Polyak-Ruppert-Averaged Q-Learning is Statistically Efficient

Xiang Li* Wenhao Yang† Jiadong Liang‡ Zhihua Zhang§ Michael I. Jordan¶

February 8, 2022

Abstract

We study synchronous Q-learning with Polyak-Ruppert averaging (a.k.a., averaged Q-learning) in a γ-discounted MDP. We establish a functional central limit theorem (FCLT) for the averaged iteration $\bar{\boldsymbol{Q}}_T$ and show its standardized partial-sum process converges weakly to a rescaled Brownian motion. Furthermore, we show that $\bar{\boldsymbol{Q}}_T$ is actually a regular asymptotically linear (RAL) estimator for the optimal Q-value function $\boldsymbol{Q}^*$ with the most efficient influence function. This implies the averaged Q-learning iteration has the smallest asymptotic variance among all RAL estimators. In addition, we present a nonasymptotic analysis for the ℓ_∞ error $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty$, showing that for polynomial step sizes it matches the instance-dependent lower bound as well as the optimal minimax complexity lower bound. In short, our theoretical analysis shows that averaged Q-learning is statistically efficient.

1 Introduction

Q-learning [Watkins, 1989], as a model-free approach seeking the optimal Q-function of an Markov decision process (MDP), is perhaps the most popular learning algorithm in reinforcement learning (RL) [Sutton and Barto, 2018]. Study of its sample efficiency has included both asymptotic analysis [Jaakkola et al., 1993, Tsitsiklis, 1994, Borkar and Meyn, 2000] and nonasymptotic analysis [Szepesvári et al., 1998, Even-Dar et al., 2003, Beck and Srikant, 2012, Zhang et al., 2021, Chen et al., 2020b]. In recent years results have tightened and it is now known that the sample efficiency of Q-learning is on the order of $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}\times\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right)$ tight up to a log factor [Li et al., 2021a, 2020b].

The minimax lower bound is $\Omega\left(\frac{|\mathcal{S}\times\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$ [Azar et al., 2013], and thus one of the remaining tasks for the theory of Q-learning is to close the gap on the effective horizon $(1-\gamma)^{-1}$. The gap can be closed via other, more complex, algorithms [Lattimore and Hutter, 2014, Sidford et al., 2018a,b, Wainwright, 2019c], but ideally we would close the gap via standard Q-learning, or a simple variant, using standard tools. Indeed, many RL algorithms can be viewed through the lens of stochastic approximation (SA), a general iterative framework for solving root-finding problems [Robbins and

*School of Mathematical Sciences, Peking University; email: lx10077@pku.edu.cn.
†Academy for Advanced Interdisciplinary Studies, Peking University; email: yangwenhaosms@pku.edu.cn.
‡School of Mathematical Sciences, Peking University; email: jdliang@pku.edu.cn.
§School of Mathematical Sciences, Peking University; email: zhzhang@math.pku.edu.cn.
¶Department of Statistics, Department of Electrical Engineering and Computer Sciences, UC Berkeley; email: jordan@cs.berkeley.edu.

Monro, 1951]. Q-learning 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 (3) for definition). A general way to stabilize and accelerate SA algorithms is via averaging, specifically Polyak-Ruppert averaging [Polyak and Juditsky, 1992], which is known to accelerate policy evaluation [Mou et al., 2020a,b] and exhibits superior empirical performance [Lillicrap et al., 2016, Anschel et al., 2017]. Hence, it is natural to ask whether Q-learning with Polyak-Ruppert averaging (a.k.a., averaged Q-learning) could close the gap with respect to the effective horizon $(1-\gamma)^{-1}$.

We will give an affirmative answer to the question. Furthermore, we will give both asymptotic and nonasymptotic analysis of averaged Q-learning in the setting of a γ-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]. Unlike policy evaluation where the underlying structure is linear in nature and the goal is essentially to solve a linear system, Q-learning is inherently nonlinear, nonsmooth and nonstationary.[1] In this more challenging problem, we can not use classical SA theory directly.[2] Therefore, we develop a new analysis for averaged Q-learning that establishes its asymptotic statistical properties and provides a nonasymptotic analysis for finite samples.

