsymbol{I} - \eta_i\boldsymbol{G}) - \prod_{i=j}^{t-1}(\boldsymbol{I} - \eta_i\boldsymbol{G})\right)\boldsymbol{G}^{-1} - \prod_{t=j}^{T}(\boldsymbol{I} - \eta_t\boldsymbol{G})\boldsymbol{G}^{-1} \\
&= \sum_{t=j+1}^{T}(\eta_j - \eta_t)\prod_{i=j+1}^{t-1}(\boldsymbol{I} - \eta_i\boldsymbol{G}) - \prod_{t=j}^{T}(\boldsymbol{I} - \eta_t\boldsymbol{G})\boldsymbol{G}^{-1} \\
&:= \boldsymbol{M}_{T,j}^{(1)} + \boldsymbol{M}_{T,j}^{(2)}.
\end{aligned} \tag{44}$$

On the one hand,

$$\left\|\boldsymbol{M}_{T,j}^{(2)}\right\|_\infty \leq \|\boldsymbol{G}^{-1}\|_\infty\prod_{t=j}^{T}\|\boldsymbol{I} - \eta_t\boldsymbol{G}\|_\infty \leq \frac{\prod_{t=j}^{T}(1-\widetilde{\eta}_t)}{1-\gamma} \leq \frac{(1-\widetilde{\eta}_T)^{T-j+1}}{1-\gamma}.$$

On the other hand,

$$
\begin{aligned}
\left\|\boldsymbol{M}_{T,j}^{(1)}\right\|_\infty &= \left\|\sum_{t=j+1}^{T}(\eta_t-\eta_j)\prod_{i=j+1}^{t-1}(\boldsymbol{I}-\eta_i\boldsymbol{G})\right\|_\infty \\
&\le \sum_{t=j+1}^{T}|\eta_t-\eta_j|\exp\left(-\sum_{i=j+1}^{t-1}\widetilde{\eta}_i\right) \\
&\le \sum_{t=j+1}^{T}\sum_{k=j}^{t-1}|\eta_{k+1}-\eta_k|\exp\left(-\sum_{i=j+1}^{t-1}\widetilde{\eta}_i\right) \\
&\overset{(a)}{\le} \sum_{t=j+1}^{T}\sum_{k=j}^{t-1}\frac{\alpha}{k}\eta_k\exp\left(-\sum_{i=j+1}^{t-1}\widetilde{\eta}_i\right) \\
&\overset{(b)}{\le} \frac{e\alpha}{(1-\gamma)j}\sum_{t=j+1}^{T}\widetilde{m}_{j,t-1}\exp\left(-\widetilde{m}_{j,t-1}\right) = \frac{e\alpha}{(1-\gamma)j}\sum_{t=j}^{T-1}\widetilde{m}_{j,t}\exp\left(-\widetilde{m}_{j,t}\right) \\
&\overset{(c)}{\le} \frac{ec\alpha}{(1-\gamma)j}\left[\frac{2^{\frac{1}{1-\alpha}}}{(1-\alpha)^{\frac{3}{2}}}\frac{1}{(1-\gamma)^{\frac{1}{1-\alpha}}}+\frac{2^{\frac{1}{1-\alpha}}}{1-\gamma}(j-1)^{\alpha}\right],
\end{aligned}
$$

where (a) uses the fact that for $\eta_t = t^{-\alpha}$, we have

$$
\frac{\eta_t-\eta_{t+1}}{\eta_t} = 1-\left(1-\frac{1}{t+1}\right)^{\alpha} \le 1-\exp(-\frac{\alpha}{t}) \le \frac{\alpha}{t},
$$

where we use $\ln(1+x)\ge x/(1+x)$ in the first inequality and $\ln(1+x)\le x$ in the second inequality. (b) uses the notation $\widetilde{m}_{j,t} := \sum_{i=j}^{t}\widetilde{\eta}_i$ and $\exp(\widetilde{\eta}_j)\le\exp(1)=e$. (c) uses the following lemma.

Lemma C.8. *Let $\widetilde{m}_{j,t} := \sum_{i=j}^{t}\widetilde{\eta}_i$ and recall $\widetilde{\eta}_i = (1-\gamma)i^{-\alpha}$. Then $T\ge j\ge 1$, for some constant $c>1$,*

$$
\sum_{t=j}^{T}\widetilde{m}_{j,t}\exp\left(-\widetilde{m}_{j,t}\right) \le c\left[\frac{2^{\frac{1}{1-\alpha}}}{(1-\alpha)^{\frac{3}{2}}}\frac{1}{(1-\gamma)^{\frac{1}{1-\alpha}}}+\frac{2^{\frac{1}{1-\alpha}}}{1-\gamma}(j-1)^{\alpha}\right].
$$

Therefore,

$$
\begin{aligned}
\frac{1}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T-\boldsymbol{G}^{-1}\|_\infty^2 &\le \frac{2}{T}\sum_{j=1}^{T}\left[\left\|\boldsymbol{M}_{T,j}^{(1)}\right\|_\infty^2+\left\|\boldsymbol{M}_{T,j}^{(2)}\right\|_\infty^2\right] \\
&\le \frac{2c}{T}\sum_{j=1}^{T}\left[\frac{\alpha^2}{j^2}\frac{2^{\frac{2}{1-\alpha}}}{(1-\alpha)^3}\frac{1}{(1-\gamma)^{2+\frac{2}{1-\alpha}}}+\frac{\alpha^2 2^{\frac{2}{1-\alpha}}}{(1-\gamma)^4}\frac{1}{j^{2(1-\alpha)}}+\frac{(1-\widetilde{\eta}_T)^{2(T-j+1)}}{(1-\gamma)^2}\right] \\
&\le \frac{c\alpha^2 2^{2+\frac{2}{1-\alpha}}}{T}\left[\frac{1}{(1-\alpha)^3}\frac{1}{(1-\gamma)^{2+\frac{2}{1-\alpha}}}+\frac{1}{(1-\gamma)^4}\sum_{j=1}^{T}\frac{1}{j^{2(1-\alpha)}}\right]+\frac{1}{(1-\gamma)^2}\frac{1}{T\widetilde{\eta}_T}.
\end{aligned}
$$

□

Proof of Lemma C.8. Clearly we have

$$
\frac{1-\gamma}{1-\alpha}\left((t+1)^{1-\alpha}-j^{1-\alpha}\right) \le \widetilde{m}_{j,t} = \sum_{i=j}^{t}\widetilde{\eta}_i \le \frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha}-(j-1)^{1-\alpha}\right).
$$

Then $\widetilde{m}_{j,t} \le \frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right) \le \widetilde{m}_{j-1,t-1}$. Hence,

$$\begin{aligned}
\sum_{t=j}^{T} \widetilde{m}_{j,t} \exp\left(-\widetilde{m}_{j,t}\right) &= \sum_{t=j}^{T} \widetilde{m}_{j,t} \exp(-\widetilde{m}_{j-1,t-1}) \exp(\widetilde{\eta}_{j-1} - \widetilde{\eta}_t) \\
&= e \sum_{t=j}^{T} \frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right) \exp\left(-\frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right)\right) \\
&\le 2e \int_{j-1}^{\infty} \frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right) \exp\left(-\frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right)\right) dt \\
&\overset{(a)}{=} \frac{2e}{1-\gamma} \int_{0}^{\infty} y \exp\left(-y\right) \left(\frac{1-\alpha}{1-\gamma} y + (j-1)^{1-\alpha}\right)^{\frac{\alpha}{1-\alpha}} dt \\
&\overset{(b)}{\le} \frac{e \max\{2^{\frac{\alpha}{1-\alpha}}, 2\}}{1-\gamma} \int_{0}^{\infty} y \exp\left(-y\right) \left[\left(\frac{1-\alpha}{1-\gamma} y\right)^{\frac{\alpha}{1-\alpha}} + (j-1)^{\alpha}\right] dt \\
&\overset{(c)}{\le} \frac{e 2^{\frac{1}{1-\alpha}}}{1-\gamma} \left[\left(\frac{1-\alpha}{1-\gamma}\right)^{\frac{\alpha}{1-\alpha}} \Gamma\left(1 + \frac{1}{1-\alpha}\right) + (j-1)^{\alpha}\right] \\
&\overset{(d)}{\le} \frac{\sqrt{2\pi} e^{\frac{3}{2}} 2^{\frac{1}{1-\alpha}}}{(1-\alpha)^{\frac{3}{2}}} \frac{1}{(1-\gamma)^{\frac{1}{1-\alpha}}} + \frac{e 2^{\frac{1}{1-\alpha}}}{1-\gamma} (j-1)^{\alpha},
\end{aligned}$$

where (a) uses the change of variable $y = \frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right)$, (b) uses $(a+b)^p \le \max\{2^{p-1}, 1\}(a^p + b^p)$ for any $p > 0$, (c) uses $\max\left\{2^{\frac{\alpha}{1-\alpha}}, 2\right\} \le 2^{\frac{1}{1-\alpha}}$, (d) uses $\Gamma\left(1 + \frac{1}{1-\alpha}\right) = \frac{1}{1-\alpha}\Gamma\left(\frac{1}{1-\alpha}\right)$ and $(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right) \le \frac{\sqrt{2\pi e}}{\sqrt{1-\alpha}}$ from (65). □

C.4.4 Proof of Lemma C.7

Proof of Lemma C.7. By Lemma B.1 and Lemma C.3, it follows that

$$\begin{aligned}
\mathbb{E}\|\mathcal{T}_5\|_{\sup} = \mathbb{E} \sup_{r\in[0,1]} \|\mathcal{T}_5(r)\|_\infty &\le \frac{\gamma}{\sqrt{T}} \mathbb{E} \sup_{r\in[0,1]} \sum_{j=1}^{\lfloor Tr \rfloor} \left\| \boldsymbol{A}_j^{\lfloor Tr \rfloor} (\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*}) \boldsymbol{\Delta}_{j-1} \right\|_\infty \\
&\le \frac{\gamma}{\sqrt{T}} \mathbb{E} \sum_{j=1}^{T} \sup_{r\in[0,1]} \left\| \boldsymbol{A}_j^{\lfloor Tr \rfloor} \right\|_\infty \left\| (\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*}) \boldsymbol{\Delta}_{j-1} \right\|_\infty \\
&\le \gamma L C_0 \cdot \frac{1}{\sqrt{T}} \mathbb{E} \sum_{j=1}^{T} \|\boldsymbol{\Delta}_{j-1}\|_\infty^2 .
\end{aligned}$$

Here we use $\sup_{r\in[0,1]} \left\| \boldsymbol{A}_j^{\lfloor Tr \rfloor} \right\|_\infty \le C_0$ due to Lemma C.2. □

D UNIFORM NEGLIGIBILITY OF NOISE RECURSION

Definition D.1 (Hurwitz matrix)**.** *We say $-\boldsymbol{G} \in \mathbb{R}^{d\times d}$ is a Hurwitz (or stable) matrix if $\mathrm{Re}\lambda_i(\boldsymbol{G}) > 0$ for $i \in [d]$. Here $\lambda_i(\cdot)$ denotes the i-th eigenvalue.*

Lemma D.1 (A generalization of Lemma B.7 in [Li et al., 2022])**.** *Let $\{\varepsilon_t\}_{t\ge 0}$ be a martingale difference sequence adapting to the filtration $\mathcal{F}_t$. Define an auxiliary sequence $\{\boldsymbol{y}_t\}_{t\ge 0}$ as following: $\boldsymbol{y}_0 = \boldsymbol{0}$ and for $t \ge 0$,*

$$\boldsymbol{y}_{t+1} = (\boldsymbol{I} - \eta_t \boldsymbol{G})\boldsymbol{y}_t + \eta_t \boldsymbol{\epsilon}_t. \tag{45}$$

It is easy to verify that

$$\boldsymbol{y}_{t+1} = \sum_{j=0}^{t} \left(\prod_{i=j+1}^{t} (\boldsymbol{I} - \eta_i \boldsymbol{G})\right) \eta_j \boldsymbol{\epsilon}_j. \tag{46}$$

Let $\{\eta_t\}_{t\geq 0}$ satisfy Assumption 3.3. If $-\boldsymbol{G} \in \mathbb{R}^{d\times d}$ is Hurwitz, and $\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^4 < \infty$, then we have that

$$\frac{1}{\sqrt{T}} \sup_{0\leq t\leq T} \frac{\|\boldsymbol{y}_{t+1}\|}{\eta_{t+1}} \xrightarrow{p} 0.$$

Proof of Lemma D.1. In the sequel, we denote $\check{\boldsymbol{y}}_t = \frac{\boldsymbol{y}_t}{\sqrt{\eta_{t-1}}}$. We will also use $a \precsim b$ to denote $a \leq Cb$ for unimportant positive constants C with the specific value of C changing according to the context. Then the update rule (45) can be rewritten as

$$\begin{aligned} \check{\boldsymbol{y}}_{t+1} &= \frac{\boldsymbol{y}_{t+1}}{\sqrt{\eta_t}} = \frac{1}{\sqrt{\eta_t}}(\boldsymbol{y}_t - \eta_t \boldsymbol{G}\boldsymbol{y}_t + \eta_t \boldsymbol{\epsilon}_t) \\ &= \check{\boldsymbol{y}}_t + \left(\sqrt{\frac{\eta_{t-1}}{\eta_t}} - 1\right)\check{\boldsymbol{y}}_t - \sqrt{\eta_t\eta_{t-1}}\boldsymbol{G}\check{\boldsymbol{y}}_t + \sqrt{\eta_t}\boldsymbol{\epsilon}_t. \end{aligned} \tag{47}$$

Step 1: Divide the time interval. For a specific $\lambda > 0$, we divide the the time interval $[0, T]$ into several disjoint portions with the t_k the k-th endpoint such that $\sum_{t_k}^{t_{k+1}-1} \eta_s \geq \lambda$. In particular, $\{t_k\}_{k\geq 0}$ is defined iteratively by $t_0 = 0$ and

$$t_{k+1} = \min\left\{n : \sum_{s=t_k}^{n-1} \eta_s \geq \lambda\right\} \wedge \{T\}.$$

Clearly, K is the number of portions and we have $0 = t_0 < t_1 < \cdots < t_K = T$. Since $\sum_{t=1}^{\infty} \eta_t = \infty$, we know that $K \to \infty$ as $T \to \infty$ What's more, K is upper bounded by $\frac{1}{\lambda}\sum_{t=0}^{T} \eta_t$ due to

$$\sum_{t=0}^{T-1} \eta_t = \sum_{k=0}^{K-1}\sum_{s=t_k}^{t_{k+1}-1} \eta_s \geq \lambda K. \tag{48}$$

The fact $\sup_{t\leq T} \frac{\|\boldsymbol{y}_t\|}{\eta_{t-1}} \leq \sup_{t\leq T} \frac{\|\check{\boldsymbol{y}}_t\|}{\sqrt{\eta_T}}$ implies we have for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{\sqrt{T}}\sup_{t\geq T}\frac{\|\boldsymbol{y}_t\|}{\eta_{t-1}} > \epsilon\right) \leq \mathbb{P}\left(\sup_{t\leq T}\|\check{\boldsymbol{y}}_t\| > \epsilon\sqrt{T\eta_T}\right). \tag{49}$$

Lemma D.2. *Let $\{\boldsymbol{y}_t\}_{t\geq 0}$ be defined in the way of (45). If $-\boldsymbol{G} \in \mathbb{R}^{d\times d}$ is Hurwitz and $\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^p < \infty$ for $p \geq 4$, then the sequence $\{\boldsymbol{y}_t\}_{t\geq 0}$ is $(L^4, \sqrt{\eta_t})$-consistency, that is, there exists a universal constant $C_4 > 0$ such that $\mathbb{E}\|\boldsymbol{y}_t\|^4 \leq C_4\eta_t^2$ for all $t \geq 0$.*

The proof of Lemma D.2 is deferred in Section D.1. Lemma D.2 implies that $\sup_{t\geq 0}\mathbb{E}\|\check{\boldsymbol{y}}_t\|^4 \precsim 1$. Let $\mathcal{B} := \left\{\sup_{1\leq k\leq K}\|\check{\boldsymbol{y}}_{t_k}\| \leq \epsilon\sqrt{T\eta_T}\right\}$ be the event where all $\|\check{\boldsymbol{y}}_{t_k}\|$'s are smaller than $\epsilon\sqrt{T\eta_T}$ for $1 \leq k \leq K$. By the union bound and Markov inequality,

$$\mathbb{P}(\mathcal{B}^c) \leq \sum_{k=1}^{K}\mathbb{P}\left(\|\check{\boldsymbol{y}}_{t_k}\| > \epsilon\sqrt{T\eta_T}\right) \leq \sum_{k=1}^{K}\frac{\mathbb{E}\|\check{\boldsymbol{y}}_{t_k}\|^4}{\epsilon^4(T\eta_T)^2} \leq \sup_{t\geq 0}\mathbb{E}\|\check{\boldsymbol{y}}_t\|^4 \cdot \frac{K}{\epsilon^4(T\eta_T)^2} \precsim \frac{\sum_{t=0}^{T}\eta_t}{\lambda\epsilon^4(T\eta_T)^2} \to 0.$$

Here the last inequality uses (48) and the condition on $\{\eta_t\}_{t\geq 0}$ that $\frac{\sum_{t=0}^{T}\eta_t}{(T\eta_T)^2} \to 0$ due to $\frac{\sum_{t=0}^{T}\eta_t}{T\eta_T} \leq C$ and $\eta_T T \to \infty$. The above result implies for given $\lambda, \epsilon > 0$, the event $\mathcal{B}$ holds with probability approaching one. Hence, we focus our analysis on the event $\mathcal{B}$. Conditioning on the event $\mathcal{B}$, we split our target event into several disjoint events whose probability will be analyzed latter.

$$
\begin{aligned}
\mathbb{P}\left(\sup_{t\le T}\|\check{\boldsymbol{y}}_t\|>3\epsilon\sqrt{T\eta_T}\right) &\le \mathbb{P}(\mathcal{B}^c)+\mathbb{P}\left(\sup_{t\le T}\|\check{\boldsymbol{y}}_t\|>\epsilon\sqrt{T\eta_T};\mathcal{B}\right)\\
&\le \mathbb{P}(\mathcal{B}^c)+\sum_{k=0}^{K-1}\mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}-1]}\|\check{\boldsymbol{y}}_t\|>3\epsilon\sqrt{T\eta_T};\mathcal{B}\right)\\
&\le \mathbb{P}(\mathcal{B}^c)+\sum_{k=0}^{K-1}\mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}-1]}\|\check{\boldsymbol{y}}_t-\check{\boldsymbol{y}}_{t_k}\|>2\epsilon\sqrt{T\eta_T};\mathcal{B}\right)\\
&\le \mathbb{P}(\mathcal{B}^c)+\sum_{k=0}^{K-1}\mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}-1]}\|\check{\boldsymbol{y}}_t-\check{\boldsymbol{y}}_{t_k}\|>2\epsilon\sqrt{T\eta_T}\right)\\
&:= \mathbb{P}(\mathcal{B}^c)+\sum_{k=1}^{K-1}\mathcal{P}_k. \qquad (50)
\end{aligned}
$$

Step 2: Bound each $\mathcal{P}_k$. Leveraging (47) recursively implies for given $r<t$,

$$
\check{\boldsymbol{y}}_t-\check{\boldsymbol{y}}_r=\sum_{s=r}^{t-1}\left\{\left(\sqrt{\frac{\eta_{s-1}}{\eta_s}}-1\right)\check{\boldsymbol{y}}_s-\sqrt{\eta_{s-1}\eta_s}\boldsymbol{G}\check{\boldsymbol{y}}_s+\sqrt{\eta_s}\boldsymbol{\epsilon}_s\right\}.
$$

As a result,

$$
\begin{aligned}
\mathcal{P}_k &= \mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}-1]}\left\|\sum_{s=t_k}^{t-1}\left\{\left(\sqrt{\frac{\eta_{s-1}}{\eta_s}}-1\right)\check{\boldsymbol{y}}_s-\sqrt{\eta_{s-1}\eta_s}\boldsymbol{G}\check{\boldsymbol{y}}_s+\sqrt{\eta_s}\boldsymbol{\epsilon}_s\right\}\right\|>2\epsilon\sqrt{T\eta_T}\right)\\
&\le \mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}-1]}\left\|\sum_{s=t_k}^{t-1}\left\{\left(\sqrt{\frac{\eta_{s-1}}{\eta_s}}-1\right)\check{\boldsymbol{y}}_s-\sqrt{\eta_{s-1}\eta_s}\boldsymbol{G}\check{\boldsymbol{y}}_s\right\}\right\|>\epsilon\sqrt{T\eta_T}\right)\\
&\quad+\mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}-1]}\left\|\sum_{s=t_k}^{t-1}\sqrt{\eta_s}\boldsymbol{\epsilon}_s\right\|>\epsilon\sqrt{T\eta_T}\right)\\
&=: \mathcal{P}_k^{(1)}+\mathcal{P}_k^{(2)}. \qquad (51)
\end{aligned}
$$

In the following, we highlight the dependence on T and λ and use $\precsim$ to omit universal constants.