1.1 Our Contribution

We develop a "sandwich" argument to decompose the error $\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^* := \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Q}_t - \boldsymbol{Q}^*$ into several terms, 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. This analysis method may be of independent interest.

Our theoretical findings are summarized as follows. On the asymptotic side, we show that averaged Q-learning enjoys asymptotic normality, $\sqrt{T}(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) \xrightarrow{d} \mathcal{N}(0, \mathsf{Var}_{\boldsymbol{Q}})$, with $\mathsf{Var}_{\boldsymbol{Q}}$ (see (8)) the asymptotic variance. Our results accommodate a polynomial step size as well as a more generally decaying step size (see Assumption 3.3). Furthermore, we establish a functional central limit theorem (FCLT) showing that the standardized 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, $\mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D(r)$, where $\lfloor \cdot \rfloor$ is the floor function and $\boldsymbol{B}_D(\cdot)$ is the standard D-dimensional Brownian motion on $[0,1]$. Such a FCLT 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 $\boldsymbol{Q}^*$. Such a FCLT may open a door to statistical inference for RL.

As a supplementary result, we establish a semiparametric efficiency lower bound for any regular asymptotically linear (RAL) estimator (see Definition 3.1 for detail) of the optimal Q-value function $\boldsymbol{Q}^*$. We further show that $\bar{\boldsymbol{Q}}_T$ is the most efficient RAL estimator with the smallest asymptotic variance, confirming its optimality in the asymptotic regime.

[1] Here the nonstationarity means the maintained policy π_t changes with iteration t. By contrast, $\pi_t \equiv \pi_b$ for all iteration t in policy evaluation where π_b is the target policy.

[2] On the asymptotic side, Polyak and Juditsky [1992] provided a set of conditions under which a given SA algorithm, when combined with Polyak-Ruppert averaging, is guaranteed to have optimal asymptotics. For Q-learning, it is unclear whether we can find a well-behaved Lyapunov function that satisfies those conditions. On the nonasymptotic side, Moulines and Bach [2011] provided a finite-sample analysis for SGD with Polyak-Ruppert averaging on smooth loss functions, which is not suitable for Q-learning.

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 $\mathcal{O}(\sqrt{\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln|\mathcal{S}\times\mathcal{A}|}{T}})$ 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 $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty = \mathcal{O}((1-\gamma)^{-3})$, imply 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].

1.2 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 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 in 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 in 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.

2 Preliminaries

Discounted infinite-horizon MDPs. We consider an infinite-horizon MDP as represented by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, R, r)$. Here $\mathcal{S}$ is the state space with $S = |\mathcal{S}|$ the cardinality and $\mathcal{A}$ is the action space with $A = |\mathcal{A}|$ the cardinality. $\gamma \in (0,1)$ is the discount factor. $P\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ represents the probability transition kernel, i.e., $P(s'|s,a)$ is the probability of transiting to s' from a given state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. $R\colon \mathcal{S} \times \mathcal{A} \to [0,\infty)$ stands 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.

Value function and Q-function. A deterministic policy π maps each $s \in \mathcal{S}$ to a single action $a \in \mathcal{A}$. In a γ-discounted MDP, a common objective is to maximize the expected long-term reward. For a given deterministic policy $\pi\colon \mathcal{S} \to \mathcal{A}$, the expected long-term reward is measured by a value function V^π and a Q-function Q^π defined as follows:

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

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here $\tau = \{(s_t,a_t)\}_{t\geq 0}$ is a trajectory of the MDP induced by the policy π and the expectation $\mathbb{E}_{\tau\sim\pi}(\cdot)$ is taken with respect to the randomness of the trajectory τ. The optimal value function V^* and optimal Q-function Q^* are defined as $V^*(s) = \max_\pi V^\pi(s)$ and $Q^*(s,a) = \max_\pi Q^\pi(s,a)$, respectively, for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Q-learning. We assume access to a generative model [Kearns and Singh, 1999, Sidford et al., 2018a, Li et al., 2021a]. In particular, in each 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 $(s,a) \in \mathcal{S} \times \mathcal{A}$.[3] Note that $r_t(s,a)$ is identically distributed as $R(s,a)$ with the expectation $r(s,a)$. The synchronous Q-learning algorithm maintains a Q-function estimate, $Q_t \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, for all $t \geq 0$ and updates all entries of the Q-function estimate via the following update rule:

$$Q_t = (1-\eta_t) Q_{t-1} + \eta_t \widehat{\mathcal{T}}_t(Q_{t-1}), \tag{1}$$

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

$$\widehat{\mathcal{T}}_t(Q)(s,a) = r_t(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a'), \text{ where } r_t(s,a) \sim R(s,a) \text{ and } s_t = s_t(s,a) \sim P(\cdot|s,a), \tag{2}$$

for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Clearly, $\widehat{\mathcal{T}}_t$ is an unbiased estimate of the celebrated Bellman operator $\mathcal{T}$ given by

$$\mathcal{T}(Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q(s', a') \text{ for all } (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{3}$$

The optimal Q-function Q^* is the unique fixed point of the Bellman operator, $\mathcal{T}(Q^*) = Q^*$. 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] show 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 more difficult than policy evaluation because of 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 (1). 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].

Matrix notation. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{D \times D}$, we use $\|\boldsymbol{A}\|_\infty$ to denote the infinity operator norm of $\boldsymbol{A}$, i.e., $\|\boldsymbol{A}\|_\infty = \max_{\|\boldsymbol{x}\|_\infty = 1} \|\boldsymbol{A}\boldsymbol{x}\|_\infty = \max_{i \in [D]} \sum_{j \in [D]} |\boldsymbol{A}(i,j)|$. We use $\|\boldsymbol{A}\|_{\max}$ to denote the max norm, i.e., $\|\boldsymbol{A}\|_{\max} = \max_{i,j \in [D]} |\boldsymbol{A}(i,j)|$. We use $\mathrm{diag}(\boldsymbol{A})$ to denote the diagonal matrix obtained by removing all off-diagonal entries in $\boldsymbol{A}$.

For simplicity, we define $D = |\mathcal{S} \times \mathcal{A}| = SA$. 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

$$\boldsymbol{P}^\pi := \boldsymbol{P}\boldsymbol{\Pi}^\pi \quad \text{and} \quad \boldsymbol{P}_\pi := \boldsymbol{\Pi}^\pi \boldsymbol{P}, \tag{4}$$

where $\boldsymbol{\Pi}^\pi \in \{0,1\}^{S \times D}$ is a projection matrix associated with the deterministic policy π:

$$\boldsymbol{\Pi}^\pi = \mathrm{diag}\{\boldsymbol{e}_{\pi(1)}^\top, \boldsymbol{e}_{\pi(2)}^\top, \cdots, \boldsymbol{e}_{\pi(S)}^\top\}, \tag{5}$$

[3] This means that $s_t(s,a)$ (as well as $r_t(s,a)$) are independent over different $(s,a) \in \mathcal{S} \times \mathcal{A}$.

with $\boldsymbol{e}_i$ the i-th standard basis vector. We use $\boldsymbol{r}_t \in \mathbb{R}^D$ to represent the random reward R generated at iteration t such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, the (s,a)-th entry of $\boldsymbol{r}_t$ is given by $\boldsymbol{r}_t(s,a) \overset{d.}{=} R(s,a)$. The vector $\boldsymbol{r} = \mathbb{E}\boldsymbol{r}_t \in \mathbb{R}^D$ denotes the expected reward. Similarly $\boldsymbol{P}_t \in \mathbb{R}^{D\times S}$ is the empirical transition matrix at iteration t with each row containing only one nonzero entry. We have $\mathbb{E}\boldsymbol{P}_t = \boldsymbol{P}$. Analogously, 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 $V^\pi, V^*, Q^\pi, Q^*, Q_t, \bar{Q}_t$.

Bellman noise. We define $\boldsymbol{Z}_t \in \mathbb{R}^D$ as the Bellman noise (vector) at the t-th iteration. The (s,a)-th entry of $\boldsymbol{Z}_t$ 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 (2) using samples collected at iteration t as an estimate of the population Bellman operator (3).