We consider to bound $\mathcal{P}_k^{(1)}$ first. Since $\frac{\eta_t}{\eta_{t-1}}=1-o(\eta_{t-1})$, we have $\sqrt{\frac{\eta_{t-1}}{\eta_t}}-1=\frac{1}{\sqrt{1-o(\eta_{t-1})}}-1=o(\eta_{t-1})$. Hence, there exists a universal positive $C>0$ such that $\left\|\left(\sqrt{\frac{\eta_{t-1}}{\eta_t}}-1\right)\check{\boldsymbol{y}}_t-\sqrt{\eta_{t-1}\eta_t}\boldsymbol{G}\check{\boldsymbol{y}}_t\right\|\le C\eta_t\|\check{\boldsymbol{y}}_t\|$ for all $t\ge 0$. As a result,

$$
\begin{aligned}
\mathcal{P}_k^{(1)} &= \mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}-1]}\sum_{s=t_k}^{t-1}\left\|\left(\sqrt{\frac{\eta_{s-1}}{\eta_s}}-1\right)\check{\boldsymbol{y}}_s-\sqrt{\eta_{s-1}\eta_s}\boldsymbol{G}\check{\boldsymbol{y}}_s\right\|>\epsilon\sqrt{T\eta_T}\right)\\
&\le \mathbb{P}\left(\sum_{s=t_k}^{t_{k+1}-1}C_0\eta_s\|\check{\boldsymbol{y}}_s\|>\epsilon\sqrt{T\eta_T}\right)\precsim\frac{1}{\epsilon^2T\eta_T}\cdot\mathbb{E}\left(\sum_{s=t_k}^{t_{k+1}-1}\eta_s\|\check{\boldsymbol{y}}_s\|\right)^2\\
&\le \frac{1}{\epsilon^2T\eta_T}\left\{\left(\sum_{s=t_k}^{t_{k+1}-1}\eta_s\right)\left(\sum_{s=t_k}^{t_{k+1}-1}\eta_s\mathbb{E}\|\check{\boldsymbol{y}}_s\|^2\right)\right\}\\
&\le \frac{\sup_{t\ge0}\mathbb{E}\|\check{\boldsymbol{y}}_t\|^2}{\epsilon^2T\eta_T}\left(\sum_{s=t_k}^{t_{k+1}-1}\eta_s\right)^2\precsim\frac{1}{\epsilon^2T\eta_T}\left(\sum_{s=t_k}^{t_{k+1}-1}\eta_s\right)^2. \qquad (52)
\end{aligned}
$$

Let $K_0=\max\{m\ge0:\eta_m\ge\lambda\}$. Since η_t decreases in t and converges to 0, we know K_0 also decreases in λ. If $t_k\le K_0$, we have $t_{k+1}=t_k+1$ and thus $\sum_{s=t_k}^{t_{k+1}-1}\eta_s=\eta_{t_k}\le\eta_0$; otherwise, $\sum_{s=t_k}^{t_{k+1}-1}\eta_s\le 2\lambda$ by definition. Summing over $\mathcal{P}_k^{(1)}$ from

0 to $K-1$ and using (52) yield

$$\begin{aligned}
\sum_{k=0}^{K-1}\mathcal{P}_k^{(1)} &= \sum_{k=0}^{K_0-1}\mathcal{P}_k^{(1)} + \sum_{k=K_0}^{K-1}\mathcal{P}_k^{(1)} \precsim \sum_{k=0}^{K_0-1}\frac{\eta_0^2}{\epsilon^2 T\eta_T} + \sum_{k=K_0}^{K-1}\frac{4\lambda^2}{\epsilon^2 T\eta_T} \\
&\le \frac{1}{\epsilon^2 T\eta_T}\left(K_0\eta_0^2 + 4K\lambda^2\right) \overset{(48)}{\precsim} \frac{1}{\epsilon^2 T\eta_T}\left(K_0 + \lambda\sum_{t=0}^{T}\eta_t\right) \\
&\precsim \frac{K_0}{\epsilon^2 T\eta_T} + \frac{\lambda}{\epsilon^2}.
\end{aligned}$$

The last inequality uses $\sum_{t=0}^{T}\eta_t \le CT\eta_T$ for all $T \ge 1$. For a given λ, letting $T \to \infty$ can make the first term go to zero. Then letting $\lambda \to 0$ make the second term vanish too. Hence, we have

$$\lim_{\lambda\to 0}\lim_{T\to\infty}\sum_{k=0}^{K-1}\mathcal{P}_k^{(1)} = 0.$$

Next, we consider to bound $\mathcal{P}_k^{(2)}$. To than end, we will use the Burkholder inequality which relates a martingale with its quadratic variation.

Lemma D.3 (Burkholder's inequality [Burkholder, 1988])**.** *Fix any $p \ge 2$. For a martingale difference $\{\boldsymbol{x}_t\}_{t\in[T]}$ in a real (or complex) Hilbert space, each with finite L^p-norm, one has*

$$\mathbb{E}\left\|\sum_{t=1}^{T}\boldsymbol{x}_t\right\|^p \le B_p\mathbb{E}\left(\sum_{t=1}^{T}\|\boldsymbol{x}_t\|^2\right)^{\frac{p}{2}}$$

where B_p is a universal positive constant depending only on p.

Hence,

$$\begin{aligned}
\mathcal{P}_k^{(2)} &= \mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}-1]}\left\|\sum_{s=t_k}^{t-1}\sqrt{\eta_s}\boldsymbol{\epsilon}_s\right\| > \epsilon\sqrt{T\eta_T}\right) \\
&\le \frac{1}{\epsilon^4(T\eta_T)^2}\mathbb{E}\sup_{t\in[t_k,t_{k+1}-1]}\left\|\sum_{s=t_k}^{t-1}\sqrt{\eta_s}\boldsymbol{\epsilon}_s\right\|^4 \\
&\overset{(a)}{\precsim} \frac{1}{\epsilon^4(T\eta_T)^2}\mathbb{E}\left(\sum_{s=t_k}^{t_{k+1}-1}\eta_s\|\boldsymbol{\epsilon}_s\|^2\right)^2 \\
&\overset{(b)}{\precsim} \frac{\left(\sum_{s=t_k}^{t_{k+1}-1}\eta_s\right)^2}{\epsilon^4(T\eta_T)^2}\sum_{s=t_k}^{t_{k+1}-1}\frac{\eta_s}{\sum_{l=t_k}^{t_{k+1}-1}\eta_l}\mathbb{E}\|\boldsymbol{\epsilon}_s\|^4 \\
&\overset{(c)}{\precsim} \frac{1}{\epsilon^4(T\eta_T)^2}\left(\sum_{s=t_k}^{t_{k+1}-1}\eta_s\right)^2
\end{aligned}$$

where (a) uses Lemma D.3; (b) uses Jensen's inequality; and (c) uses $\sup_{t\ge 0}\mathbb{E}\|\boldsymbol{\epsilon}_t\|^4 < \infty$. As before, we will discuss two cases depending on whether η_t is larger than λ or not. It is equivalent to whether t_k is greater than K_0. Similar to the argument in bounding $\sum_{k=0}^{K-1} \mathcal{P}_k^{(1)}$, we have

$$\begin{aligned}
\sum_{k=0}^{K-1} \mathcal{P}_k^{(2)} &= \sum_{k=0}^{K_0-1} \mathcal{P}_k^{(2)} + \sum_{k=K_0}^{K-1} \mathcal{P}_k^{(2)} \\
&\precsim \frac{1}{\epsilon^4 (T\eta_T)^2} \left(\sum_{k=0}^{K_0-1} \eta_0^2 + \sum_{k=K_0}^{K-1} 2^{p/2} \lambda^2 \right) \\
&\precsim \frac{1}{\epsilon^4 (T\eta_T)^2} \left(K_0 + K\lambda^2 \right) \\
&\precsim \frac{K_0}{\epsilon^4 (T\eta_T)^2} + \frac{\lambda}{\epsilon^4} \cdot \frac{\sum_{t=0}^{T} \eta_t}{(T\eta_T)^2} \\
&\precsim \frac{K_0}{\epsilon^4 (T\eta_T)^2} + \frac{\lambda}{\epsilon^4} \cdot \frac{C}{T\eta_T}
\end{aligned}$$

where the last inequality uses $\sum_{t=0}^{T} \eta_t \le CT\eta_T$ for all $T \ge 1$. From the last inequality, letting $T \to \infty$ makes these two terms converge to zero. Hence, we have

$$\lim_{\lambda \to 0} \lim_{T \to \infty} \sum_{k=0}^{K-1} \mathcal{P}_k^{(2)} = 0.$$

Step 3: Putting the pieces together. Therefore,

$$\begin{aligned}
\lim_{T\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{T}} \sup_{t \ge T} \frac{\|\boldsymbol{y}_t\|}{\eta_{t-1}} > \epsilon \right) &\overset{(49)}{\le} \lim_{T\to\infty} \mathbb{P}\left(\sup_{t \le T} \|\check{\boldsymbol{y}}_t\| > \epsilon \sqrt{T\eta_T} \right) \\
&\overset{(50)}{\le} \lim_{T\to\infty} \left(\mathbb{P}(\mathcal{B}^c) + \sum_{k=1}^{K-1} \mathcal{P}_k \right) = \lim_{T\to\infty} \sum_{k=1}^{K-1} \mathcal{P}_k \\
&\overset{(51)}{\le} \lim_{T\to\infty} \sum_{k=1}^{K-1} \left(\mathcal{P}_k^{(1)} + \mathcal{P}_k^{(2)} \right).
\end{aligned}$$

Since the probability of the left-hand side has nothing to do with λ, letting $\lambda \to 0$ gives

$$\lim_{T\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{T}} \sup_{t \ge T} \frac{\|\boldsymbol{y}_t\|}{\eta_{t-1}} > \epsilon \right) \le \lim_{\lambda \to 0} \lim_{T\to\infty} \sum_{k=1}^{K-1} \left(\mathcal{P}_k^{(1)} + \mathcal{P}_k^{(2)} \right) = 0.$$

□

D.1 Proof of Lemma D.2

For the proof in the section, we will consider random variables (or matrices) in the complex field $\mathbb{C}$. Hence, we will introduce new notations for them. For a vector $\boldsymbol{v} \in \mathbb{C}$ (or a matrix $\boldsymbol{U} \in \mathbb{C}^{d\times d}$), we use $\boldsymbol{v}^{\mathrm{H}}$ (or $\boldsymbol{U}^{\mathrm{H}}$) to denote its Hermitian transpose or conjugate transpose. For any two vectors $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{C}$, with a slight abuse of notation, we use $\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \boldsymbol{v}^{\mathrm{H}}\boldsymbol{u}$ to denote the inner product in $\mathbb{C}$. For simplicity, for a complex matrix $\boldsymbol{U} \in \mathbb{C}^{d\times d}$, we use $\|\boldsymbol{U}\|$ to denote the its operator norm introduced by the complex inner product $\langle \cdot, \cdot \rangle$. When $\boldsymbol{U} \in \mathbb{R}^{d\times d}$, $\|\boldsymbol{U}\|$ is reduced to the spectrum norm.

Proof of Lemma D.2. By Lemma D.4, $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$ for two non-singular matrices $\boldsymbol{U}, \boldsymbol{D} \in \mathbb{C}^{d\times d}$ that satisfies $2\mu \cdot \boldsymbol{I} \preceq \boldsymbol{D} + \boldsymbol{D}^{\mathrm{H}}$ with $\mu := \min_{i \in [d]} \lambda_i(\boldsymbol{G})$ for simplicity.

Lemma D.4 (Property of Hurwitz matrices, Lemma 1 in [Mou et al., 2020a]). *If $-\boldsymbol{G} \in \mathbb{R}^{d\times d}$ be a Hurwitz matrix (i.e., $\mathrm{Re}\lambda_i(\boldsymbol{G}) > 0$ for all $i \in [d]$), there exists a non-degenerate matrix $\boldsymbol{U} \in \mathbb{C}^{d\times d}$ such that $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$ for some matrix $\boldsymbol{D} \in \mathbb{C}^{d\times d}$ that satisfies*

$$2 \min_{i\in[d]} \lambda_i(\boldsymbol{G}) \cdot \boldsymbol{I} \preceq \boldsymbol{D} + \boldsymbol{D}^{\mathrm{H}}$$

where $\boldsymbol{D}^{\mathrm{H}}$ denotes the conjugate transpose or Hermitian transpose.

Notice that

$$
\begin{aligned}
\|\boldsymbol{U}^{-1}\boldsymbol{y}_{t+1}\|^2 &= \|\boldsymbol{U}^{-1}\left[(\boldsymbol{I}-\eta_t\boldsymbol{G})\boldsymbol{y}_t+\eta_t\boldsymbol{\epsilon}_t\right]\|^2\\
&= \|\boldsymbol{U}^{-1}(\boldsymbol{I}-\eta_t\boldsymbol{G})\boldsymbol{y}_t\|^2+\eta_t^2\|\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\|^2+2\eta_t\mathrm{Re}\langle\boldsymbol{U}^{-1}(\boldsymbol{I}-\eta_t\boldsymbol{G})\boldsymbol{y}_t,\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\rangle\\
&\le \|\boldsymbol{I}-\eta_t\boldsymbol{D}\|^2\|\boldsymbol{U}^{-1}\boldsymbol{y}_t\|^2+\eta_t^2\|\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\|^2+2\eta_t\mathrm{Re}\langle(\boldsymbol{I}-\eta_t\boldsymbol{D})\boldsymbol{U}^{-1}\boldsymbol{y}_t,\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\rangle.
\end{aligned}
$$

We then bound $\|\boldsymbol{I}-\eta_t\boldsymbol{D}\|$ as following.

$$
\begin{aligned}
\|\boldsymbol{I}-\eta_t\boldsymbol{D}\|^2 &= \sup_{\boldsymbol{v}\in\mathbb{C}^d,\|\boldsymbol{v}\|=1} \boldsymbol{v}^{\mathrm{H}}(\boldsymbol{I}-\eta_t\boldsymbol{D})^{\mathrm{H}}(\boldsymbol{I}-\eta_t\boldsymbol{D})\boldsymbol{v}\\
&= \sup_{\boldsymbol{v}\in\mathbb{C}^d,\|\boldsymbol{v}\|=1}\left(\|\boldsymbol{v}\|^2-\eta_t\boldsymbol{v}^{\mathrm{H}}(\boldsymbol{D}^{\mathrm{H}}+\boldsymbol{D})\boldsymbol{v}+\eta_t^2\boldsymbol{v}^{\mathrm{H}}\boldsymbol{D}^{\mathrm{H}}\boldsymbol{D}\boldsymbol{v}\right)\\
&\le 1-2\eta_t\mu+\eta_t^2\|\boldsymbol{D}\|^2.
\end{aligned}
$$

For simplicity, we define

$$
h_t=\frac{\|\boldsymbol{U}^{-1}\boldsymbol{y}_t\|^2}{\eta_t}.
$$

Then we have

$$
\begin{aligned}
h_{t+1}=\frac{\|\boldsymbol{U}^{-1}\boldsymbol{y}_{t+1}\|^2}{\eta_{t+1}} &\le \left(1-2\eta_t\mu+\eta_t^2\|\boldsymbol{D}\|^2\right)\frac{\eta_t}{\eta_{t+1}}h_t+\frac{\eta_t^2}{\eta_{t+1}}\|\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\|^2\\
&\quad+\frac{2\eta_t}{\eta_{t+1}}\mathrm{Re}\langle(\boldsymbol{I}-\eta_t\boldsymbol{D})\boldsymbol{U}^{-1}\boldsymbol{y}_t,\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\rangle.\\
&:= \left(1-2\eta_t\mu+\eta_t^2\|\boldsymbol{D}\|^2\right)\frac{\eta_t}{\eta_{t+1}}h_t+z_t
\end{aligned}
\tag{53}
$$

where for simplicity we denote

$$
z_t=\frac{\eta_t^2}{\eta_{t+1}}\|\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\|^2+\frac{2\eta_t}{\eta_{t+1}}\mathrm{Re}\langle(\boldsymbol{I}-\eta_t\boldsymbol{D})\boldsymbol{U}^{-1}\boldsymbol{y}_t,\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\rangle.
$$

Taking the second-order moment on the both sides of (53), we obtain

$$
\begin{aligned}
\mathbb{E}h_{t+1}^2 &\le \left[\left(1-2\eta_t\mu+\eta_t^2\|\boldsymbol{D}\|^2\right)\frac{\eta_t}{\eta_{t+1}}\right]^2\mathbb{E}h_t^2+\mathbb{E}|z_t|^2\\
&\quad+2\left[\left(1-2\eta_t\mu+\eta_t^2\|\boldsymbol{D}\|^2\right)\frac{\eta_t}{\eta_{t+1}}\right]\mathbb{E}h_tz_t.
\end{aligned}
$$

Due to $\eta_{t+1}=(1-o(\eta_t))\eta_t$ and $\eta_t=o(1)$, there exists $t_0>0$ so that for any $t\ge t_0$, $\eta_t\le 2\eta_{t+1}$ and

$$
0<\frac{\eta_t}{\eta_{t+1}}(1-2\eta_t\mu+\eta_t^2\|\boldsymbol{D}\|^2)=(1+o(1))(1+o(\eta_t))^2(1-2\eta_t\mu+o(\eta_t))\le 1-\mu\eta_t<1.
$$

By Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}|z_t|^2 &\le 2\left(4\eta_t^2\mathbb{E}\|\boldsymbol{U}^{-1}\epsilon_t\|^4+16\mathbb{E}|\mathrm{Re}\langle(\boldsymbol{I}-\eta_t\boldsymbol{D})\boldsymbol{U}^{-1}\boldsymbol{y}_t,\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\rangle|^2\right)\\
&\precsim \eta_t^2\mathbb{E}\|\boldsymbol{U}^{-1}\epsilon_t\|^4+\mathbb{E}\|\boldsymbol{U}^{-1}\boldsymbol{y}_t\|^2\cdot\|\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\|^2\\
&= \eta_t^2\mathbb{E}\|\boldsymbol{U}^{-1}\epsilon_t\|^4+\mathbb{E}(\eta_th_t)\cdot\|\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\|^2\\
&\precsim \eta_t^2\mathbb{E}\|\boldsymbol{U}^{-1}\epsilon_t\|^4+\eta_t\sqrt{\mathbb{E}h_t^2\cdot\mathbb{E}\|\boldsymbol{U}^{-1}\epsilon_t\|^4}.
\end{aligned}
$$

Since $\mathbb{E}[\mathrm{Re}\langle(\boldsymbol{I}-\eta_t\boldsymbol{D})\boldsymbol{U}^{-1}\boldsymbol{y}_t,\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\rangle|\mathcal{F}_t]=0$, it follows that

$$
\mathbb{E}h_tz_t=\mathbb{E}h_t\frac{\eta_t^2}{\eta_{t+1}}\|\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\|^2\le 2\eta_t\cdot\mathbb{E}h_t\|\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\|^2\le 2\eta_t\left(\mathbb{E}\|\boldsymbol{U}^{-1}\boldsymbol{\epsilon}_t\|^4\right)^{\frac{1}{2}}\left(\mathbb{E}h_t^2\right)^{\frac{1}{2}}
$$

where the last inequality follows from Hölder's inequality. Notice that $\sup_{t\ge 0}\mathbb{E}\|\boldsymbol{\epsilon}_t\|^4\precsim 1$ by assumption. Putting the pieces together, we have that there exists some $c>0$ such that

$$
\mathbb{E}h_{t+1}^2\le(1-\mu\eta_t)\mathbb{E}h_t^2+c\left(\eta_t\left(\mathbb{E}h_t^2\right)^{\frac{1}{2}}+\eta_t^2\right).
$$

By induction, one can show that

$$\mathbb{E}h_t^2 \le \frac{c+\sqrt{c^2+4c\mu\eta_0}}{2\mu} \lesssim 1$$

of which the right hand side is the solution of the quadratic equation $\mu x = c\left(\sqrt{x}+\eta_0\right)$. Since $\boldsymbol{U}$ is non-singular, $\mathbb{E}h_t^2 \lesssim 1$ is equivalent to $\mathbb{E}\|\boldsymbol{y}_t\|^4\eta_t^{-2} \lesssim 1$. □

E A CONVERGENCE RESULT

Denote $\boldsymbol{\Delta}_t = \boldsymbol{Q}_t - \boldsymbol{Q}^*$ as the error of the Q-function estimate $\boldsymbol{Q}_t$ in the t-th iteration. In this section, we study both asymptotic and non-asymptotic convergence of $\frac{1}{T}\sum_{t=0}^T \mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2$.

E.1 For General Step Sizes

We first show that $\frac{1}{T}\sum_{t=0}^T \mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 = o\left(\frac{1}{\sqrt{T}}\right)$ when using the general step size in Assumption 3.3.

Theorem E.1. *Under Assumption 3.1 and using the general step size in Assumption 3.3, we have*

$$\lim_{T\to\infty} \frac{1}{\sqrt{T}} \sum_{t=0}^T \mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 = 0. \tag{54}$$

Proof of Theorem E.1. We will make use of the convergence result in [Chen et al., 2020b].

Theorem E.2 (Theorem 2.1 and Corollary 2.1.3 in [Chen et al., 2020b])**.** *Consider the algorithm $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t(\mathcal{H}(\boldsymbol{x}_t) - \boldsymbol{x}_t + \boldsymbol{\varepsilon}_t)$ and $\boldsymbol{x}^*$ is the solution of $\mathcal{H}(\boldsymbol{x}) = \boldsymbol{x}$. Assume (i) $\|\mathcal{H}(\boldsymbol{x}) - \mathcal{H}(\boldsymbol{y})\|_\infty \le \gamma\|\boldsymbol{x}-\boldsymbol{y}\|_\infty$ for any $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^D$; (ii) $\mathbb{E}[\boldsymbol{\varepsilon}_t|\mathcal{F}_t] = \boldsymbol{0}$ and $\mathbb{E}[\|\boldsymbol{\varepsilon}_t\|_\infty^2|\mathcal{F}_t] \le A + B\|\boldsymbol{x}_t\|_\infty^2$ and (iii) η_t is positive and non-increasing. If $\eta_0 \le \frac{\alpha_2}{\alpha_3}$, it follows that*

$$\mathbb{E}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|_\infty^2 \le \alpha_1\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_\infty^2 \prod_{j=0}^t (1-\alpha_2\eta_j) + \alpha_4(A + 2B\|\boldsymbol{x}^*\|_\infty^2)\sum_{j=1}^t \eta_j^2 \prod_{i=j+1}^t (1-\alpha_2\eta_i).$$

where

$$\alpha_1 \le \frac{3}{2}, \alpha_2 \ge \frac{1-\gamma}{2}, \alpha_3 \le \frac{32e(B+2)\log D}{1-\gamma}, \alpha_4 \le \frac{16e\log D}{1-\gamma}.$$

Recall the update rule is $\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r}_t + \gamma\boldsymbol{P}_t\boldsymbol{V}_{t-1}) = \boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r} + \gamma\boldsymbol{P}\boldsymbol{V}_{t-1} - \boldsymbol{Q}_{t-1} + \boldsymbol{\varepsilon}_t)$ where $\boldsymbol{\varepsilon}_t = \boldsymbol{r}_t - \boldsymbol{r} + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}_{t-1}$. Let $\mathcal{F}_t = \sigma(\{(\boldsymbol{r}_\tau, \boldsymbol{P}_\tau)\}_{0\le\tau<t})$. Hence, $\mathbb{E}[\boldsymbol{\varepsilon}_t|\mathcal{F}_t] = \boldsymbol{0}$ and $\mathbb{E}[\|\boldsymbol{\varepsilon}_t\|_\infty^2|\mathcal{F}_t] \le 2\mathbb{E}\|\boldsymbol{r}_t - \boldsymbol{r}\|_\infty^2 + 2\gamma^2\mathbb{E}\|\boldsymbol{P}_t - \boldsymbol{P}\|_\infty^2\|\boldsymbol{V}_{t-1}\|_\infty^2 := A + B\|\boldsymbol{Q}_{t-1}\|_\infty^2$ where the last equation uses $A = 2\mathbb{E}\|\boldsymbol{r}_t - \boldsymbol{r}\|_\infty^2, B = 2\gamma^2\mathbb{E}\|\boldsymbol{P}_t - \boldsymbol{P}\|_\infty^2$ and $\|\boldsymbol{V}_{t-1}\|_\infty = \|\boldsymbol{Q}_{t-1}\|_\infty$. Then setting $\widetilde{\eta}_t = (1-\gamma)\eta_t$, by Theorem E.2, we have

$$\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \le 2\|\boldsymbol{\Delta}_0\|_\infty^2 \prod_{j=1}^t (1-0.5\widetilde{\eta}_j) + C_1\sum_{j=1}^t \eta_j \cdot 0.5\widetilde{\eta}_j \prod_{i=j+1}^t (1-0.5\widetilde{\eta}_i). \tag{55}$$

where

$$C_1 = \frac{32e\log D}{(1-\gamma)^2}(A + 2B\|\boldsymbol{Q}^*\|_\infty^2).$$

To simplify the notation, we denote

$$\widetilde{\eta}_{(t,T)} = \begin{cases} \prod_{j=1}^T (1-0.5\widetilde{\eta}_j), & \text{if } t=0 \\ 0.5\widetilde{\eta}_t \prod_{j=t+1}^T (1-0.5\widetilde{\eta}_j), & \text{if } 0<t<T \\ 0.5\widetilde{\eta}_T, & \text{if } t=T. \end{cases} \tag{56}$$

It is clear that we have $\sum_{t=0}^T \widetilde{\eta}_{(t,T)} = 1$. Then it follows that

$$\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \le 2\|\boldsymbol{\Delta}_0\|_\infty^2\widetilde{\eta}_{(0,T)} + C_1\sum_{j=1}^t \eta_j \cdot \widetilde{\eta}_{(j,t)}.$$

Therefore, it follows that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbb{E}\|\mathbf{\Delta}_t\|_\infty^2 \le \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[2\widetilde{\eta}_{(0,t)}\|\mathbf{\Delta}_0\|_\infty^2 + C_1\sum_{s=1}^{t}\widetilde{\eta}_{(s,t)}\eta_{s-1}\right] = \frac{2\|\mathbf{\Delta}_0\|_\infty^2}{\sqrt{T}}\sum_{t=1}^{T}\widetilde{\eta}_{(0,t)} + \frac{C_1}{\sqrt{T}}\sum_{t=1}^{T}\sum_{s=1}^{t}\widetilde{\eta}_{(s,t)}\eta_{s-1}.$$

Recall that Assumption 3.3 requires the step size satisfies

(C1) $0 \le \sup_t \eta_t \le 1, \eta_t \downarrow 0$ and $t\eta_t \uparrow \infty$ when $t \to \infty$;

(C2) $\frac{\eta_{t-1}-\eta_t}{\eta_{t-1}} = o(\eta_{t-1})$ for all $t \ge 1$;

(C3) $\frac{1}{\sqrt{T}}\sum_{t=0}^{T}\eta_t \to 0$ when $T \to \infty$.

Noticing $t\eta_t \uparrow \infty$ due to (C1), we must have $\sum_{t=1}^{T}\widetilde{\eta}_t - \frac{1}{4}\ln T \to +\infty$ and thus implies

$$\sqrt{T}\widetilde{\eta}_{(0,T)} = \sqrt{T}\prod_{t=1}^{T}(1-0.5\widetilde{\eta}_t) \le \exp\left(\frac{1}{2}\ln T - 2\sum_{t=1}^{T}\widetilde{\eta}_t\right) \to 0,$$

which, together with the Stolz–Cesaro theorem, implies $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\widetilde{\eta}_{(0,t)}^2 \to 0$.

On the other hand, by Lemma E.1 and (C3), it follows that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sum_{s=1}^{t}\widetilde{\eta}_{(s,t)}\eta_{s-1} = \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\eta_{s-1}\cdot\sum_{t=s}^{T}\widetilde{\eta}_{(s,t)} \le \frac{c}{\sqrt{T}}\sum_{t=1}^{T}\eta_{t-1} \to 0.$$

Lemma E.1. *There exists some $c > 0$ such that $\sum_{l=t}^{T}\widetilde{\eta}_{(t,l)} \le c$ for any $T \ge t \ge 1$. Here $\{\widetilde{\eta}_{(t,l)}\}_{l\ge t\ge 0}$ is defined in (56) and $\{\widetilde{\eta}_t\}_{t\ge 0}$ satisfies Assumption 3.3.*

Putting all pieces together, we have established (54). □

Proof of Lemma E.1. We define $\widetilde{m}_{t,l} := \sum_{i=t}^{l}\widetilde{\eta}_i$. Due to $t\widetilde{\eta}_t \uparrow \infty$, we have $t\widetilde{\eta}_t \le i\widetilde{\eta}_i$ for all $i \ge t$ and thus

$$\widetilde{m}_{t,l} := \sum_{i=t}^{l}\widetilde{\eta}_i \ge t\widetilde{\eta}_t\sum_{i=t}^{l}\frac{1}{i} \ge t\widetilde{\eta}_t\left(\ln\frac{l}{t} - \frac{1}{2t}\right) = -\frac{\widetilde{\eta}_t}{2} + t\widetilde{\eta}_t\ln\frac{l}{t}.$$

Since $t\widetilde{\eta}_t \uparrow \infty$, there exists some $t_0 > 0$ such that any $t \ge t_0$, we have $t\widetilde{\eta}_t \ge 2$. Therefore, we have for all $l \ge t \ge t_0$,

$$\frac{1}{\widetilde{\eta}_l} \le \frac{l}{t\widetilde{\eta}_t} \le \frac{1}{\widetilde{\eta}_t}\exp\left(\frac{\widetilde{m}_{t,l}+\frac{\widetilde{\eta}_t}{2}}{t\widetilde{\eta}_t}\right) \le \frac{\sqrt{e}}{\widetilde{\eta}_t}\exp\left(\frac{\widetilde{m}_{t,l}}{2}\right). \tag{57}$$

In the following, we will discuss three cases.

- If $T \ge t \ge t_0$, by definition, it follows that

$$\begin{aligned}\sum_{l=t}^{T}\widetilde{\eta}_{(t,l)} &= \sum_{l=t}^{T}\widetilde{\eta}_t\prod_{j=t+1}^{l}(1-\widetilde{\eta}_j) \le \frac{\widetilde{\eta}_t}{1-\widetilde{\eta}_t}\sum_{l=t}^{T}\exp\left(-\widetilde{m}_{t,l}\right)\\ &\overset{(a)}{\le} \frac{\widetilde{\eta}_t}{1-\widetilde{\eta}_t}\sum_{l=t}^{T}\widetilde{\eta}_l\cdot\frac{\sqrt{e}}{\widetilde{\eta}_t}\exp\left(-\frac{\widetilde{m}_{t,l}}{2}\right)\\ &\overset{(b)}{\le} \frac{\sqrt{e}}{\gamma}\sum_{l=t}^{T}\widetilde{\eta}_l\exp\left(-\frac{\widetilde{m}_{t,l}}{2}\right) \overset{(c)}{\le} \frac{2\sqrt{e}}{\gamma},\end{aligned}$$

  where (a) follows from (57); (b) uses $1-\widetilde{\eta}_t \ge 1-\widetilde{\eta}_0 = \gamma$; and (c) uses $\sum_{l=t}^{T}\widetilde{\eta}_l\exp\left(-\frac{\widetilde{m}_{t,l}}{2}\right) \le \int_0^\infty \exp(-x/2)dx = 2$ due to $\widetilde{m}_{t,l} \uparrow \infty$ as $l \to \infty$.

- If $T \ge t_0 \ge t$, by definition, $\widetilde{\eta}_{(t,l)} = \widetilde{\eta}_{(t,t_0)}\widetilde{\eta}_{(t_0,l)}/\widetilde{\eta}_{t_0} \le C_2\widetilde{\eta}_{(t_0,l)}$ where $C_2 = \sup_{0\le t\le t_0}\widetilde{\eta}_{(t,t_0)}/\widetilde{\eta}_{t_0}$. Then we have $\sum_{l=t}^T \widetilde{\eta}_{(t,l)} = \sum_{l=t}^{t_0}\widetilde{\eta}_{(t,l)} + \sum_{l=t_0}^T \widetilde{\eta}_{(t,l)} \le t_0 + C_2\sum_{l=t_0}^T \widetilde{\eta}_{(t_0,l)} \le t_0 + C_2\frac{2\sqrt{e}}{\gamma}$.
- If $t_0 \ge T \ge t$, we have $\sum_{l=t}^T \widetilde{\eta}_{(t,l)} \le t_0$.

Putting the three cases together, we can set $c = t_0 + 2\max\{C_0, 1\}\sqrt{e}/\gamma$ which ensures that $\sum_{l=t}^T \widetilde{\eta}_{(t,l)} \le c$ for any $T \ge t \ge 1$. □

E.2 For Two Specific Step Sizes

To obtain an $\log D$ dependence (which implies the rewards are distributed either sub-gaussian or sub-exponential), we use a almost-surely bounded rewards assumption as follows.

Assumption E.1. *We assume $0 \le R(s,a) \le 1$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.*

Theorem E.3. *Under Assumption E.1, there exist some positive constant $c > 0$ such that*

- *If $\eta_t = \frac{1}{1+(1-\gamma)t}$, it follows that*

$$\frac{1}{T}\sum_{t=0}^T \mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \le c\left[\frac{\|\boldsymbol{\Delta}_0\|_\infty^2}{(1-\gamma)^2}\frac{1}{T} + \frac{\ln(2eD)}{(1-\gamma)^5}\frac{\ln^2(eT)}{T}\right].$$

- *If $\eta_t = t^{-\alpha}$ with $\alpha \in (0,1)$ for $t \ge 1$ and $\eta_0 = 1$, it follows that*

$$\frac{1}{T}\sum_{t=0}^T \mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \le c\left[\frac{\Delta_0}{\sqrt{1-\alpha}(1-\gamma)^{\frac{1}{1-\alpha}}}\frac{1}{T} + \frac{\ln(2eD)}{(1-\alpha)(1-\gamma)^4}\frac{1}{T^\alpha}\right],$$

where

$$\Delta_0 = 3\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{48\gamma^2\ln(2eD)}{(1-\gamma)^3}\left(\frac{2\alpha}{1-\gamma}\right)^{\frac{1}{1-\alpha}}.$$

E.3 Proof of Theorem E.3

Our proof is divided into three steps. The first is a upper bound for $\|\boldsymbol{\Delta}_t\|_\infty$ provided by Lemma E.2: $\|\boldsymbol{\Delta}_t\|_\infty \le a_t + b_t + \|\boldsymbol{N}_t\|_\infty$, As a result, $\|\boldsymbol{\Delta}_t\|_\infty^2 \le 3(a_t^2 + b_t^2 + \|\boldsymbol{N}_t\|_\infty^2)$. Lemma E.2 follows from Theorem 1 in [Wainwright, 2019b] which views Q-learning as a cone-contractive operator and establishes a ℓ_∞-norm bound.

Lemma E.2 (Theorem 1 in [Wainwright, 2019b])**.** *For any sequence of step sizes $\{\eta_t\}_{t\ge0}$ in the interval $(0,1)$, the iterates $\{\boldsymbol{\Delta}_t\}_{t\ge0}$ satisfies the sandwich relation*

$$-(a_t+b_t)\mathbf{1} + \boldsymbol{N}_t \le \boldsymbol{\Delta}_t \le (a_t+b_t)\mathbf{1} + \boldsymbol{N}_t \tag{58}$$

where $\{a_t\}_{t\ge0}, \{b_t\}_{t\ge0}$ are non-negative scalars and $\{\boldsymbol{N}_t\}_{t\ge0}$ are random vectors collecting noise terms from empirical Bellman operators. The three sequences are defined in a recursive way: they are initialized as $a_0 = \|\boldsymbol{\Delta}_0\|_\infty, b_0 = 0$ and $\boldsymbol{N}_0 = \mathbf{0}$ and satisfy the following recursion:

$$\begin{aligned} a_t &= (1-\eta_t(1-\gamma))a_{t-1} \\ b_t &= (1-\eta_t(1-\gamma))b_{t-1} + \eta_t\gamma\|\boldsymbol{N}_{t-1}\|_\infty \\ \boldsymbol{N}_t &= (1-\eta_t)\boldsymbol{N}_{t-1} + \eta_t\boldsymbol{Z}_t, \end{aligned}$$

where $\boldsymbol{Z}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*$ is the empirical Bellman error at iteration t.

The second step is to bound $\mathbb{E}\|\boldsymbol{N}_T\|_\infty^2$ which is an autoregressive process of independent Bellman noise terms. One can prove the result following a similar argument of Lemma 2 in [Wainwright, 2019b].

Lemma E.3. *Under Assumption E.1 and assuming $(1-\eta_t)\eta_{t-1} \le \eta_t$ for any $t \ge 1$, we have*

$$\mathbb{E}\|\boldsymbol{N}_t\|_\infty^2 \le \frac{2\eta_t\ln(2eD)}{(1-\gamma)^2}.$$

The final step is to establish the dependence of $\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2$ on $\{\eta_t\}_{t\geq 0}$. Wainwright [2019b] finds it is crucial to set η_t to be proportional to $1/(1-\gamma)$ to ensure the sample complexity has polynomial dependence on $1/(1-\gamma)$. We then set $\widetilde{\eta}_t = (1-\gamma)\eta_t$ as the rescaled step size. We first redefine

$$\widetilde{\eta}_{(t,T)} = \begin{cases} \prod_{j=1}^T (1-\widetilde{\eta}_j), & \text{if } t=0 \\ \widetilde{\eta}_t \prod_{j=t+1}^T (1-\widetilde{\eta}_j), & \text{if } 0<t<T \\ \widetilde{\eta}_T, & \text{if } t=T. \end{cases} \tag{59}$$

It is clear that we have $\sum_{t=0}^T \widetilde{\eta}_{(t,T)} = 1$.

Lemma E.4. *Under Assumption 3.1, if $(1-\eta_t)\eta_{t-1} \leq \eta_t$ for any $t \geq 1$, then we have*

$$\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 \leq 3\widetilde{\eta}_{(0,T)}^2\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{6\gamma^2\ln(2eD)}{(1-\gamma)^4}\sum_{t=1}^T \widetilde{\eta}_{(t,T)}\eta_{t-1} + \frac{6\ln(2eD)}{(1-\gamma)^2}\eta_T, \tag{60}$$

where $\{\widetilde{\eta}_{(t,T)}\}_{T\geq t\geq 0}$ defined in (59) and $\{\boldsymbol{N}_t\}_{t\geq 0}$ is defined in Lemma E.2.

Proof of Lemma E.4. By the recursion of $\{a_t\}_{t\geq 0}$ and $\{b_t\}_{t\geq 0}$ in Lemma E.2, it follows that

$$a_T = \prod_{t=1}^T (1-\widetilde{\eta}_t)\|\boldsymbol{\Delta}_0\|_\infty = \widetilde{\eta}_{(0,T)}\|\boldsymbol{\Delta}_0\|_\infty$$
$$b_T = \gamma\sum_{t=1}^T\prod_{j=t+1}^T (1-\widetilde{\eta}_j)\eta_t\|\boldsymbol{N}_{t-1}\|_\infty = \frac{\gamma}{1-\gamma}\sum_{t=1}^T \widetilde{\eta}_{(t,T)}\|\boldsymbol{N}_{t-1}\|_\infty.$$

Hence, $a_T^2 = \widetilde{\eta}_{(0,T)}^2\|\boldsymbol{\Delta}_0\|_\infty^2$ and

$$\mathbb{E}b_T^2 = \frac{\gamma^2}{(1-\gamma)^2}\mathbb{E}\left(\sum_{t=1}^T \widetilde{\eta}_{(t,T)}\|\boldsymbol{N}_{t-1}\|_\infty\right)^2 \overset{(a)}{\leq} \frac{\gamma^2}{(1-\gamma)^2}\sum_{t=1}^T \widetilde{\eta}_{(t,T)}\mathbb{E}\|\boldsymbol{N}_{t-1}\|_\infty^2$$

where (a) uses $\sum_{t=1}^T \widetilde{\eta}_{(t,T)} = 1-\widetilde{\eta}_{(0,T)} \leq 1$ and Jensen's inequality.

Therefore,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 &\leq 3(a_T^2 + \mathbb{E}b_T^2 + \mathbb{E}\|\boldsymbol{N}_T\|_\infty^2) \\ &\leq 3\widetilde{\eta}_{(0,T)}^2\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{3\gamma^2}{(1-\gamma)^2}\sum_{t=1}^T \widetilde{\eta}_{(t,T)}\mathbb{E}\|\boldsymbol{N}_{t-1}\|_\infty^2 + 3\mathbb{E}\|\boldsymbol{N}_T\|_\infty^2.\end{aligned} \tag{61}$$

Given the condition $(1-\eta_t)\eta_{t-1} \leq \eta_t$, we can apply Lemma E.3 which implies

$$\mathbb{E}\|\boldsymbol{N}_t\|_\infty^2 \leq \frac{2\eta_t\ln(2eD)}{(1-\gamma)^2}.$$

Plugging these bounds into (61) yields (60). □

With these lemmas, we are ready to prove the following theorem.

Theorem E.4. *Under Assumption 3.1, we have the following bounds for $\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2$. Here $c>0$ is a universal positive constant and might be overwritten (and thus different) in different statements. The specific value of different c's can be found in our proof.*

- *If $\eta_t = \frac{1}{1+(1-\gamma)t}$, it follows that for all $T \geq 1$,*

$$\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 \leq \frac{12\|\boldsymbol{\Delta}_0\|_\infty^2}{(1-\gamma)^2}\frac{1}{(1+T)^2} + \frac{12\gamma^2\ln(2eD)}{(1-\gamma)^5}\frac{\ln(eT)}{T}.$$

- *If* $\eta_t = t^{-\alpha}$ *with* $\alpha \in (0,1)$ *for* $t \ge 1$ *and* $\eta_0 = 1$*, it follows that for all* $T \ge 1$*,*

$$\mathbb{E}\|\mathbf{\Delta}_T\|_\infty^2 \le \Delta_0 \exp\left(-\frac{1-\gamma}{1-\alpha}\left((1+T)^{1-\alpha}-1\right)\right) + \frac{114\ln(2eD)}{(1-\gamma)^4}\frac{1}{T^\alpha}$$

, where

$$\Delta_0 = 3\|\mathbf{\Delta}_0\|_\infty^2 + \frac{48\gamma^2\ln(6D)}{(1-\gamma)^3}\left(\frac{2\alpha}{1-\gamma}\right)^{\frac{1}{1-\alpha}}.$$

Proof of Theorem E.4. We discuss the two cases separately.

(I) Linearly rescaled step size. If we use a linear rescaled step size, i.e., $\eta_t = \frac{1}{1+(1-\gamma)t}$ (equivalently $\widetilde{\eta}_t = \frac{1-\gamma}{1+(1-\gamma)t}$), then we have (i) $1-\eta_t \le 1-\widetilde{\eta}_t = \frac{1+(1-\gamma)(t-1)}{1+(1-\gamma)t} = \widetilde{\eta}_t/\widetilde{\eta}_{t-1} = \eta_t/\eta_{t-1}$ for $t \ge 1$ and (ii) $\widetilde{\eta}_{(t,T)} \le \widetilde{\eta}_T$. It implies Lemma E.4 is applicable. Notice that $\sum_{t=1}^T \widetilde{\eta}_{t-1} \le 1 + \sum_{t=1}^{T-1}\frac{1}{t} \le 1+\ln(T-1) \le \ln(eT)$ and $\ln\frac{(1-\gamma)(T+1)}{2} \le \ln\frac{1+(1-\gamma)(T+1)}{1+(1-\gamma)} = \int_1^{T+1}\frac{1-\gamma}{1+(1-\gamma)t}dt \le \sum_{t=1}^T\frac{1-\gamma}{1+(1-\gamma)t} = \sum_{t=1}^T\widetilde{\eta}_t$. Hence,

$$\widetilde{\eta}_{(0,T)}^2 = \prod_{t=1}^T(1-\widetilde{\eta}_t)^2 \le \exp\left(-2\sum_{t=1}^T\widetilde{\eta}_t\right) \le \frac{4}{(1-\gamma)^2}\frac{1}{(1+T)^2}$$

$$\sum_{t=1}^T\widetilde{\eta}_{(t,T)}\eta_{t-1} = \frac{1}{1-\gamma}\sum_{t=1}^T\widetilde{\eta}_{(t,T)}\widetilde{\eta}_{t-1} \le \frac{\widetilde{\eta}_T}{1-\gamma}\sum_{t=1}^T\widetilde{\eta}_{t-1} \le \frac{\widetilde{\eta}_T\ln(eT)}{1-\gamma}.$$

Finally, plugging these inequalities into (60), we have

$$\mathbb{E}\|\mathbf{\Delta}_T\|_\infty^2 \le \frac{12\|\mathbf{\Delta}_0\|_\infty^2}{(1-\gamma)^2}\frac{1}{(1+T)^2} + \frac{12\gamma^2\ln(2eD)}{(1-\gamma)^5}\frac{\ln(eT)}{T}. \tag{62}$$

(II) Polynomial step size. If we choose a polynomial step size, i.e., $\eta_t = t^{-\alpha}$ with $\alpha \in (0,1)$ for $t \ge 1$ and $\eta_0 = 1$, then we again have $1-\eta_t = 1-\frac{1}{t^\alpha} \le \left(\frac{t-1}{t}\right)^\alpha = \eta_t/\eta_{t-1}$ for $t \ge 1$, which implies Lemma E.3 is applicable. Note that

$$\frac{(T+1)^{1-\alpha}-(t+1)^{1-\alpha}}{1-\alpha} = \int_{t+1}^{T+1}j^{-\alpha}dj \le \sum_{j=t+1}^T j^{-\alpha} \le \int_t^T j^{-\alpha}dj = \frac{T^{1-\alpha}-t^{1-\alpha}}{1-\alpha}, \tag{63}$$

which implies that $\sum_{t=1}^T\eta_t \ge \sum_{t=1}^T t^{-\alpha} \ge \frac{1}{1-\alpha}\left((T+1)^{1-\alpha}-1\right)$ and $(T+1)^{1-\alpha} \le 1+T^{1-\alpha}$. Hence,

$$\widetilde{\eta}_{(0,T)}^2 = \prod_{t=1}^T(1-\widetilde{\eta}_t)^2 \le \exp\left(-2(1-\gamma)\sum_{t=1}^T\eta_t\right) \le \exp\left(-2\frac{1-\gamma}{1-\alpha}\left((1+T)^{1-\alpha}-1\right)\right).$$

Additionally, using $\eta_{t-1} \le 2\eta_t$ for all $t \ge 1$ and (63), we have,

$$\begin{aligned}\frac{\widetilde{\eta}_{(t,T)}}{1-\gamma}\eta_{t-1} = \prod_{j=t+1}^T(1-\widetilde{\eta}_j)\eta_t\eta_{t-1} &\le 8\prod_{j=t+1}^T(1-\widetilde{\eta}_j)\eta_{t+1}^2 \le 8\exp\left(-\sum_{j=t+1}^T\widetilde{\eta}_j\right)\eta_{t+1}^2\\ &\le 8\exp\left(-\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)\frac{\exp\left(\frac{1-\gamma}{1-\alpha}(t+1)^{1-\alpha}\right)}{(t+1)^{2\alpha}},\end{aligned}$$

which implies

$$\begin{aligned}\frac{1}{1-\gamma}\sum_{t=1}^T\widetilde{\eta}_{(t,T)}\eta_{t-1} &\le \frac{1}{1-\gamma}\sum_{t=1}^{T-1}\widetilde{\eta}_{(t,T)}\eta_{t-1} + \eta_T\eta_{T-1} \le \frac{1}{1-\gamma}\sum_{t=1}^{T-1}\widetilde{\eta}_{(t,T)}\eta_{t-1} + \eta_T^2\\ &\le 8\sum_{t=2}^T\exp\left(-\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)\frac{\exp\left(\frac{1-\gamma}{1-\alpha}t^{1-\alpha}\right)}{t^{2\alpha}} + \frac{2}{T^{2\alpha}}.\end{aligned}$$

At the the end of this subsection, we will prove that

Lemma E.5. *For any $\alpha \in (0,1)$ and $\beta > 0$, it follows that*

$$\sum_{t=1}^{T} \frac{\exp\left(\frac{1-\gamma}{1-\alpha} t^{1-\alpha}\right)}{t^{\beta}} \le \left(\frac{\beta}{1-\gamma}\right)^{\frac{1}{1-\alpha}} \exp\left(\frac{1-\gamma}{1-\alpha}\right) + \frac{\beta}{(1-\gamma)\alpha} \frac{\exp\left(\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)}{(1+T)^{\beta-\alpha}}. \tag{64}$$

By setting $\beta = 2\alpha$, we have

$$\sum_{t=1}^{T} \frac{\exp\left(\frac{1-\gamma}{1-\alpha} t^{1-\alpha}\right)}{t^{2\alpha}} \le \left(\frac{2\alpha}{1-\gamma}\right)^{\frac{1}{1-\alpha}} \exp\left(\frac{1-\gamma}{1-\alpha}\right) + \frac{2}{1-\gamma} \frac{\exp\left(\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)}{(1+T)^{\alpha}}.$$

Therefore,

$$\frac{1}{1-\gamma}\sum_{t=1}^{T} \widetilde{\eta}_{(t,T)}\eta_{t-1} \le 8\left(\frac{2\alpha}{1-\gamma}\right)^{\frac{1}{1-\alpha}} \exp\left(-\frac{1-\gamma}{1-\alpha}\left((1+T)^{1-\alpha}-1\right)\right) + \frac{16}{1-\gamma}\frac{1}{(1+T)^{\alpha}} + \frac{2}{T^{2\alpha}}.$$

Putting together the pieces, we can safely conclude that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 &\le 3\|\boldsymbol{\Delta}_0\|_\infty^2 \exp\left(-2\frac{1-\gamma}{1-\alpha}\left((T+1)^{1-\alpha}-1\right)\right) + \frac{6\ln(2eD)}{(1-\gamma)^2}\frac{1}{T^{\alpha}} + \frac{96\gamma^2\ln(2eD)}{(1-\gamma)^4}\frac{1}{(1+T)^{\alpha}} \\
&\quad + \frac{12\gamma^2\ln(2eD)}{(1-\gamma)^3}\frac{1}{T^{2\alpha}} + \frac{48\gamma^2\ln(2eD)}{(1-\gamma)^3}\exp\left(-\frac{1-\gamma}{1-\alpha}\left((1+T)^{1-\alpha}-1\right)\right)\left(\frac{2\alpha}{1-\gamma}\right)^{\frac{1}{1-\alpha}} \\
&\le \Delta_0 \exp\left(-\frac{1-\gamma}{1-\alpha}\left((1+T)^{1-\alpha}-1\right)\right) + \frac{114\ln(2eD)}{(1-\gamma)^4}\frac{1}{T^{\alpha}},
\end{aligned}$$

where

$$\Delta_0 = 3\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{48\gamma^2\ln(6D)}{(1-\gamma)^3}\left(\frac{2\alpha}{1-\gamma}\right)^{\frac{1}{1-\alpha}}.$$

□

Proof of Lemma E.5. We do this via a similar argument of Lemma 4 in [Wainwright, 2019b]. Let $f(t) = \frac{\exp\left(\frac{1-\gamma}{1-\alpha}t^{1-\alpha}\right)}{t^{\beta}}$. By taking derivatives, we find that $f(t)$ is decreasing in t on the interval $[0, t^*]$ and increasing for $[t^*, \infty)$, where $t^* = \left(\frac{\beta}{1-\gamma}\right)^{\frac{1}{1-\alpha}}$. Hence,

$$\sum_{t=1}^{T} f(t) \le \begin{cases} Tf(1) & \text{if } T \le \lfloor t^* \rfloor, \\ \lfloor t^* \rfloor f(1) + \int_{t^*}^{T+1} f(t)dt & \text{if } T > \lfloor t^* \rfloor. \end{cases}$$

Using integrating by parts, it follows that

$$\begin{aligned}
I^* := \int_{t^*}^{T+1} f(t)dt &= \left.\frac{\exp\left(\frac{1-\gamma}{1-\alpha}t^{1-\alpha}\right)}{(1-\gamma)t^{\beta-\alpha}}\right|_{t^*}^{T+1} + \frac{\beta-\alpha}{1-\gamma}\int_{t^*}^{T+1}\frac{\exp\left(\frac{1-\gamma}{1-\alpha}t^{1-\alpha}\right)}{t^{1+\beta-\alpha}}dt \\
&\le \frac{\exp\left(\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)}{(1-\gamma)(1+T)^{\beta-\alpha}} + \frac{\beta-\alpha}{1-\gamma}\int_{t^*}^{T+1}\frac{f(t)}{t^{1-\alpha}}dt \\
&\le \frac{\exp\left(\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)}{(1-\gamma)(1+T)^{\beta-\alpha}} + \frac{\beta-\alpha}{1-\gamma}\frac{1}{(t^*)^{1-\alpha}}\int_{t^*}^{T+1} f(t)dt \\
&= \frac{\exp\left(\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)}{(1-\gamma)(1+T)^{\beta-\alpha}} + \frac{\beta-\alpha}{\beta}I^*,
\end{aligned}$$

where the last equality uses definition of t^* and I^*. Hence, we have

$$I^* = \int_{t^*}^{T+1} f(t)dt \le \frac{\beta}{(1-\gamma)\alpha}\frac{\exp\left(\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)}{(1+T)^{\beta-\alpha}}.$$

Putting together the pieces, we have shown that if $T > \lfloor t^* \rfloor$,

$$\sum_{t=1}^{T} f(t) \leq t^* f(1) + I^* = \left(\frac{\beta}{1-\gamma}\right)^{\frac{1}{1-\alpha}} \exp\left(\frac{1-\gamma}{1-\alpha}\right) + \frac{\beta}{(1-\gamma)\alpha} \frac{\exp\left(\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)}{(1+T)^{\beta-\alpha}}.$$

If $T \leq \lfloor t^* \rfloor$, then

$$\sum_{t=1}^{T} f(t) \leq \lfloor t^* \rfloor f(1) \leq t^* f(1) = \left(\frac{\beta}{1-\gamma}\right)^{\frac{1}{1-\alpha}} \exp\left(\frac{1-\gamma}{1-\alpha}\right).$$

Thus we have proved the inequality is true for any choice of T. □

Based on Theorem E.4, we now can prove Theorem E.3 by averaging the individual error bounds.

Proof of Theorem E.3. The result directly follows from Theorem E.4.

- For the first item, we already have $\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 \leq \frac{12\|\boldsymbol{\Delta}_0\|_\infty^2}{(1-\gamma)^2}\frac{1}{(1+T)^2} + \frac{12\gamma^2 \ln(2eD)}{(1-\gamma)^5}\frac{\ln(eT)}{T}$. Using $\sum_{t=1}^{\infty} t^{-2} = \frac{\pi^2}{6}$ and $\sum_{t=1}^{T} t^{-1} \leq 1 + \ln T = \ln(eT)$, we have for some universal constant $c > 0$,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T} \mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 &\leq \frac{1}{T}\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{1}{T}\sum_{t=1}^{T}\left[\frac{12\|\boldsymbol{\Delta}_0\|_\infty^2}{(1-\gamma)^2}\frac{1}{(1+t)^2} + \frac{12\gamma^2\ln(2eD)}{(1-\gamma)^5}\frac{\ln(eT)}{T}\right] \\ &= c\left[\frac{\|\boldsymbol{\Delta}_0\|_\infty^2}{(1-\gamma)^2}\frac{1}{T} + \frac{\ln(2eD)}{(1-\gamma)^5}\frac{\ln^2(eT)}{T}\right].\end{aligned}$$

- For the second item, we have $\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 \leq \Delta_0 \exp\left(-\frac{1-\gamma}{1-\alpha}\left((1+T)^{1-\alpha}-1\right)\right) + \frac{114\ln(2eD)}{(1-\gamma)^4}\frac{1}{T^\alpha}$ with $\Delta_0 = 3\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{48\gamma^2\ln(2eD)}{(1-\gamma)^3}\left(\frac{2\alpha}{1-\gamma}\right)^{\frac{1}{1-\alpha}}$. Notice that

$$\begin{aligned}\sum_{t=2}^{\infty}\exp\left(-\frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha}-1\right)\right) &\leq \int_1^\infty \exp\left(-\frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha}-1\right)\right)dt \\ &\overset{(a)}{=} \frac{\exp\left(\frac{1-\gamma}{1-\alpha}\right)}{1-\gamma}\int_0^\infty e^{-x}\left(\frac{1-\alpha}{1-\gamma}x\right)^{\frac{\alpha}{1-\alpha}}dx \\ &\overset{(b)}{=} \frac{\exp\left(\frac{1-\gamma}{1-\alpha}\right)(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma(\frac{1}{1-\alpha})}{(1-\gamma)^{\frac{1}{1-\alpha}}} \\ &\overset{(c)}{\leq} \frac{\sqrt{2\pi e}}{\sqrt{1-\alpha}}\frac{1}{(1-\gamma)^{\frac{1}{1-\alpha}}}\end{aligned}$$

  and $\sum_{t=1}^{T} t^{-\alpha} \leq \int_0^T t^{-\alpha}dt = \frac{T^{1-\alpha}}{1-\alpha}$. Here (a) uses the change of variable $x = \frac{1-\gamma}{1-\alpha}t^{1-\alpha}$ and (b) uses the definition of gamma function $\Gamma(z) = \int_0^\infty e^{-x}x^{z-1}dx$. Finally (c) follows from a numeral inequality about gamma function. Since $\Gamma(1+x) < \sqrt{2\pi}\left(\frac{x+1/2}{e}\right)^{x+1/2}$ for any $x > 0$ (see Theorem 1.5 of [Batir, 2008]), then

$$\Gamma\left(\frac{1}{1-\alpha}\right) \leq \sqrt{2\pi}\left(\frac{1+\alpha}{2(1-\alpha)}\right)^{\frac{1+\alpha}{2(1-\alpha)}}\exp\left(-\frac{1+\alpha}{2(1-\alpha)}\right),$$

  which implies that

$$\exp\left(\frac{1-\gamma}{1-\alpha}\right)(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right) \leq \frac{\sqrt{2\pi e}}{\sqrt{1-\alpha}}. \tag{65}$$

Therefore,

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \leq \frac{1}{T}\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{1}{T}\sum_{t=1}^{T}\left[\Delta_0 \exp\left(-\frac{1-\gamma}{1-\alpha}\left((1+t)^{1-\alpha}-1\right)\right) + \frac{114\ln(2eD)}{(1-\gamma)^4}\frac{1}{t^\alpha}\right]$$
$$\leq c\left[\frac{\Delta_0}{\sqrt{1-\alpha}(1-\gamma)^{\frac{1}{1-\alpha}}}\frac{1}{T} + \frac{\ln(2eD)}{(1-\alpha)(1-\gamma)^4}\frac{1}{T^\alpha}\right].$$

□

F PROOF OF THEOREM 5.1

In the section, we provide the proof for our finite-sample analysis of averaged Q-learning in the ℓ_∞-norm. Our main idea is similar to Appendix C. The average Q-learning estimator $\bar{\boldsymbol{Q}}_T$ has the error

$$\bar{\boldsymbol{\Delta}}_T := \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Delta}_t = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Q}_t - \boldsymbol{Q}^*). \tag{66}$$

Using two auxiliary sequences $\{\boldsymbol{\Delta}_t^1\}_{t\geq 0}$ and $\{\boldsymbol{\Delta}_t^2\}_{t\geq 0}$ defined in Lemma C.1, we similarly define

$$\bar{\boldsymbol{\Delta}}_T^1 := \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Delta}_t^1 \text{ and } \bar{\boldsymbol{\Delta}}_T^2 := \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Delta}_t^2.$$

Because $\boldsymbol{\Delta}_t^2 \leq \boldsymbol{\Delta}_t \leq \boldsymbol{\Delta}_t^1$ coordinate-wise, it is valid that

$$\bar{\boldsymbol{\Delta}}_T^2 \leq \bar{\boldsymbol{\Delta}}_T \leq \bar{\boldsymbol{\Delta}}_T^1. \tag{67}$$

As a result, $\mathbb{E}\|\bar{\boldsymbol{\Delta}}_T\|_\infty \leq \mathbb{E}\max\{\|\bar{\boldsymbol{\Delta}}_T^1\|_\infty, \|\bar{\boldsymbol{\Delta}}_T^2\|_\infty\}$. Hence, bounding $\|\bar{\boldsymbol{\Delta}}_T\|_\infty$ in expectation is reduced to bound the maximum between $\|\bar{\boldsymbol{\Delta}}_T^1\|_\infty$ and $\|\bar{\boldsymbol{\Delta}}_T^2\|_\infty$. Given $\bar{\boldsymbol{\Delta}}_T^1$ and $\bar{\boldsymbol{\Delta}}_T^2$ are defined in a similar way (see Lemma C.1), they share a similar error decomposition.

F.1 Error Decomposition

Setting $r = 1$ in (36), we obtain

$$\bar{\boldsymbol{\Delta}}_T^1 = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Delta}_t^1 = \frac{1}{\eta_0 T}(\boldsymbol{A}_0^T - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{T}\sum_{j=1}^{T}\boldsymbol{A}_j^T\left(\boldsymbol{Z}_j + \gamma\boldsymbol{D}_{j-1}^1\right).$$

Similar to (38), we decompose $\bar{\boldsymbol{\Delta}}_T^1$ into five separate terms

$$\begin{aligned}\bar{\boldsymbol{\Delta}}_T^1 &= \frac{1}{\eta_0 T}(\boldsymbol{A}_0^T - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{T}\sum_{j=1}^{T}\boldsymbol{G}^{-1}\boldsymbol{Z}_j + \frac{1}{T}\sum_{j=1}^{T}(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j \\ &\quad + \frac{\gamma}{T}\sum_{j=1}^{T}\boldsymbol{A}_j^T(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*) + \frac{\gamma}{T}\sum_{j=1}^{T}\boldsymbol{A}_j^T(\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{j-1} \\ &:= \mathcal{T}_0 + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4.\end{aligned} \tag{68}$$

Here one should distinguish $\mathcal{T}_i$ with ψ_i, the former a random variable and the latter a random function. Comparing (35) and (40), we find that $\boldsymbol{D}_{j-1}^1 = \boldsymbol{D}_{j-1}^2 + (\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{j-1}$. Repeating the same argument to $\bar{\boldsymbol{\Delta}}_T^2$, we obtain

$$
\begin{aligned}
\bar{\boldsymbol{\Delta}}_T^2 = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\Delta}_t^2 &= \frac{1}{\eta_0 T}(\boldsymbol{A}_0^T - \eta_0 \boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{T}\sum_{j=1}^T \boldsymbol{A}_j^T\left(\boldsymbol{Z}_j + \gamma \boldsymbol{D}_{j-1}^2\right) \\
&= \frac{1}{\eta_0 T}(\boldsymbol{A}_0^T - \eta_0 \boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{T}\sum_{j=1}^T \boldsymbol{G}^{-1}\boldsymbol{Z}_j + \frac{1}{T}\sum_{j=1}^T (\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j \\
&\quad + \frac{\gamma}{T}\sum_{j=1}^T \boldsymbol{A}_j^T(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*) \\
&= \mathcal{T}_0 + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3.
\end{aligned} \tag{69}
$$

Here $\{\mathcal{T}_i\}_{i=0}^3$ are exactly the same as in (68). Putting the pieces together, we have

$$
\mathbb{E}\|\bar{\boldsymbol{\Delta}}_T\|_\infty \le \mathbb{E}\max\{\|\bar{\boldsymbol{\Delta}}_T^1\|_\infty, \|\bar{\boldsymbol{\Delta}}_T^2\|_\infty\} \le \sum_{i=0}^4 \mathbb{E}\|\mathcal{T}_i\|_\infty. \tag{70}
$$

F.2 Bounding the Separate Terms

For $\|\mathcal{T}_0\|_\infty$. Recall that $C_0 = \sup_{T\ge j\ge 0}\|\boldsymbol{A}_j^T\|_\infty$. Since $\eta_0 = 1 \le C_0$, it is obvious that

$$
\|\mathcal{T}_0\|_\infty = \frac{1}{\eta_0 T}\|(\boldsymbol{A}_0^T - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0\|_\infty \le \frac{1}{\eta_0 T}(\|\boldsymbol{A}_0^T\|_\infty + \eta_0)\|\boldsymbol{\Delta}_0\|_\infty \le \frac{2C_0}{1-\gamma}\frac{1}{T}. \tag{71}
$$

For $\|\mathcal{T}_1\|_\infty$. We apply (85) in Lemma H.1 to bound $\mathcal{T}_1 := \frac{1}{T}\sum_{j=1}^T \boldsymbol{G}^{-1}\boldsymbol{Z}_j$. Indeed, by setting $\boldsymbol{B}_j \equiv \boldsymbol{I}, \boldsymbol{X}_j = \frac{1}{T}\boldsymbol{G}^{-1}\boldsymbol{Z}_j$, we have $B = 1, X = \frac{1}{(1-\gamma)^2 T}$ and $\|\boldsymbol{W}_T\|_\infty \le \frac{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty}{T}$ defined therein. Hence,

$$
\mathbb{E}\|\mathcal{T}_1\|_\infty \le 6\sqrt{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln(2D)}{T}} + \frac{4\ln(6D)}{3(1-\gamma)^2 T}. \tag{72}
$$

For $\|\mathcal{T}_2\|_\infty$. We also apply (85) in Lemma H.1 to analyze $\mathcal{T}_2 := \frac{1}{T}\sum_{j=1}^T(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j$. Indeed, by setting $\boldsymbol{B}_j = \boldsymbol{A}_j^T - \boldsymbol{G}^{-1}, \boldsymbol{X}_j = \frac{1}{T}\boldsymbol{Z}_j$, we have $B = 2C_0, X = \frac{1}{(1-\gamma)T}$ and $\|\boldsymbol{W}_T\|_\infty \le \frac{1}{T^2}\sum_{j=1}^T \|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_\infty^2 \|\mathrm{Var}(\boldsymbol{Z})\|_\infty$ defined therein. Hence,

$$
\mathbb{E}\|\mathcal{T}_2\|_\infty \le 6\sqrt{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}\sqrt{\frac{\ln(2D)}{T}}\sqrt{\frac{1}{T}\sum_{j=1}^T\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_\infty^2} + \frac{8C_0\ln(6D)}{3(1-\gamma)T}. \tag{73}
$$

For $\|\mathcal{T}_3\|_\infty$. We apply (86) in Lemma H.1 to analyze $\mathcal{T}_3 := \frac{\gamma}{T}\sum_{j=1}^T \boldsymbol{A}_j^T(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)$. Because $\mathcal{T}_3$ is more complex than $\mathcal{T}_1$ and $\mathcal{T}_2$, we defer the detailed proof in Appendix F.5.

Lemma F.1.

$$
\mathbb{E}\|\mathcal{T}_3\|_\infty \le 4\gamma C_0\sqrt{\frac{\ln(2DT^2)}{T}}\cdot\sqrt{\frac{1}{T}\sum_{j=1}^T \mathbb{E}\|\boldsymbol{\Delta}_{j-1}\|_\infty^2} + \frac{32\gamma C_0\ln(3DT^2)}{3(1-\gamma)T}. \tag{74}
$$

where C_0 is the uniform bound given in Lemma C.3 and $D = |\mathcal{S}\times\mathcal{A}|$.

For $\|\mathcal{T}_4\|_\infty$. We have already analyzed $\mathcal{T}_4 := \frac{\gamma}{T}\sum_{j=1}^T \boldsymbol{A}_j^T(\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{j-1}$ in Lemma C.7. It follows that

$$
\mathbb{E}\|\mathcal{T}_4\|_\infty = \frac{1}{\sqrt{T}}\mathbb{E}\|\psi_5(1)\|_\infty \le \frac{1}{\sqrt{T}}\mathbb{E}\|\psi_5\|_{\sup} \le \gamma L C_0\cdot\frac{1}{T}\sum_{j=1}^T \mathbb{E}\|\boldsymbol{\Delta}_{j-1}\|_\infty^2. \tag{75}
$$

Remark F.1. *Under Assumption 3.1 3.2 and 3.3, we assert that $\sqrt{T}\mathbb{E}\|\mathcal{T}_i\| = o(1)$ for $i = 0, 2, 3, 4$. It is handy to verify $\sqrt{T}\|\mathcal{T}_0\| = o(1)$. Lemma C.2 implies $\frac{1}{T}\sum_{j=1}^T\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_\infty^2 = o(1)$, by which we conclude $\sqrt{T}\mathbb{E}\|\mathcal{T}_2\| = o(1)$. Theorem E.1 shows $\frac{1}{\sqrt{T}}\sum_{t=0}^T\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \to 0$ when we use the general step size. We then know that both $\sqrt{T}\mathbb{E}\|\mathcal{T}_3\|$ and $\sqrt{T}\mathbb{E}\|\mathcal{T}_4\|$ converge to zero when T goes to infinity.*

F.3 Specific Rates for Two Step Sizes

(I) Linearly rescaled step size. If we use a linear rescaled step size, i.e., $\eta_t = \frac{1}{1+(1-\gamma)t}$ (equivalently $\widetilde{\eta}_t = \frac{1-\gamma}{1+(1-\gamma)t}$), then Lemma C.3 and Lemma C.4 give

$$C_0 = \frac{2}{1-\gamma}\ln(1+(1-\gamma)T) = \mathcal{O}\left(\frac{\ln T}{1-\gamma}\right) \text{ and } \frac{1}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_\infty^2 \le \frac{25}{(1-\gamma)^2}.$$

Hiding constant factors in c, Theorem E.3 gives

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \le c\left[\frac{\|\boldsymbol{\Delta}_0\|_\infty^2}{(1-\gamma)^2}\frac{1}{T} + \frac{\ln(2eD)}{(1-\gamma)^5}\frac{\ln^2(eT)}{T}\right].$$

Hence, combining these bounds with (71), (72), (73), (74), and (75), we have

$$\begin{aligned}
\mathbb{E}\|\bar{\boldsymbol{\Delta}}_T\|_\infty &= \mathcal{O}\Bigg(\frac{\ln T}{(1-\gamma)^2 T} + \sqrt{\frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}}\sqrt{\frac{\ln D}{T}} + \frac{\ln D}{(1-\gamma)^2}\frac{\ln T}{T} \\
&\quad + \frac{\gamma\ln T\sqrt{\ln(DT)}}{(1-\gamma)^3}\left(\frac{1}{T} + \sqrt{\frac{\ln D}{1-\gamma}\frac{\ln T}{T}}\right) + \frac{\gamma\ln(DT)}{(1-\gamma)^2}\frac{\ln T}{T} \\
&\quad + + \frac{\gamma L\ln T}{1-\gamma}\left(\frac{1}{(1-\gamma)^4}\frac{1}{T} + \frac{\ln D}{(1-\gamma)^5}\frac{\ln^2 T}{T}\right)\Bigg) \\
&= \mathcal{O}\left(\sqrt{\frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}}\sqrt{\frac{\ln D}{T}}\right) + \widetilde{\mathcal{O}}\left(\frac{L}{(1-\gamma)^6}\frac{1}{T}\right),
\end{aligned}$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides polynomial dependence on logarithmic terms namely $\ln D$ and $\ln T$. Here we use $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty \le \frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}$ to simplify the final inequality.

(II) Polynomial step size. If we choose a polynomial step size, i.e., $\eta_t = t^{-\alpha}$ with $\alpha \in (0.5, 1)$ for $t \ge 1$ and $\eta_0 = 1$, then hiding constant factors in c, Lemma C.3 and Lemma C.4 give

$$C_0 = \mathcal{O}\left(\frac{1}{(1-\gamma)^{\frac{1}{1-\alpha}}}\right)$$

$$\sqrt{\frac{1}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_\infty^2} = \mathcal{O}\left(\frac{1}{(1-\gamma)^{1+\frac{1}{1-\alpha}}}\frac{1}{\sqrt{T}} + \frac{1}{(1-\gamma)^2}\frac{1}{T^{1-\alpha}} + \frac{1}{(1-\gamma)^{\frac{3}{2}}}\frac{1}{T^{\frac{1-\alpha}{2}}}\right),$$

where $\mathcal{O}(\cdot)$ hides constant factors on α. Theorem E.3 gives

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \le \mathcal{O}\left(\frac{\ln D}{(1-\gamma)^{3+\frac{1}{1-\alpha}}}\frac{1}{T} + \frac{\ln D}{(1-\gamma)^4}\frac{1}{T^\alpha}\right).$$

Hence, combining these bounds with (71), (72), (73), (74), and (75), we have

$$\begin{aligned}
\mathbb{E}\|\bar{\boldsymbol{\Delta}}_T\|_\infty = \mathcal{O}\Bigg(& \frac{1}{(1-\gamma)^{1+\frac{1}{1-\alpha}}T} + \sqrt{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}} + \frac{\ln(D)}{(1-\gamma)^2 T} \\
& + \sqrt{\frac{\ln D}{(1-\gamma)^2 T}}\left(\frac{1}{(1-\gamma)^{1+\frac{1}{1-\alpha}}}\frac{1}{\sqrt{T}} + \frac{1}{(1-\gamma)^2}\frac{1}{T^{1-\alpha}} + \frac{1}{(1-\gamma)^{\frac{3}{2}}}\frac{1}{T^{\frac{1-\alpha}{2}}} \right) \\
& + \frac{\gamma}{(1-\gamma)^{\frac{1}{1-\alpha}}}\sqrt{\frac{\ln(DT)}{T}}\left(\frac{\sqrt{\ln D}}{(1-\gamma)^{1.5+\frac{1}{2(1-\alpha)}}}\frac{1}{\sqrt{T}} + \frac{\sqrt{\ln D}}{(1-\gamma)^2}\frac{1}{T^{\frac{\alpha}{2}}} \right) \\
& + \frac{\gamma}{(1-\gamma)^{1+\frac{1}{1-\alpha}}}\frac{\ln DT}{T} + \frac{\gamma L}{(1-\gamma)^{\frac{1}{1-\alpha}}}\left(\frac{\ln D}{(1-\gamma)^{3+\frac{1}{1-\alpha}}}\frac{1}{T} + \frac{\ln D}{(1-\gamma)^4}\frac{1}{T^\alpha} \right)\Bigg) \\
= \mathcal{O}\Bigg(& \sqrt{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}} + \frac{\sqrt{\ln D}}{(1-\gamma)^3}\frac{1}{T^{1-\frac{\alpha}{2}}}\Bigg) + \widetilde{\mathcal{O}}\left(\frac{L}{(1-\gamma)^{3+\frac{2}{1-\alpha}}}\frac{1}{T} + \frac{\gamma L}{(1-\gamma)^{4+\frac{1}{1-\alpha}}}\frac{1}{T^\alpha} \right),
\end{aligned}$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides polynomial dependence on logarithmic terms, namely $\ln D$ and $\ln T$. Here we use $\|\mathrm{Var}(\boldsymbol{Z})\|_\infty \le \frac{1}{(1-\gamma)^2}$, $T^{-\frac{1+\alpha}{2}} \le T^{-\alpha}$ to simplify the final inequality.

F.4 A Useful Inequality

The following is a useful inequality which will be used frequently in the subsequent proof.

Lemma F.2. *For any matrices $\boldsymbol{A}, \boldsymbol{V}$ with a compatible order, we have*

$$\|\mathrm{diag}(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^\top)\|_\infty \le \|\boldsymbol{V}\|_{\max}\|\boldsymbol{A}\|_\infty^2, \tag{76}$$

where $\|\boldsymbol{V}\|_{\max} = \max_{i,k}|\boldsymbol{V}(i,k)|$.

Proof of Lemma F.2. For any diagonal entry i, it follows that

$$\begin{aligned}
|(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^\top)(i,i)| &= \left|\sum_l (\boldsymbol{A}\boldsymbol{V})(i,l)\boldsymbol{A}(i,l)\right| = \left|\sum_l\sum_k \boldsymbol{A}(i,k)\boldsymbol{V}(k,l)\boldsymbol{A}(i,l)\right| \\
&\le \sum_l\sum_k |\boldsymbol{A}(i,k)|\cdot|\boldsymbol{V}(k,l)|\cdot|\boldsymbol{A}(i,l)| \\
&\le \|\boldsymbol{V}\|_{\max}\sum_k|\boldsymbol{A}(i,k)|\cdot\sum_l|\boldsymbol{A}(i,l)| \\
&\le \|\boldsymbol{V}\|_{\max}\|\boldsymbol{A}\|_\infty^2.
\end{aligned}$$

□

F.5 Proof of Lemma F.1

Proof of Lemma F.1. Recall that $\mathcal{T}_3 = \frac{\gamma}{T}\sum_{j=1}^T \boldsymbol{A}_j^T(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)$ and $\mathcal{F}_j$ is the σ-field generated by all randomness before (and including) iteration j. We will apply Lemma H.1 to prove our lemma. Using the notation defined therein, we set $\boldsymbol{X}_j = \frac{\gamma}{T}(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)$ and $\boldsymbol{B}_j = \boldsymbol{A}_j^T$. Clearly, $\{\boldsymbol{X}_j\}_{j\ge 0}$ is a martingale difference sequence since $\mathbb{E}[\boldsymbol{X}_j|\mathcal{F}_{j-1}] = \frac{\gamma}{T}\mathbb{E}[\boldsymbol{P}_j - \boldsymbol{P}|\mathcal{F}_{j-1}](\boldsymbol{V}_{j-1} - \boldsymbol{V}^*) = \boldsymbol{0}$. As a result, $X = \frac{4\gamma}{T(1-\gamma)}$, $B = C_0$, $D = |\mathcal{S}\times\mathcal{A}|$ and $\boldsymbol{U}_j = \mathrm{Var}[\boldsymbol{X}_j|\mathcal{F}_{j-1}]$.[8]

Recall that $\boldsymbol{W}_T = \mathrm{diag}(\sum_{j=1}^T \boldsymbol{B}_j\boldsymbol{U}_j\boldsymbol{B}_j^\top)$. To upper bound $\mathbb{E}\|\boldsymbol{W}_T\|_\infty$, we aim to find a upper bound for $\|\boldsymbol{W}_T\|_\infty$. We first note that

$$\|\boldsymbol{W}_T\|_\infty = \left\|\mathrm{diag}\left(\sum_{j=1}^T \boldsymbol{B}_j\boldsymbol{U}_j\boldsymbol{B}_j^\top\right)\right\|_\infty \le \sum_{j=1}^T \left\|\mathrm{diag}\left(\boldsymbol{B}_j\boldsymbol{U}_j\boldsymbol{B}_j^\top\right)\right\|_\infty \le \sum_{j=1}^T \|\boldsymbol{B}_j\|_\infty^2\|\boldsymbol{U}_j\|_{\max}.$$

[8]To distinguish $\mathrm{Var}[\boldsymbol{X}_j|\mathcal{F}_{j-1}]$ and the value function $\boldsymbol{V}_j$, we use $\boldsymbol{U}_j$ to denote the conditional variance.

Here the last inequality uses (76). To bound $\|\boldsymbol{U}_j\|_{\max}$, we find that for any $i \neq k$, $\boldsymbol{U}_j(i,k) = \mathbb{E}[\boldsymbol{e}_i^\top \boldsymbol{X}_j \boldsymbol{X}_j^\top \boldsymbol{e}_k | \mathcal{F}_{j-1}] = 0$ due to each coordinate of $\boldsymbol{X}_j$ are independent conditioning on $\mathcal{F}_{j-1}$. Hence,

$$
\begin{aligned}
\|\boldsymbol{U}_j\|_{\max} &= \max_{i,k} |\boldsymbol{U}_j(i,k)| = \max_i |\boldsymbol{U}_j(i,i)| = \left\|\mathbb{E}[\mathrm{diag}(\boldsymbol{X}_j\boldsymbol{X}_j^\top)|\mathcal{F}_{j-1}]\right\|_\infty \\
&\le \mathbb{E}\left[\left\|\mathrm{diag}(\boldsymbol{X}_j\boldsymbol{X}_j^\top)\right\|_\infty \middle| \mathcal{F}_{j-1}\right] \overset{(a)}{\le} \mathbb{E}[\|\boldsymbol{X}_j\|_\infty^2 |\mathcal{F}_{j-1}] \\
&= \frac{\gamma^2}{T^2}\mathbb{E}[\|(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)\|_\infty^2 |\mathcal{F}_{j-1}] \\
&\le \frac{\gamma^2}{T^2}\|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_\infty^2 \mathbb{E}\|\boldsymbol{P}_j - \boldsymbol{P}\|_\infty^2 \overset{(b)}{\le} \frac{4\gamma^2}{T^2}\|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_\infty^2,
\end{aligned}
$$

where (a) again uses (76) and (b) uses $\|\boldsymbol{P}_j - \boldsymbol{P}\|_\infty \le \|\boldsymbol{P}_j\|_\infty + \|\boldsymbol{P}\|_\infty = 2$.

Putting the pieces together, we have

$$
\|\boldsymbol{W}_T\|_\infty \le \frac{4\gamma^2}{T}\sum_{j=1}^T \|\boldsymbol{B}_j\|_\infty^2 \|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_\infty^2 \le \frac{4\gamma^2 C_0^2}{T^2}\sum_{j=1}^T \|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_\infty^2,
$$

where we use $\sup_j \|\boldsymbol{B}_j\|_\infty \le B = \frac{C_0}{1-\gamma}$. The rest follows from (86) in Lemma H.1 by plugging the corresponding B, X, D and σ^2 and the inequality $\|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_\infty \le \|\boldsymbol{Q}_{j-1} - \boldsymbol{Q}^*\|_\infty = \|\boldsymbol{\Delta}_{j-1}\|_\infty$. □

G PROOF OF THE INFORMATION-THEORETIC LOWER BOUND

G.1 Proof of Theorem 4.1

The semiparametric model $\mathcal{P}_\theta \in \mathcal{P}_P \times \mathcal{P}_R$ described in Section 4 is described through an infinite-dimensional parameter $\theta = (\boldsymbol{P}, R)$, which is partitioned into a finite-dimensional parameter $\boldsymbol{P} \in \mathbb{R}^{D\times S}$ and an infinite-dimensional parameter R. The reason why R is infinite dimensional is because we don't specify the probability model of each $R(s,a)$, which is equivalent to considering the class of all p.d.f.'s on the interval $[0,1]$, which is infinite dimensional. The parameter of interest is a smooth function of θ, denoted by $\beta(\theta) = \boldsymbol{Q}^* \in \mathbb{R}^D$. To compute the semiparametric Cramer-Rao lower bound (see Definition 4.7 of [Vermeulen, 2011]), we need to compute

$$
\sup_{\mathcal{P}_\gamma \subset \mathcal{P}} \boldsymbol{\Gamma}(\gamma_0)\boldsymbol{I}(\gamma_0)^{-1}\boldsymbol{\Gamma}^\top(\gamma_0), \tag{77}
$$

where $\mathcal{P}_\gamma$ is any parametric submodel containing the truth, i.e., $\mathcal{P}_{\gamma_0} = \mathcal{P}_\theta$. Hence, under one kind of parameterization, the true model $\mathcal{P}_\theta$ can be recovered by setting $\gamma = \gamma_0$ in the parametric submodel $\mathcal{P}_\gamma$. Here, $\boldsymbol{\Gamma}(\gamma_0) = \frac{\partial \boldsymbol{Q}^*}{\partial \gamma}|_{\gamma=\gamma_0}$ is the score and $\boldsymbol{I}(\gamma_0)$ is the corresponding Fisher information matrix. Let $\gamma_0(R)$ (resp. $\gamma_0(\boldsymbol{P})$) be the finite-dimensional part of γ_0 that relates with R (resp. $\boldsymbol{P}$). Due to the (variational) independence between $\boldsymbol{P}$ and R, $\gamma_0(\boldsymbol{P})$ doesn't intersect with $\gamma_0(R)$. Hence, (77) can be divided into two parts

$$
\begin{aligned}
&\sup_{\mathcal{P}_\gamma(\boldsymbol{P})\subset \mathcal{P}_P} \boldsymbol{\Gamma}(\gamma_0(\boldsymbol{P}))\boldsymbol{I}(\gamma_0(\boldsymbol{P}))^{-1}\boldsymbol{\Gamma}^\top(\gamma_0(\boldsymbol{P})) + \sup_{\mathcal{P}_\gamma(R)\subset \mathcal{P}_R} \boldsymbol{\Gamma}(\gamma_0(R))\boldsymbol{I}(\gamma_0(R))^{-1}\boldsymbol{\Gamma}^\top(\gamma_0(R)) \\
\overset{(*)}{=}&\boldsymbol{\Gamma}(\boldsymbol{P})\boldsymbol{I}(\boldsymbol{P})^{-1}\boldsymbol{\Gamma}^\top(\boldsymbol{P}) + \sup_{\mathcal{P}_\gamma(R)\subset \mathcal{P}_R} \boldsymbol{\Gamma}(\gamma_0(R))\boldsymbol{I}(\gamma_0(R))^{-1}\boldsymbol{\Gamma}^\top(\gamma_0(R)),
\end{aligned}
$$

where $\mathcal{P}_\gamma(R)$ (resp. $\mathcal{P}_\gamma(\boldsymbol{P})$) denotes the parametric submodel depending only on R (resp. $\boldsymbol{P}$). The equality $(*)$ follows because in the case the parametric model $\mathcal{P}_P$ is the full model and the parametric Cramer-Rao lower bound is not affected by any one-to-one reparameterization. Here, $\boldsymbol{\Gamma}(\boldsymbol{P}) = \frac{\partial \boldsymbol{Q}^*}{\partial \boldsymbol{P}}$ and $\boldsymbol{I}(\boldsymbol{P})$ is the (constrained) information matrix.

In the following, we will first handle the parametric part (i.e., the transition kernel P) by computing the (constrained) information matrix and then cope with the nonparametric part (i.e., the random reward R) by using semiparametric tools. Combining the two parts together, we find that the semiparametric efficiency bound is

$$
\begin{aligned}
&\frac{1}{T}\cdot(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-1}\mathrm{Var}(\gamma\boldsymbol{P}_j\boldsymbol{V}^*)(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-\top} + \frac{1}{T}\cdot(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-1}\mathrm{Var}(\boldsymbol{r}_j)(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-\top} \\
=&\frac{1}{T}\cdot(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-1}\mathrm{Var}(\boldsymbol{Z}_j)(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-\top},
\end{aligned}
$$

using the notation $\boldsymbol{Z}_j = \boldsymbol{r}_j + \gamma\boldsymbol{P}_j\boldsymbol{V}^*$ and the independence of $\boldsymbol{r}_j$ and $\boldsymbol{P}_j$.

G.1.1 Parametric Part

We first investigate the Cramer-Rao lower bound for estimating $\boldsymbol{Q}^*$ using samples from $\{\boldsymbol{P}_t\}_{t\in[T]}$ whose distribution is determined by $\boldsymbol{P}\in\mathcal{P}$ with $\mathcal{P}$ defined in (15). Note that $\boldsymbol{P}\in\mathcal{P}$ is linearly constrained, i.e.,

$$\boldsymbol{h}(\boldsymbol{P})=0,$$

where $\boldsymbol{h}:\mathbb{R}^{D\times S}\to\mathbb{R}^D$ with its $(\tilde{s},\tilde{a})$-th coordinate of $\boldsymbol{h}$ given by

$$h_{\tilde{s},\tilde{a}}(\boldsymbol{P})=\sum_{s,a,s'}P(s'|s,a)1_{\{(s,a)=(\tilde{s},\tilde{a})\}}-1. \tag{78}$$

Hence, we encounter the Cramer-Rao lower bound for constrained parameters. Let $\boldsymbol{C}_T(\boldsymbol{P})$ is the inverse Fisher information matrix using T i.i.d. samples under the constraint $\boldsymbol{h}(\boldsymbol{P})=0$. Hence, $\boldsymbol{C}_T(\boldsymbol{P})=\frac{\boldsymbol{C}_1(\boldsymbol{P})}{T}$ and the constrained Cramer-Rao lower bound [Moore Jr, 2010] is

$$\boldsymbol{\Gamma}(\boldsymbol{P})\boldsymbol{I}(\boldsymbol{P})^{-1}\boldsymbol{\Gamma}^\top(\boldsymbol{P})=\left(\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}\right)^\top\boldsymbol{C}_T(\boldsymbol{P})\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}=\frac{1}{T}\cdot\left(\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}\right)^\top\boldsymbol{C}_1(\boldsymbol{P})\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}, \tag{79}$$

where $\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}$ is the partial derivatives computed ignoring the linear constraint $\boldsymbol{h}(\boldsymbol{P})=0$.

To give a precise formulation of the bound (79), we first compute $\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}$.

Lemma G.1. *Under Assumption 3.2, $\boldsymbol{Q}^*$ is differentiable w.r.t. $\boldsymbol{P}$ with the partial derivatives given by*

$$\frac{\partial Q^*(s,a)}{\partial P(s'|\tilde{s},\tilde{a})}=\gamma V^*(s')\cdot(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}((s,a),(\tilde{s},\tilde{a})).$$

We then compute $\boldsymbol{C}_1(\boldsymbol{P})$ via the following lemma.

Lemma G.2. *The (s,a)-th row of the random matrix $\boldsymbol{P}_t$ is given by $P_t(s'|s,a)=1_{\{s_t(s,a)=s'\}}$ where $s_t(s,a)$ is the generated next-state from (s,a) at iteration t with probability given as the (s,a)-th row of $\boldsymbol{P}$. Hence $\boldsymbol{P}=\mathbb{E}\boldsymbol{P}_t$ and $\boldsymbol{P}$ belongs to the following parametric space*

$$\mathcal{P}=\left\{\boldsymbol{P}\in\mathbb{R}^{D\times S}:\ P(s'|s,a)\geq 0\,\text{for all}\ (s,a,s')\ \text{and}\ \boldsymbol{h}(\boldsymbol{P})=\boldsymbol{0}\right\},$$

with $\boldsymbol{h}$ defined in (78). The constrained inverse Fisher information matrix $\boldsymbol{C}_1(\boldsymbol{P})$ is

$$\boldsymbol{C}_1(\boldsymbol{P})=\operatorname{diag}\left(\left\{\operatorname{diag}(P(\cdot|s,a))-P(\cdot|s,a)P(\cdot|s,a)^\top)\right\}_{(s,a)}\right).$$

By Lemma G.1 and G.2, we have

$$\begin{aligned}\left(\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}\right)^\top\boldsymbol{C}_1(\boldsymbol{P})\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}((s,a),(\bar{s},\bar{a}))&=\sum_{(\tilde{s},\tilde{a})}\gamma^2\boldsymbol{G}^{-1}((s,a),(\tilde{s},\tilde{a}))\boldsymbol{G}^{-1}((\bar{s},\bar{a}),(\tilde{s},\tilde{a}))\\&\quad\cdot\left(\sum_{\tilde{s}'}V^*(\tilde{s}')^2P(\tilde{s}'|\tilde{s},\tilde{a})-(\sum_{\tilde{s}'}V^*(\tilde{s}')P(\tilde{s}'|\tilde{s},\tilde{a}))^2\right).\end{aligned}$$

The Cramer-Rao lower bound is thus equal to

$$\left(\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}\right)^\top\boldsymbol{C}_T(\boldsymbol{P})\frac{\partial\boldsymbol{Q}^*}{\partial\boldsymbol{P}}=T\cdot(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}\operatorname{Var}(\gamma\boldsymbol{P}_j\boldsymbol{V}^*)(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-\top}.$$

At the end of this part, we provide the deferred proof for Lemma G.1 and G.2.

Proof of Lemma G.1. Notice that $\boldsymbol{Q}^*=\boldsymbol{R}+\gamma\boldsymbol{P}\boldsymbol{V}^*$. Then by the chain rule, we have

$$\begin{aligned}\frac{\partial Q^*(s,a)}{\partial P(s'|s,a)}&=\gamma V^*(s')+\gamma\sum_{s_1}P(s_1|s,a)\frac{\partial V^*(s_1)}{\partial P(s'|s,a)},\\\frac{\partial Q^*(s,a)}{\partial P(s'|\tilde{s},\tilde{a})}&=\gamma\sum_{s_1}P(s_1|\tilde{s},\tilde{a})\frac{\partial V^*(s_1)}{\partial P(s'|\tilde{s},\tilde{a})}\ \text{for any}\ (s,a)\neq(\tilde{s},\tilde{a}).\end{aligned}$$

Assumption 3.2 implies the optimal policy π^* is unique. Hence, using $V^*(s_1) = \max_a Q^*(s_1, a) = Q^*(s_1, \pi^*(s_1))$, we have

$$\frac{\partial V^*(s_1)}{\partial P(s'|s,a)} = \frac{\partial Q^*(s_1, \pi^*(s_1))}{\partial P(s'|s,a)}.$$

Notice that $\boldsymbol{P}^*((s,a),(\tilde{s},\tilde{a})) = P(\tilde{s}|s,a)1_{\{\tilde{a}=\pi^*(\tilde{s})\}}$. Putting all the pieces together and solving $\{\frac{\partial Q^*(s,a)}{\partial P(s'|\tilde{s},\tilde{a})}\}_{s,a,s',\tilde{s},\tilde{a}}$ from the linear system, we have

$$\frac{\partial Q^*(s,a)}{\partial P(s'|\tilde{s},\tilde{a})} = \gamma V^*(s') \cdot (\boldsymbol{I} - \gamma \boldsymbol{P}^*)^{-1}((s,a),(\tilde{s},\tilde{a})).$$

□

Proof of Lemma G.2. We write our the log-likelihood of sample $\boldsymbol{P}_t$ as

$$\log f_{\boldsymbol{P}}(\boldsymbol{P}_t) = \sum_{s,a,s'} 1_{\{s_t(s,a)=s'\}} \log P(s'|s,a),$$

which implies $\frac{\partial}{\partial \boldsymbol{P}} \log f_{\boldsymbol{P}}(\boldsymbol{P}_t) \in \mathbb{R}^{S^2A}$ with the (s,a,s')-th entry given by

$$\frac{\partial \log f_{\boldsymbol{P}}(\boldsymbol{P}_t)}{\partial P(s'|s,a)} = \frac{1_{\{s_t(s,a)=s'\}}}{P(s'|s,a)}. \tag{80}$$

By definition of the Fisher information matrix, we have

$$\boldsymbol{I}_1(\boldsymbol{P}) = \mathbb{E}\left\{ \frac{\partial}{\partial \boldsymbol{P}} \log f_{\boldsymbol{P}}(\boldsymbol{P}_t) \left[\frac{\partial}{\partial \boldsymbol{P}} \log f_{\boldsymbol{P}}(\boldsymbol{P}_t) \right]^\top \right\} \in \mathbb{R}^{S^2A \times S^2A},$$

which implies

$$\boldsymbol{I}_1(\boldsymbol{P})((s,a,s'),(\tilde{s},\tilde{a},\tilde{s}')) = \begin{cases} \frac{1_{\{s'=\tilde{s}'\}}}{P(s'|s,a)} & \text{if } (s,a) = (\tilde{s},\tilde{a}), \\ 1 & \text{if } (s,a) \neq (\tilde{s},\tilde{a}). \end{cases}$$

By definition of $\boldsymbol{h}(\boldsymbol{P})$, we rearrange $\boldsymbol{h}(\boldsymbol{P})$ into an $S^2A \times SA$ matrix given by

$$\boldsymbol{H}(\boldsymbol{P})((s,a,s'),(\tilde{s},\tilde{a})) := \frac{\partial h_{\tilde{s},\tilde{a}}(\boldsymbol{P})}{\partial P(s'|s,a)} = 1_{\{(\tilde{s},\tilde{a})=(s,a)\}}.$$

Let $\boldsymbol{U}(\boldsymbol{P}) \in \mathbb{R}^{S^2A \times (S^2A - SA)}$ be the orthogonal matrix whose column space is the orthogonal complement of the column space of $\boldsymbol{H}(\boldsymbol{P})$, which stands for $\boldsymbol{H}(\boldsymbol{P})^\top \boldsymbol{U}(\boldsymbol{P}) = \boldsymbol{0}$ and $\boldsymbol{U}(\boldsymbol{P})^\top \boldsymbol{U}(\boldsymbol{P}) = \boldsymbol{I}$. Using results in [Moore Jr, 2010], the constrained CRLB is

$$\boldsymbol{C}_1(\boldsymbol{P}) = \boldsymbol{U}(\boldsymbol{P}) \left(\boldsymbol{U}(\boldsymbol{P})^\top \boldsymbol{I}_1(\boldsymbol{P}) \boldsymbol{U}(\boldsymbol{P}) \right)^{-1} \boldsymbol{U}(\boldsymbol{P})^\top.$$

We define an auxiliary matrix $\boldsymbol{X} \in \mathbb{R}^{SA \times S^2A}$ satisfying

$$\boldsymbol{X}((s,a),(\tilde{s},\tilde{a},\tilde{s}')) = -\frac{1}{2} \cdot 1_{\{(s,a)\neq(\tilde{s},\tilde{a})\}}.$$

By $\boldsymbol{H}(\boldsymbol{P})^\top \boldsymbol{U}(\boldsymbol{P}) = \boldsymbol{0}$, we have

$$\begin{aligned} \boldsymbol{C}_1(\boldsymbol{P}) &= \boldsymbol{U}(\boldsymbol{P}) \left(\boldsymbol{U}(\boldsymbol{P})^\top (\boldsymbol{H}(\boldsymbol{P})\boldsymbol{X} + \boldsymbol{I}_1(\boldsymbol{P}) + \boldsymbol{X}^\top \boldsymbol{U}(\boldsymbol{P})^\top) \boldsymbol{U}(\boldsymbol{P}) \right)^{-1} \boldsymbol{U}(\boldsymbol{P})^\top \\ &:= \boldsymbol{U}(\boldsymbol{P}) \left(\boldsymbol{U}(\boldsymbol{P})^\top \boldsymbol{D}(\boldsymbol{P}) \boldsymbol{U}(\boldsymbol{P}) \right)^{-1} \boldsymbol{U}(\boldsymbol{P})^\top, \end{aligned}$$

where $\boldsymbol{D}(\boldsymbol{P})((s,a,s'),(s,a,s')) = 1/P(s'|s,a)$ and takes value 0 elsewhere. Now we reformulate $\boldsymbol{D}(\boldsymbol{P})$ as a block diagonal matrix $\boldsymbol{D}(\boldsymbol{P}) = \mathrm{diag}(\{\boldsymbol{D}_{(s,a)}\}_{(s,a)}) := \mathrm{diag}(\{1/P(\cdot|s,a)\}_{(s,a)})$ where $\boldsymbol{D}_{(s,a)}$ is a diagonal matrix with $\boldsymbol{D}_{(s,a)}(s',s') = 1/P(s'|s,a)$. Similarly, we have $\boldsymbol{H}(\boldsymbol{P}) = \mathrm{diag}(\{\boldsymbol{1}_S\}_{(s,a)})$, where $\boldsymbol{1}_S$ is an all-1 vector with dimension S, and $\boldsymbol{U}(\boldsymbol{P}) = \text{diag}(\{\boldsymbol{U}_{(s,a)}\}_{(s,a)})$, where $\boldsymbol{U}_{(s,a)} \in \mathbb{R}^{S \times S-1}$ satisfying $\boldsymbol{U}_{(s,a)}^\top \mathbf{1}_S = \mathbf{0}$. In this way, $\boldsymbol{C}_1(\boldsymbol{P})$ has a equivalent block diagonal formulation

$$\boldsymbol{C}_1(\boldsymbol{P}) = \text{diag}\left(\left\{\boldsymbol{U}_{(s,a)}\left(\boldsymbol{U}_{(s,a)}^\top \boldsymbol{D}_{(s,a)} \boldsymbol{U}_{(s,a)}\right)^{-1} \boldsymbol{U}_{(s,a)}^\top\right\}_{(s,a)}\right).$$

For each block (s,a) of $\boldsymbol{C}_1(\boldsymbol{P})$, the submatrix is exactly the constrained Cramer-Rao bound of a multinomial distribution $\boldsymbol{P}_{s,a} = \{P(\cdot|s,a)\}$, which is equal to $\text{diag}(\boldsymbol{P}_{s,a}) - \boldsymbol{P}_{s,a}\boldsymbol{P}_{s,a}^\top$. Therefore,

$$\boldsymbol{C}_1(\boldsymbol{P}) = \text{diag}\left(\left\{\text{diag}(P(\cdot|s,a)) - P(\cdot|s,a)P(\cdot|s,a)^\top\right)\right\}_{(s,a)}\right).$$

□

G.1.2 Nonparametric Part

Next, we move on discussing the efficiency on rewards. Unlike $\boldsymbol{P}_t$ that is generated according to a parametric model, the generating mechanism of $\boldsymbol{r}_t$ can be arbitrary. In other words, a finite dimensional parametric space is not enough to cover the possible distributions of $\boldsymbol{r}_t$. Thus, semiparametric theory is needed here. Fortunately, our interest parameter $\boldsymbol{Q}^* = (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\boldsymbol{r}$ is linear in $\boldsymbol{r} := \mathbb{E}\boldsymbol{r}_t$, implying only the expectation of $\boldsymbol{r}_t$ matters. In semiparametric theory [Van der Vaart, 2000, Tsiatis, 2006], the efficienct influence function for mean estimation is exatly the random variable minus its expectation. Lemma G.3 shows it is still true in our case.

Lemma G.3. *Let Assumption 3.2 hold. Given a random sample $\boldsymbol{r}_t$, the most efficient influence function for estimating $\boldsymbol{Q}^*(s,a)$ for any (s,a) is*

$$\phi(s,a) = (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}(\boldsymbol{r}_t - \boldsymbol{r})(s,a),$$

where $\boldsymbol{r} = \mathbb{E}\boldsymbol{r}_t$. Hence, the semiparametric efficiency bound of estimating $\boldsymbol{Q}^*$ with $\{\boldsymbol{r}_t\}_{t\in[T]}$ is

$$\sup_{\mathcal{P}_\gamma(R) \subset \mathcal{P}_R} \boldsymbol{\Gamma}(\gamma_0(R))\boldsymbol{I}(\gamma_0(R))^{-1}\boldsymbol{\Gamma}^\top(\gamma_0(R)) = \frac{1}{T} \cdot (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\text{Var}(\boldsymbol{r}_t)(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}.$$

Proof of Lemma G.3. As $r_t(s,a)$ are independent with different (s',a') pairs, we can only consider randomness of one pair (s,a).

Firstly, we consider a submodel family $\mathcal{P}_{R_\varepsilon}$ of $\mathcal{P}_R$ that is parameterized by ε such that when $\varepsilon = 0$, we recover the distribution of $R(s,a)$. That is $\mathcal{P}_{R_\varepsilon} = \{R_\varepsilon : \varepsilon \in [-\delta, \delta] \text{ and } R(s,a) = R_\varepsilon(s,a)|_{\varepsilon=0}\}$. This can be achieved by manipulating density functions of each $R(s,a)$. It is clear that $\mathcal{P}_{R_\varepsilon}$ is a parametric family on rewards and we can make use of results in parametric statistics for our purpose. By definition, we have for (s,a),

$$\begin{aligned}\left.\frac{\partial Q^*(s,a)}{\partial \varepsilon}\right|_{\varepsilon=0} &= \frac{\partial}{\partial \varepsilon}\left(\mathbb{E}r_t(s,a) + \gamma \sum_{s'} P(s'|s,a)Q^*(s',\pi^*(s'))\right)\Bigg|_{\varepsilon=0} \\ &= \left.\frac{\partial \mathbb{E}r_t(s,a)}{\partial \varepsilon}\right|_{\varepsilon=0} + \gamma \sum_{s'} P(s'|s,a)\left.\frac{\partial Q^*(s',\pi^*(s'))}{\partial \varepsilon}\right|_{\varepsilon=0}.\end{aligned}$$

For any $(\tilde{s},\tilde{a}) \neq (s,a)$, we have

$$\left.\frac{\partial Q^*(\tilde{s},\tilde{a})}{\partial \varepsilon}\right|_{\varepsilon=0} = \gamma \sum_{s'} P(s'|\tilde{s},\tilde{a})\left.\frac{\partial Q^*(s',\pi^*(s'))}{\partial \varepsilon}\right|_{\varepsilon=0}.$$

Recursively expanding the above terms like what we have done in Lemma G.1, we have

$$\left.\frac{\partial Q^*(\tilde{s},\tilde{a})}{\partial \varepsilon}\right|_{\varepsilon=0} = \left.\frac{\partial \mathbb{E}r_t(s,a)}{\partial \varepsilon}\right|_{\varepsilon=0} \cdot (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}((\tilde{s},\tilde{a}),(s,a)).$$

Let F_ε denote the cumulative distribution function of $R_\varepsilon(s,a)$. Then we have

$$\begin{aligned}\frac{\partial \mathbb{E} r_t(s,a)}{\partial \varepsilon}\bigg|_{\varepsilon=0} &= \int r_t(s,a)\frac{\partial}{\partial\varepsilon} dF_\varepsilon\bigg|_{\varepsilon=0}\\ &= \int (r_t(s,a)-r(s,a))\frac{\partial}{\partial\varepsilon}\log dF_\varepsilon\bigg|_{\varepsilon=0} dF_0,\end{aligned}$$

where $r(s,a)=\mathbb{E}r_t(s,a)$ and $\frac{\partial}{\partial\varepsilon}\log dF_\varepsilon$ is the score function. Therefore,

$$\frac{\partial Q^*(\tilde{s},\tilde{a})}{\partial\varepsilon}\bigg|_{\varepsilon=0} = \int \phi(\tilde{s},\tilde{a})\frac{\partial}{\partial\varepsilon}\log dF_\varepsilon\bigg|_{\varepsilon=0} dF_0, \tag{81}$$

where

$$\phi(\tilde{s},\tilde{a}) = (r_t - r)(s,a)\cdot(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}((\tilde{s},\tilde{a}),(s,a)).$$

Since the parametric submodel family $\mathcal{R}_\varepsilon$ is arbitrary, we conclude that the efficient influence function of $Q^*(\tilde{s},\tilde{a})$ is $\phi(\tilde{s},\tilde{a})$ by Theorem 2.2 in [Newey, 1990]. Finally, as $r_t(s,a)$ is independent with each other $r_t(s',a')$'s, our final result is obtained by summing the above equation over all (s,a). □

G.2 Proof of Theorem 4.2

Proof of Theorem 4.2. Recall that $\bar{\boldsymbol{\Delta}}_T = \frac{1}{T}\sum_{t=1}^T(\boldsymbol{Q}_T-\boldsymbol{Q}^*)$. Combining (67), (68) and (69), we have

$$\sqrt{T}(\mathcal{T}_0+\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3) \le \bar{\boldsymbol{\Delta}}_T^1 \le \sqrt{T}\bar{\boldsymbol{\Delta}}_T \le \sqrt{T}\bar{\boldsymbol{\Delta}}_T^2 \le \sqrt{T}(\mathcal{T}_0+\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3+\mathcal{T}_4),$$

where the inequality holds coordinate-wise. In Appendix F.2, we have analyze $\mathbb{E}\|\mathcal{T}_i\|_\infty$ with explicit upper bounds. It is easy to verify that $\sqrt{T}\mathbb{E}\|\mathcal{T}_i\| = o(1)$ for $i=0,2,3,4$ (see Remark F.1). Hence,

$$\bar{\boldsymbol{\Delta}}_T = \sqrt{T}\mathcal{T}_1 + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{T}}\sum_{t=1}^T(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}\boldsymbol{Z}_t + o_{\mathbb{P}}(1) := \frac{1}{\sqrt{T}}\sum_{t=1}^T\boldsymbol{\phi}(\boldsymbol{r}_t,\boldsymbol{P}_t)+o_{\mathbb{P}}(1),$$

where $\boldsymbol{Z}_t = (\boldsymbol{r}_t-\boldsymbol{r})+\gamma(\boldsymbol{P}_t-\boldsymbol{P})\boldsymbol{V}^*$ is the Bellman noise at iteration t. This implies $\bar{\boldsymbol{Q}}_T$ is asymptotically linear with the influence function $\boldsymbol{\phi}(\boldsymbol{r}_t,\boldsymbol{P}_t) := (\boldsymbol{I}-\gamma\boldsymbol{P}^*)^{-1}\boldsymbol{Z}_t$.

The remaining issue is to prove regularity. By definition, a RAL estimator is regular for a semiparametric model $\mathcal{P}=\mathcal{P}_P\times\mathcal{P}_R$ if it is a RAL estimator for every parametric submodel $\mathcal{P}_\gamma=\mathcal{P}_P\times\mathcal{P}_{R_\varepsilon}\subset\mathcal{P}$ where $\gamma=(\boldsymbol{P},\varepsilon)$ is the finite-dimensional parameter controlling $\mathcal{P}_\gamma$. In a parametric submodel $\mathcal{P}_P\times\mathcal{P}_{R_\varepsilon}$, by Theorem 2.2 in [Newey, 1990], for the asymptotically linear estimator $\bar{\boldsymbol{Q}}_T$ of $\boldsymbol{Q}^*$ which has the influence function

$$\boldsymbol{\phi}(\boldsymbol{r}_t,\boldsymbol{P}_t) = (\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}\left[(\boldsymbol{r}_t-\boldsymbol{r})+\gamma(\boldsymbol{P}_t-\boldsymbol{P})\boldsymbol{V}^*\right],$$

its regularity is equivalent to the equality

$$\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t,\boldsymbol{P}_t)S_\gamma^\top(\gamma_0) = \frac{\partial\boldsymbol{Q}^*}{\partial\gamma}\bigg|_{\gamma=\gamma_0}, \tag{82}$$

where $S_\gamma(\cdot)$ is the score function, $\gamma=(\boldsymbol{P}',\varepsilon)\in\mathcal{P}_P\times[-\delta,\delta]$ is the finite-dimensional parameter and $\gamma_0=(\boldsymbol{P},0)$ is the true underlying parameter. Since $\boldsymbol{P}$ and ε are variationally independent, $S_\gamma(\gamma_0)=(S_{\boldsymbol{P}}(\gamma_0),S_\varepsilon(\gamma_0))$.

For the transition kernel $\boldsymbol{P}$. Since our parametric space $\mathcal{P}_P$ has a linear constraint, it is not easy to compute the constrained score function. Hence, for $\boldsymbol{P}=\{P(s'|s,a)\}_{s,a,s'}$, we regard $\{P(s'|s,a)\}_{s,a,s'\neq s_0}$ as free parameters where $s_0\in\mathcal{S}$ is any fixed state and use it as our new parameter. For a fixed (s,a), once $P(s'|s,a)$ is determined for all $s'\neq s_0$, one can recover $P(s_0|s,a)$ by $P(s_0|s,a)=1-\sum_{s'\neq s_0}P(s'|s,a)$. In this way, each $\{P(s'|s,a)\}_{s'\neq s_0}$ lies in a open set. We still denote the set collecting all feasible $\{P(s'|s,a)\}_{s,a,s'\neq s_0}$ as $\mathcal{P}$, but readers should remember that current $\boldsymbol{P}=\{P(s'|s,a)\}_{s,a,s'\neq s_0}\in\mathbb{R}^{SA\times(S-1)}$. From (80) and under our new notation of $\boldsymbol{P}$, $S_{\boldsymbol{P}}(\gamma_0)\in\mathbb{R}^{SA(S-1)}$ with entries given by

$$S_{\boldsymbol{P}}(\gamma_0)(s,a,s') = \frac{1_{\{s_t(s,a)=s'\}}}{P(s'|s,a)} - \frac{1_{\{s_t(s,a)=s_0\}}}{P(s_0|s,a)} \text{ for any } s'\neq s_0.$$

By Lemma G.1 and the chain rule, it follows that $\frac{\partial \boldsymbol{Q}^*}{\partial \boldsymbol{P}} \in \mathbb{R}^{SA\times SA(S-1)}$ and its $(\tilde{s}, \tilde{a}, s')$-th column is

$$\gamma(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}(\cdot, (\tilde{s}, \tilde{a}))\,[V^*(s') - V^*(s_0)]\,. \tag{83}$$

Since $(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}$ has a full rank (i.e., SA), it is easy to see that $\frac{\partial \boldsymbol{Q}^*}{\partial \boldsymbol{P}}$ also has rank SA by varying $(\tilde{s}, \tilde{a})$ and fixing s', s_0 in (83). On the other hand, the $(\tilde{s}, \tilde{a}, s')$-th column of $\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)S_{\boldsymbol{P}}(\theta_0)^\top$ is

$$\begin{aligned}
(\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)S_{\boldsymbol{P}}(\gamma_0)^\top)(\cdot, (\tilde{s}, \tilde{a}, s')) &= \mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)\left[\frac{1_{\{s_t(s,a)=s'\}}}{P(s'|s,a)} - \frac{1_{\{s_t(s,a)=s_0\}}}{P(s_0|s,a)}\right] \\
&= \gamma(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\mathbb{E}(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*\left[\frac{1_{\{s_t(s,a)=s'\}}}{P(s'|s,a)} - \frac{1_{\{s_t(s,a)=s_0\}}}{P(s_0|s,a)}\right] \\
&= \gamma(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}(\cdot, (\tilde{s}, \tilde{a}))\,[V^*(s') - V^*(s_0)]\,,
\end{aligned}$$

where the last equality uses the following result. By direct calculation, the (s,a)-th entry of $\mathbb{E}(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*\left[\frac{1_{\{s_t(s,a)=s'\}}}{P(s'|s,a)} - \frac{1_{\{s_t(s,a)=s_0\}}}{P(s_0|s,a)}\right]$ is 0 for all $(s,a) \neq (\tilde{s}, \tilde{a})$ (due to independence) and the $(\tilde{s}, \tilde{a})$-th entry is $V^*(s') - V^*(s_0)$. Indeed, the $(\tilde{s}, \tilde{a})$-th entry of the mentioned matrix is

$$\begin{aligned}
&\mathbb{E}\sum_{i\in\mathcal{S}}(1_{\{s_t(s,a)=i\}} - P(i|s,a))V^*(i)\left[\frac{1_{\{s_t(s,a)=s'\}}}{P(s'|s,a)} - \frac{1_{\{s_t(s,a)=s_0\}}}{P(s_0|s,a)}\right] \\
&= \left(V^*(s') - \sum_{i\neq s_0} P(i|s,a)V^*(i)\right) + \sum_{i\in\mathcal{S}} P(i|s,a)V^*(i) = V^*(s') - V^*(s_0).
\end{aligned}$$

Therefore, combining the results for all $(\tilde{s}, \tilde{a}, s')(s' \neq s_0)$, we have

$$\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)S_{\boldsymbol{P}}(\gamma_0)^\top = \frac{\partial \boldsymbol{Q}^*}{\partial \boldsymbol{P}},$$

which implies (82) holds for the $\boldsymbol{P}$ part.

For the random reward R. Using the notation in the proof of Lemma G.3, $S_\varepsilon(\gamma_0) = \frac{\partial}{\partial\varepsilon}\log dF_\varepsilon|_{\varepsilon=0}$. By (81), we have

$$\left.\frac{\partial \boldsymbol{Q}^*}{\partial \varepsilon}\right|_{\varepsilon=0} = \mathbb{E}(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}(\boldsymbol{r}_t - \boldsymbol{r})S_\varepsilon(\gamma_0) = \mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)S_\varepsilon(\gamma_0)$$

which implies (82) holds for the ε part.

$\mathcal{P}_{R_\varepsilon}$ can be arbitrary, so (82) holds for all parametric submodels. This means $\bar{\boldsymbol{Q}}_T$ is regular for all parametric submodels and thus is regular for our semiparametric model. □

H A USEFUL CONCENTRATION INEQUALITY

We introduce a useful concentration inequality in this section. It captures the expectation and high probability concentration of a martingale difference sum in terms of $\|\cdot\|_\infty$. It uses a similar idea of Theorem 4 in Li et al. [2021a] and is built on Freedman's inequality [Freedman, 1975] and the union bound.

Lemma H.1. *Assume $\{\boldsymbol{X}_j\} \subseteq \mathbb{R}^d$ are martingale differences adapted to the filtration $\{\mathcal{F}_j\}_{j\ge 0}$ with zero conditional mean $\mathbb{E}[\boldsymbol{X}_j|\mathcal{F}_{j-1}] = \boldsymbol{0}$ and finite conditional variance $\boldsymbol{V}_j = \mathbb{E}[\boldsymbol{X}_j\boldsymbol{X}_j^\top|\mathcal{F}_{j-1}]$. Moreover, assume $\{\boldsymbol{X}_j\}_{j\ge 0}$ is uniformly bounded, i.e., $\sup_j\|\boldsymbol{X}_j\|_\infty \le X$. For any sequence of deterministic matrices $\{\boldsymbol{B}_j\}_{j\ge 0} \subseteq \mathbb{R}^{D\times d}$ satisfying $\sup_j\|\boldsymbol{B}_j\|_\infty \le B$, we define the weighted sum as*

$$\boldsymbol{Y}_T = \sum_{j=1}^T \boldsymbol{B}_j\boldsymbol{X}_j$$

and let $\boldsymbol{W}_T = \mathrm{diag}(\sum_{j=1}^T \boldsymbol{B}_j\boldsymbol{V}_j(\boldsymbol{B}_j)^\top)$ be a diagonal matrix that collects conditional quadratic variations. Then, it follows that

$$\mathbb{P}\left(\|\boldsymbol{Y}_T\|_\infty \ge \frac{2BX}{3}\ln\frac{2D}{\delta} + \sqrt{2\sigma^2\ln\frac{2D}{\delta}} \text{ and } \|\boldsymbol{W}_T\|_\infty \le \sigma^2\right) \le \delta \tag{84}$$

$$\mathbb{E}\|\boldsymbol{Y}_T\|_\infty 1_{\{\|\boldsymbol{W}_T\|_\infty \le \sigma^2\}} \le 6\sigma\sqrt{\ln(2D)} + \frac{4BX}{3}\ln(6D). \tag{85}$$

Generally, we have

$$\mathbb{E}\|\boldsymbol{Y}_T\|_\infty \le \frac{8BX}{3}\ln(3DT^2) + 2\sqrt{\mathbb{E}\|\boldsymbol{W}_T\|_\infty}\sqrt{\ln(2DT^2)}. \tag{86}$$

Proof of Lemma H.1. Fixing any $i \in [D]$, we denote the i-th row of $\boldsymbol{B}_j$ as $\boldsymbol{b}_j^\top$. For simplicity, we omit the dependence of $\boldsymbol{b}_j$ on i. Then the i-th coordinate of $\boldsymbol{Y}_T$ is $\boldsymbol{Y}_T(i) = \sum_{j=1}^T \boldsymbol{b}_j^\top \boldsymbol{X}_j$ and $\boldsymbol{W}_T(i,i) = \sum_{j=1}^T \boldsymbol{b}_j^\top \boldsymbol{V}_j \boldsymbol{b}_j$. Clearly $\{\boldsymbol{b}_j^\top \boldsymbol{X}_j\}$ is a scalar martingale difference with $\boldsymbol{W}_T(i,i) = \sum_{j=1}^T \mathbb{E}[(\boldsymbol{b}_j^\top \boldsymbol{X}_j)^2|\mathcal{F}_{j-1}]$ the quadratic variation and $|\boldsymbol{b}_j^\top \boldsymbol{X}_j| \le \|\boldsymbol{b}_j\|_1 \|\boldsymbol{X}_j\|_\infty \le \|\boldsymbol{B}_j\|_\infty \|\boldsymbol{X}_j\|_\infty = BX$ the uniform upper bound. By Freedman's inequality [Freedman, 1975], it follows that

$$\mathbb{P}(|\boldsymbol{Y}_T(i)| \ge \tau \text{ and } \boldsymbol{W}_T(i,i) \le \sigma^2) \le 2\exp\left(-\frac{\tau^2/2}{\sigma^2 + BX\tau/3}\right).$$

Then by the union bound, we have

$$\begin{aligned}
\mathbb{P}(\|\boldsymbol{Y}_T\|_\infty \ge \tau \text{ and } \|\boldsymbol{W}_T\|_\infty \le \sigma^2) &= \mathbb{P}\left(\max_{i\in[D]} |\boldsymbol{Y}_T(i)| \ge \tau \text{ and } \max_{i\in[D]} |\boldsymbol{W}_T(i,i)| \le \sigma^2\right) \\
&\le \sum_{i\in[D]} \mathbb{P}\left(|\boldsymbol{Y}_T(i)| \ge \tau \text{ and } \max_{i\in[D]} |\boldsymbol{W}_T(i,i)| \le \sigma^2\right) \\
&\le \sum_{i\in[D]} \mathbb{P}\left(|\boldsymbol{Y}_T(i)| \ge \tau \text{ and } |\boldsymbol{W}_T(i,i)| \le \sigma^2\right) \\
&\le 2D\exp\left(-\frac{\tau^2/2}{\sigma^2 + BX\tau/3}\right).
\end{aligned} \tag{87}$$

Solving for τ such that the right-hand side of (87) is equal to δ gives

$$\tau = \frac{BX}{3}\ln\frac{2D}{\delta} + \sqrt{\left(\frac{BX}{3}\ln\frac{2D}{\delta}\right)^2 + 2\sigma^2\ln\frac{2D}{\delta}}.$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ gives an upper bound on τ and provides the high probability result.

The tail bound of $\|\boldsymbol{Y}_T\|_\infty 1_{\{\|\boldsymbol{W}_T\|_\infty \le \sigma^2\}}$ has already been derived in (87). For the expectation result, we refer to the conclusion of Exercise 2.8 (a) in [Wainwright, 2019a] which implies that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{Y}_T\|_\infty 1_{\{\|\boldsymbol{W}_T\|_\infty \le \sigma^2\}} &\le 2\sigma(\sqrt{\pi} + \sqrt{\ln(2D)}) + \frac{4BX}{3}(1 + \ln(2D)) \\
&\le 6\sigma\sqrt{\ln(2D)} + \frac{4BX}{3}\ln(6D),
\end{aligned}$$

where the last inequality uses $\sqrt{a} + \sqrt{b} \le \sqrt{2(a+b)}$.

For the last result, we aim to bound $\mathbb{E}\|\boldsymbol{Y}_T\|_\infty$ without the condition $\|\boldsymbol{W}_T\|_\infty \le \sigma^2$ for some positive number σ. We first assert that there exists a trivial upper bound for $\|\boldsymbol{W}_T\|_\infty$ which is $\|\boldsymbol{W}_T\|_\infty \le TB^2X^2$. This is because

$$\|\boldsymbol{W}_T\|_\infty = \left\|\text{diag}\left(\sum_{j=1}^T \boldsymbol{B}_j\boldsymbol{V}_j(\boldsymbol{B}_j)^\top\right)\right\|_\infty \le \sum_{j=1}^T \left\|\text{diag}\left(\boldsymbol{B}_j\boldsymbol{V}_j(\boldsymbol{B}_j)^\top\right)\right\|_\infty \overset{(a)}{\le} \|\boldsymbol{V}_j\|_{\max}\|\boldsymbol{B}_j\|_\infty^2 \overset{(b)}{\le} TB^2X^2,$$

where (a) uses Lemma F.2 and (b) is due to $\|\boldsymbol{V}_j\|_{\max} \le X^2$ for all $j \in [T]$. However, if we set $\sigma^2 = TB^2X^2$ in (85), the resulting expectation bound of $\mathbb{E}\|\boldsymbol{Y}_T\|_\infty$ has a poor dependence on T.

To refine the dependence, we adapt and modify the argument of Theorem 4 in Li et al. [2021a]. For any positive integer K, we define

$$\mathcal{H}_K = \left\{\|\boldsymbol{Y}_T\|_\infty \ge \frac{2BX}{3}\ln\frac{2DK}{\delta} + \sqrt{4\max\left\{\|\boldsymbol{W}_T\|_\infty, \frac{TB^2X^2}{2^K}\right\}\ln\frac{2DK}{\delta}}\right\}$$

and claim that we have $\mathbb{P}(\mathcal{H}_K) \le \delta$. We observe that the event $\mathcal{H}_K$ is contained within the union of the following K events: $\mathcal{H}_K \subseteq \cup_{k\in[K]}\mathcal{B}_k$ where for $0 \le k < K$, $\mathcal{B}_k$ is defined to be

$$\mathcal{B}_k = \left\{ \|\boldsymbol{Y}_T\|_\infty \ge \frac{2BX}{3}\ln\frac{2DK}{\delta} + \sqrt{2\frac{TB^2X^2}{2^{k-1}}\ln\frac{2DT}{\delta}} \text{ and } \frac{TB^2X^2}{2^k} \le \|\boldsymbol{W}_T\|_\infty \le \frac{TB^2X^2}{2^{k-1}} \right\}$$
$$\mathcal{B}_K = \left\{ \|\boldsymbol{Y}_T\|_\infty \ge \frac{2BX}{3}\ln\frac{2DK}{\delta} + \sqrt{2\frac{TB^2X^2}{2^{K-1}}\ln\frac{2DT}{\delta}} \text{ and } \|\boldsymbol{W}_T\|_\infty \le \frac{TB^2X^2}{2^{K-1}} \right\}.$$

Invoking (84) with a proper $\sigma^2 = \frac{TB^2X^2}{2^{k-1}}$ and $\delta = \frac{\delta}{K}$, we have $\mathbb{P}(\mathcal{B}_k) \le \frac{\delta}{K}$ for all $k \in [K]$. Taken this result together with the union bound gives $\mathbb{P}(\mathcal{H}_K) \le \sum_{k\in[K]}\mathbb{P}(\mathcal{B}_k) \le \delta$. Then we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{Y}_T\|_\infty &= \mathbb{E}\|\boldsymbol{Y}_T\|_\infty 1_{\mathcal{H}_K} + \mathbb{E}\|\boldsymbol{Y}_T\|_\infty 1_{\mathcal{H}_K^c}\\
&\overset{(a)}{\le} TBX\mathbb{P}(\mathcal{H}_K) + \mathbb{E}\left[\frac{2BX}{3}\ln\frac{2DK}{\delta} + \sqrt{4\max\left\{\|\boldsymbol{W}_T\|_\infty, \frac{TB^2X^2}{2^K}\right\}\ln\frac{2DK}{\delta}}\right]\\
&\overset{(b)}{\le} BX + \frac{2BX}{3}\ln(2DT^2) + 2\mathbb{E}\sqrt{\max\left\{\|\boldsymbol{W}_T\|_\infty, B^2X^2\right\}\ln(2DT^2)}\\
&\overset{(c)}{\le} BX + \frac{8BX}{3}\ln(2DT^2) + 2\mathbb{E}\sqrt{\|\boldsymbol{W}_T\|_\infty\ln(2DT^2)}\\
&\overset{(d)}{\le} \frac{8BX}{3}\ln(3DT^2) + 2\sqrt{\mathbb{E}\|\boldsymbol{W}_T\|_\infty}\sqrt{\ln(2DT^2)},
\end{aligned}$$

where (a) uses $\|\boldsymbol{Y}_T\|_\infty \le TBX$, (b) follows by setting $\delta = \frac{1}{T}$ and $K = \lceil \log_2 T\rceil \le T$, (c) uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, and (d) follows from Jensen's inequality and $\exp(\frac{3}{8}) \le \frac{3}{2}$. ☐

I PROOF FOR ENTROPY REGULARIZED Q-LEARNING

In this section, we provide the counterpart results for Q-Learning with entropy. Since the proof is almost similar to that of Q-Learning, we just provide a sketch for simplicity. Recall that the matrix-form of the update rule is

$$\widetilde{\boldsymbol{Q}}_t = (1-\eta_t)\widetilde{\boldsymbol{Q}}_{t-1} + \eta_t(\boldsymbol{r}_t + \gamma\boldsymbol{P}_t\mathcal{L}_\lambda\widetilde{\boldsymbol{Q}}_{t-1}).$$

It is easy to show $\mathcal{L}_\lambda$ is a 1-contraction with respect to $\|\cdot\|_\infty$.

I.1 Convergence Under the General Step Sizes

Theorem I.1. *Under Assumption 3.1 and using the general step size in Assumption 3.3, we have*

$$\lim_{T\to\infty}\frac{1}{\sqrt{T}}\sum_{t=0}^{T}\mathbb{E}\|\widetilde{\boldsymbol{Q}}_t - \boldsymbol{Q}_\lambda^*\|_\infty^2 = 0$$

where $\boldsymbol{Q}_\lambda^*$ is the unique fixed point of the regularized Bellman equation $\boldsymbol{Q}_\lambda^* = \boldsymbol{r} + \gamma\boldsymbol{P}\mathcal{L}_\lambda\boldsymbol{Q}_\lambda^*$.

Proof of Theorem I.1. Denote $\widetilde{\boldsymbol{\Delta}}_t = \widetilde{\boldsymbol{Q}}_t - \boldsymbol{Q}_\lambda^*$ for simplicity. We will show that $\lim_{T\to\infty}\frac{1}{\sqrt{T}}\sum_{t=0}^{T}\mathbb{E}\|\widetilde{\boldsymbol{\Delta}}_t\|_\infty^2 = 0$ for the sequence generated via (16). Similar to Theorem E.1, we notice that the update rule satisfies $\widetilde{\boldsymbol{Q}}_t = \widetilde{\boldsymbol{Q}}_{t-1} + \eta_t(\boldsymbol{r} + \gamma\boldsymbol{P}\mathcal{L}_\lambda\widetilde{\boldsymbol{Q}}_{t-1} - \widetilde{\boldsymbol{Q}}_{t-1} + \boldsymbol{\varepsilon}_t)$ where $\boldsymbol{\varepsilon}_t = \boldsymbol{r}_t - \boldsymbol{r} + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\mathcal{L}_\lambda\widetilde{\boldsymbol{Q}}_{t-1}$. Hence, $\mathbb{E}[\boldsymbol{\varepsilon}_t|\mathcal{F}_t] = \boldsymbol{0}$ and $\mathbb{E}[\|\boldsymbol{\varepsilon}_t\|_\infty^2|\mathcal{F}_t] \le 2\mathbb{E}\|\boldsymbol{r}_t - \boldsymbol{r}\|_\infty^2 + 2\gamma^2\mathbb{E}\|\boldsymbol{P}_t - \boldsymbol{P}\|_\infty^2\|\mathcal{L}_\lambda\widetilde{\boldsymbol{Q}}_{t-1}\|_\infty^2 := A + B\|\widetilde{\boldsymbol{Q}}_{t-1}\|_\infty^2$ with $A = 2\mathbb{E}\|\boldsymbol{r}_t - \boldsymbol{r}\|_\infty^2, B = 2\gamma^2\mathbb{E}\|\boldsymbol{P}_t - \boldsymbol{P}\|_\infty^2$. By Theorem E.2, we arrive the same inequality as (55). Following the same analysis therein, we can show $\lim_{T\to\infty}\frac{1}{\sqrt{T}}\sum_{t=0}^{T}\mathbb{E}\|\widetilde{\boldsymbol{\Delta}}_t\|_\infty^2 = 0$ under the general step size in Assumption 3.3. ☐

I.2 Establishment of FCLT in Proof of Theorem 6.1

Proof of Theorem 6.1. Since the analysis is almost similar to that in Theorem 3.1, we just specify the differences. The three-step analysis in Section 3.2 still applies here except that we show only modify the first step.

Similar error decomposition. Let $\widetilde{\mathbf{\Delta}}_t = \widetilde{\boldsymbol{Q}}_t - \boldsymbol{Q}_\lambda^*$. By the regularized Bellman equation $\boldsymbol{Q}_\lambda^* = \boldsymbol{r} + \gamma \boldsymbol{P}\mathcal{L}_\lambda \boldsymbol{Q}_\lambda^*$, it follows that

$$
\begin{aligned}
\widetilde{\mathbf{\Delta}}_t &= (1-\eta_t)\widetilde{\mathbf{\Delta}}_{t-1} + \eta_t \left[\boldsymbol{r}_t + \gamma \boldsymbol{P}_t \mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_{t-1} - (\boldsymbol{r} + \gamma \boldsymbol{P}\mathcal{L}_\lambda \boldsymbol{Q}_\lambda^*)\right] \\
&= (1-\eta_t)\widetilde{\mathbf{\Delta}}_{t-1} + \eta_t \left[(\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t \mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_{t-1} - \boldsymbol{P}\mathcal{L}_\lambda \boldsymbol{Q}_\lambda^*)\right] \\
&= (1-\eta_t)\widetilde{\mathbf{\Delta}}_{t-1} + \eta_t \left[\widetilde{\boldsymbol{Z}}_t + \gamma \boldsymbol{P}_t(\mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_{t-1} - \mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_\lambda^*)\right] \\
&= (1-\eta_t)\widetilde{\mathbf{\Delta}}_{t-1} + \eta_t \left[\widetilde{\boldsymbol{Z}}_t + \gamma \widetilde{\boldsymbol{Z}}_t' + \gamma \boldsymbol{P}(\mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_{t-1} - \mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_\lambda^*)\right],
\end{aligned}
$$

where we use $\widetilde{\boldsymbol{Z}}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\mathcal{L}_\lambda \boldsymbol{Q}_\lambda^*$ is the regularized Bellman noise and $\widetilde{\boldsymbol{Z}}_t' = (\boldsymbol{P}_t - \boldsymbol{P})(\mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_{t-1} - \mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_\lambda^*)$ (which is still a martingale difference.)

To analyze $\mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_{t-1} - \mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_\lambda^*$, we introduce an intermediate linear operator $\mathcal{L}_\lambda^\pi$, which is defined by

$$
(\mathcal{L}_\lambda^\pi \boldsymbol{Q})(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q(s,a) - \lambda \log \pi(a|s)\right],
$$

for a given policy π and regularization coefficient λ. As a result of notation, $(\mathcal{L}_\lambda \boldsymbol{Q})(\cdot) = \sup_{\pi\in\Pi}(\mathcal{L}_\lambda^\pi \boldsymbol{Q})(\cdot)$ for all $\boldsymbol{Q} \in \mathbb{R}^D$. We assume $\mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_t = \mathcal{L}_\lambda^{\widetilde{\pi}_t} \widetilde{\boldsymbol{Q}}_t$ and $\mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_\lambda^* = \mathcal{L}_\lambda^{\pi_\lambda^*} \widetilde{\boldsymbol{Q}}_\lambda^*$. Hence,

$$
\mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_{t-1} - \mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_\lambda^* = \mathcal{L}_\lambda^{\widetilde{\pi}_{t-1}} \widetilde{\boldsymbol{Q}}_{t-1} - \mathcal{L}_\lambda^{\pi_\lambda^*} \widetilde{\boldsymbol{Q}}_\lambda^* = (\mathcal{L}_\lambda^{\widetilde{\pi}_{t-1}} - \mathcal{L}_\lambda^{\pi_\lambda^*})\widetilde{\boldsymbol{Q}}_{t-1} + \boldsymbol{P}^{\pi_\lambda^*}\widetilde{\mathbf{\Delta}}_{t-1}
$$

where the last equation uses $\mathcal{L}_\lambda^{\pi_\lambda^*}\widetilde{\boldsymbol{Q}}_{t-1} - \mathcal{L}_\lambda^{\pi_\lambda^*}\widetilde{\boldsymbol{Q}}_\lambda^* = \boldsymbol{P}^{\pi_\lambda^*}\widetilde{\mathbf{\Delta}}_{t-1}$ by definition. Putting pieces together,

$$
\widetilde{\mathbf{\Delta}}_t = \widetilde{\boldsymbol{A}}_t \mathbf{\Delta}_{t-1} + \eta_t \left[\widetilde{\boldsymbol{Z}}_t + \gamma \widetilde{\boldsymbol{Z}}_t' + \gamma \widetilde{\boldsymbol{Z}}_t''\right]
$$

where $\widetilde{\boldsymbol{A}}_t = \boldsymbol{I} - \eta_t(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi_\lambda^*})$, $\widetilde{\boldsymbol{Z}}_t' = (\boldsymbol{P}_t - \boldsymbol{P})(\mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_{t-1} - \mathcal{L}_\lambda \widetilde{\boldsymbol{Q}}_\lambda^*)$, and $\widetilde{\boldsymbol{Z}}_t'' = \boldsymbol{P}(\mathcal{L}_\lambda^{\widetilde{\pi}_{t-1}} - \mathcal{L}_\lambda^{\pi_\lambda^*})\widetilde{\boldsymbol{Q}}_{t-1}$. Recurring the last equality gives

$$
\widetilde{\mathbf{\Delta}}_t = \prod_{j=1}^t \widetilde{\boldsymbol{A}}_j \widetilde{\mathbf{\Delta}}_0 + \sum_{j=1}^t \prod_{i=j+1}^t \widetilde{\boldsymbol{A}}_i \eta_j \left(\widetilde{\boldsymbol{Z}}_j + \gamma \widetilde{\boldsymbol{Z}}_t' + \gamma \widetilde{\boldsymbol{Z}}_t''\right).
$$

Besides, using the general step size in Assumption 3.3, we can show $\frac{1}{\sqrt{T}}\sum_{t=1}^T \mathbb{E}\|\widetilde{\mathbf{\Delta}}_t\|_\infty^2 \to 0$ (in Theorem I.1).

Satisfied Lipschitz condition. In order to apply the second and third analysis in Section 3.2, we only need to show that $\|\widetilde{\boldsymbol{Z}}_t''\|_\infty \le L\|\widetilde{\mathbf{\Delta}}_{t-1}\|_\infty^2$ for an appropriate $L > 0$. Notice that $\mathcal{L}_\lambda^{\pi_\lambda^*}\widetilde{\boldsymbol{Q}}_{t-1} \le \mathcal{L}_\lambda^{\widetilde{\pi}_{t-1}}\widetilde{\boldsymbol{Q}}_{t-1}$ and $\mathcal{L}_\lambda^{\widetilde{\pi}_{t-1}}\widetilde{\boldsymbol{Q}}_\lambda^* \le \mathcal{L}_\lambda^{\pi_\lambda^*}\widetilde{\boldsymbol{Q}}_\lambda^*$ coordinately. It implies that $\widetilde{\boldsymbol{Z}}_t'' = \boldsymbol{P}(\mathcal{L}_\lambda^{\widetilde{\pi}_{t-1}} - \mathcal{L}_\lambda^{\pi_\lambda^*})\widetilde{\boldsymbol{Q}}_{t-1}$ satisfies

$$
\mathbf{0} \le \widetilde{\boldsymbol{Z}}_t'' \le \boldsymbol{P}\left[(\mathcal{L}_\lambda^{\widetilde{\pi}_{t-1}} - \mathcal{L}_\lambda^{\pi_\lambda^*})\widetilde{\boldsymbol{Q}}_{t-1} - (\mathcal{L}_\lambda^{\widetilde{\pi}_{t-1}} - \mathcal{L}_\lambda^{\pi_\lambda^*})\widetilde{\boldsymbol{Q}}_\lambda^*\right] = (\boldsymbol{P}^{\widetilde{\pi}_{t-1}} - \boldsymbol{P}^{\pi_\lambda^*})\widetilde{\mathbf{\Delta}}_{t-1}.
$$

Hence, $\|\widetilde{\boldsymbol{Z}}_t''\|_\infty \le \|(\boldsymbol{P}^{\widetilde{\pi}_{t-1}} - \boldsymbol{P}^{\pi_\lambda^*})\widetilde{\mathbf{\Delta}}_{t-1}\|_\infty \le \|\boldsymbol{P}^{\widetilde{\pi}_{t-1}} - \boldsymbol{P}^{\pi_\lambda^*}\|_\infty \|\widetilde{\mathbf{\Delta}}_{t-1}\|_\infty \le \|\Pi^{\widetilde{\pi}_{t-1}} - \Pi^{\pi_\lambda^*}\|_\infty \|\widetilde{\mathbf{\Delta}}_{t-1}\|_\infty$. By definition of Π^π, we know that

$$
\|\boldsymbol{P}^{\widetilde{\pi}_{t-1}} - \boldsymbol{P}^{\pi_\lambda^*}\|_\infty \le \sup_{s\in\mathcal{S}} \|\widetilde{\pi}_{t-1}(\cdot|s) - \pi_\lambda^*(\cdot|s)\|_\infty.
$$

On the other hand, $\widetilde{\pi}_{t-1}, \pi_\lambda$ has a closed form in terms of $\widetilde{\boldsymbol{Q}}_{t-1}$ and $\boldsymbol{Q}_\lambda^*$ respectively. Actually, we have that $\widetilde{\pi}_{t-1}(\cdot|s) \propto \exp(\widetilde{Q}_{t-1}(s,\cdot)/\lambda)$ and $\pi_\lambda^*(\cdot|s) \propto \exp(Q_\lambda^*(s,\cdot)/\lambda)$. By the following lemma, we know that $\|\pi_{t-1}(\cdot|s) - \pi_\lambda(\cdot|s)\|_\infty \le \frac{1}{\lambda}\|Q_\lambda^*(s,\cdot) - \widetilde{Q}_{t-1}(s,\cdot)\|_\infty$. As a result, we have $\|\widetilde{\boldsymbol{Z}}_t''\|_\infty \le L\|\widetilde{\mathbf{\Delta}}_{t-1}\|_\infty^2$ with $L = \frac{1}{\lambda}$.

Lemma I.1. *For any vector $\boldsymbol{v} \in \mathbb{R}^d$, let $\mathrm{softmax} : \mathbb{R}^d \to \mathbb{R}^d$ be defined by $\mathrm{softmax}(\boldsymbol{v})(i) = \exp(\boldsymbol{v}(i))/\sum_{j\in[d]}\exp(\boldsymbol{v}(j))$. Then, $\|\mathrm{softmax}(\boldsymbol{v}_1) - \mathrm{softmax}(\boldsymbol{v}_2)\|_\infty \le \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_\infty$.*

Proof of Lemma I.1. For any $\boldsymbol{v}$, it is easy to find that $\mathrm{softmax}(\boldsymbol{v}) = \frac{\partial L(\boldsymbol{v})}{\partial \boldsymbol{v}}$ where $L(\boldsymbol{v}) = \log(\sum_{j\in[d]}\exp(\boldsymbol{v}(j)))$. It is easy to show that $\left\|\frac{\partial^2 L(\boldsymbol{v})}{\partial^2 \boldsymbol{v}}\right\|_\infty \le 1$ for any $\boldsymbol{v}$. Hence, the result follows from Taylor's expansion. □

The rest proof is almost the same as that in Section 3.2. □

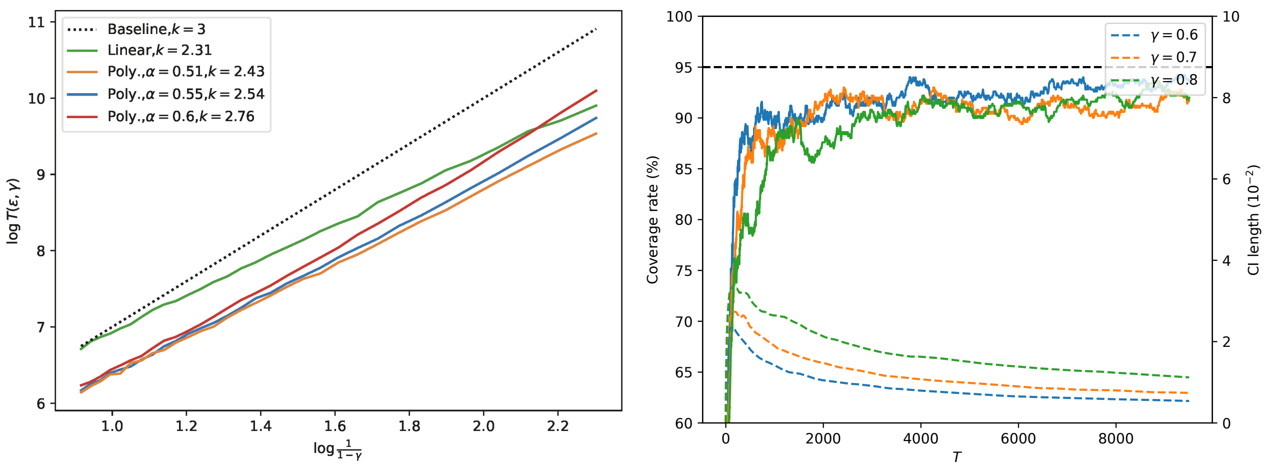

Figure 3: Left: log-log plots of the sample complexity $T(\varepsilon, \gamma)$ versus the discount complexity parameter $(1-\gamma)^{-1}$. Right: the coverage rate and the average length of the 95% confidence interval for regularized Q-Learning.

I.3 Non-asymptotic Bounds in Proof of Theorem 6.1

The error decomposition in Appendix F.1 still apply here. Let $\boldsymbol{\Delta}_t = \widetilde{\boldsymbol{Q}}_t - \boldsymbol{Q}_\lambda^*$ for simplicity. Hence, it follows that

$$\mathbb{E}\|\bar{\boldsymbol{\Delta}}_T - \boldsymbol{Q}_\lambda^*\|_\infty \le \sum_{i=0}^{4} \mathbb{E}\|\widetilde{\mathcal{T}}_i\|_\infty$$

where

$$\widetilde{\mathcal{T}}_0 = \frac{1}{\eta_0 T}(\boldsymbol{A}_0^T - \eta_0 \boldsymbol{I})\widetilde{\boldsymbol{\Delta}}_0, \widetilde{\mathcal{T}}_1 = \frac{1}{T}\sum_{j=1}^{T} \boldsymbol{G}^{-1}\widetilde{\boldsymbol{Z}}_j, \widetilde{\mathcal{T}}_2 = \frac{1}{T}\sum_{j=1}^{T}(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\widetilde{\boldsymbol{Z}}_j$$

$$\widetilde{\mathcal{T}}_3 = \frac{\gamma}{T}\sum_{j=1}^{T} \boldsymbol{A}_j^T(\boldsymbol{P}_j - \boldsymbol{P})(\mathcal{L}_\lambda \boldsymbol{Q}_{j-1} - \mathcal{L}_\lambda \boldsymbol{Q}_\lambda^*), \widetilde{\mathcal{T}}_4 = \frac{\gamma}{T}\sum_{j=1}^{T} \boldsymbol{A}_j^T(\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{j-1}.$$

Pay attention that the $\boldsymbol{A}_j^T$ used above depends on π_λ^* rather than π^* now. As argued in last subsection, Assumption 3.2 is satisfied here with $L = \frac{1}{\lambda}$.

The remaining thing are to repeat what we have done in Appendix F.2, analyzing each term $\widetilde{\mathcal{T}}_i$'s using non-asymptotic concentration inequalities. There are some important aspects to notice. First, for any j, $\|\widetilde{\boldsymbol{Z}}_j\|_\infty \le 2(1+\gamma\|\mathcal{L}_\lambda \boldsymbol{Q}_\lambda^*\|_\infty) \le 2(1+\gamma\|\boldsymbol{Q}_\lambda^*\|_\infty + \gamma\lambda \mathrm{Entropy}(\pi_\lambda^*)) \le \frac{1+\lambda\log\frac{1}{|\mathcal{A}|}}{1-\gamma} = \widetilde{\mathcal{O}}(\frac{1}{1-\gamma})$ where we use $\mathrm{Entropy}(\pi_\lambda^*) \le \log\frac{1}{|\mathcal{A}|}$ and $\|\boldsymbol{Q}^* - \boldsymbol{Q}_\lambda^*\|_\infty \le \frac{\lambda}{1-\gamma}\log\frac{1}{|\mathcal{A}|}$ (which is proved in Theorem 5 of [Yang et al., 2019]). Second, the properties of $\boldsymbol{A}_j^T$'s in Lemma C.3 and C.4 still hold with the same parameters. Finally, we have a counterpart Theorem E.4 due to Theorem 1 in [Wainwright, 2019b] also holds here. The possible difference is that $\|\widetilde{\boldsymbol{Z}}_j\|_\infty$ is bounded $\frac{\lambda}{1-\gamma}\log\frac{1}{|\mathcal{A}|}$ instead of $\frac{1}{1-\gamma}$, which is equivalent up to log factors. Hence, up to log factors, Theorem E.4 also holds for entropy regularized Q-Learning. Putting pieces together, we complete the proof.

J DETAILS OF EXPERIMENTS

The setup of MDP. According to Theorem 5.1, for sufficiently small error $\varepsilon > 0$, we expect the sample complexity $T(\varepsilon, \gamma)$ is always upper bounded by $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty$ and $\frac{1}{(1-\gamma)^3}$ at a worst case. To ensure Assumption 3.2, we consider a random MDP. In particular, for each (s, a) pair, the random reward $R(s, a) \sim \mathcal{U}(0, 1)$ is the uniformly sampled from $(0, 1)$ and the transition probability $P(s'|s, a) = u(s')/\sum_s u(s)$, where $u(s) \overset{i.i.d.}{\sim} \mathcal{U}(0, 1)$. The size of the MDP we choose is $|\mathcal{S}| = 4$, $|\mathcal{A}| = 3$. We consider 30 different values of γ equispaced between 0.6 and 0.9. For a given γ, we run Q-learning algorithm for 10^5 steps (which already ensures convergence) and repeat the process independently for 10^3 times. Finally, we average the ℓ_∞ error $\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty$ of the 10^3 independent trials as an approximation of $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty$ and compute $T(\varepsilon, \gamma)$ by definition. The polynomial step size $\eta_t = t^{-\alpha}$ uses $\alpha \in \{0.51, 0.55, 0.60\}$ and the resacled linear step size is $\eta_t = (1 + (1-\gamma)t)^{-1}$. In Figure 2, we choose $\varepsilon = e^{-4}$ and plot the results on a log-log scale. We then plot the least-squares fits through these points and the slopes of these lines are also provided in the legend.

Confirming the theoretical predictions. In the body, we show the least-squares fits through the points $\{(\log \|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty, \log T(\varepsilon, \gamma))\}_{\gamma \in \Gamma}$. As a complementary, we also show the fits through $\{(\log(1-\gamma)^{-1}, \log T(\varepsilon, \gamma))\}_{\gamma \in \Gamma}$ in Figure 3.

Online inference experiments. We visualize the empirical coverage rate and confidence interval lengths of averaged Q-Learning in Figure 1. We use the random scaling method (Algorithm 1 in [Lee et al., 2021]) to compute the weighting matrix $\boldsymbol{W}_T \in \mathbb{R}^{D \times D}$ where $\boldsymbol{W}_T = \int_0^1 \bar{\boldsymbol{\phi}}_T(r) \bar{\boldsymbol{\phi}}_T(r)^\top \mathrm{d}r$ and $\bar{\boldsymbol{\phi}}_T(r) = \boldsymbol{\phi}_T(r) - r \cdot \boldsymbol{\phi}_T(1)$. We focus on the inference of the optimal value function on the first state s_0 and the first action a_0, i.e., $Q^*(s_0, a_0)$. We use 10^4 steps of value iteration to compute the optimal value function $\boldsymbol{Q}^*$. From [Lee et al., 2021, Li et al., 2022], the asymptotic confidence interval is given by

$$\left[\bar{Q}_T(s_0, a_0) - 6.753 \sqrt{\frac{\boldsymbol{W}_T((s_0, a_0), (s_0, a_0))}{T}}, \bar{Q}_T(s_0, a_0) + 6.753 \sqrt{\frac{\boldsymbol{W}_T((s_0, a_0), (s_0, a_0))}{T}} \right].$$

We set $T = 10^4$ and discard the first 5% samples as a warm-up. This warm-up is quite important; otherwise $\boldsymbol{W}_T$ would change rapidly (as a result of fast convergence of $\boldsymbol{Q}_T$) and deteriorate the performance. The performance is measured by two statistics: the coverage rate and the average length of the 95% confidence interval. We also provide similar results for regularized Q-Learning in Figure 3.