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 $\boldsymbol{Z}$:

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

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

3 Asymptotic Properties of Averaged Q-learning

In this section, we present our investigation of the asymptotic properties of averaged Q-learning. We make three mild assumptions throughout the section. The first is that rewards are uniformly bounded and nonnegative (Assumption 3.1). The second is that the optimal policy is unique. This implies a positive optimality gap (defined in Assumption 3.2) and ensures the optimal variance is unique and identifiable. Both conditions are standard in the RL literature. The last one (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, Li et al., 2021c]. 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 (Uniformly bounded random reward)**.** *The random reward is nonnegative and uniformly bounded, i.e., for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, $0 \le R(s,a) \le 1$ almost surely.*

Assumption 3.2 (Unique optimal policy)**.** *The optimal policy is unique, which we denote by π^*. It implies a positive optimality gap defined by* $\mathrm{gap} := \min_{s\in\mathcal{S}} \min_{a\neq\pi^*(s)} |V^*(s) - Q^*(s,a)| > 0$.

Assumption 3.3. *Assume $\{\eta_t\}$ satisfies* (i) $0 \le \sup_t \eta_t \le 1, \eta_t \downarrow 0$ *and* $t\eta_t \uparrow \infty$ *as* $t \to \infty$*;* (ii) $\frac{\eta_{t-1}-\eta_t}{\eta_{t-1}} = o(\eta_{t-1})$ *and* $(1-\eta_t)\eta_{t-1} \le \eta_t$ *for* $t \ge 1$*; and* (iii) $\frac{1}{\sqrt{T}}\sum_{t=0}^{T}\eta_t \to 0$ *as* $T \to \infty$.

3.1 Central Limit Theorem (CLT)

We present a CLT for the averaged Q-learning sequence $\bar{\boldsymbol{Q}}_T := \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Q}_t$.

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

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

where the asymptotic variance is given by

$$\mathsf{Var}_{\boldsymbol{Q}} = (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\mathrm{Var}(\boldsymbol{Z})(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-\top} \in \mathbb{R}^{D\times D}. \tag{8}$$

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

Asymptotic variance. Theorem 3.1 implies that the average of sequence $\boldsymbol{Q}_t$ has an asymptotic normal distribution with $\mathsf{Var}_{\boldsymbol{Q}}$ the asymptotic variance. $\mathsf{Var}_{\boldsymbol{Q}}$ 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 (2) as an estimate of the population Bellman operator (3). Note that this implies $\|(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\| \le \sum_{t=0}^{\infty}\gamma^t \|(\boldsymbol{P}^{\pi^*})^t\|_\infty = (1-\gamma)^{-1}$. $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty$ 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. 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 3.1, $\bar{\boldsymbol{V}}_T$ enjoys a similar asymptotic normality with the asymptotic variance defined by $\mathsf{Var}_{\boldsymbol{V}}$. One can check that

$$\mathsf{Var}_{\boldsymbol{V}} = \boldsymbol{\Pi}^{\pi^*}\mathsf{Var}_{\boldsymbol{Q}}(\boldsymbol{\Pi}^{\pi^*})^\top, \tag{9}$$

where $\boldsymbol{\Pi}^{\pi^*} \in \{0,1\}^{S\times D}$ is the projection matrix associated with the deterministic optimal policy π^* (see the definition in (5)). Hence, $\mathsf{Var}_{\boldsymbol{V}}$ is formed by selecting entries from $\mathsf{Var}_{\boldsymbol{Q}}$. In particular, $\mathsf{Var}_{\boldsymbol{V}}(s,s') = \mathsf{Var}_{\boldsymbol{Q}}((s,\pi^*(s)),(s',\pi^*(s')))$ for any $s, s' \in \mathcal{S}$. The proof is deferred to Appendix C.2.

Corollary 3.1 (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.,* $\bar{V}_T(s) \in \arg\max_{a\in\mathcal{A}} \bar{Q}_T(s,a)$*. Then under Assumptions 3.1, 3.2 and 3.3,*

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

where the asymptotic variance is

$$\mathsf{Var}_{\boldsymbol{V}} = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi^*})^{-1}\mathrm{Var}(\boldsymbol{\Pi}^{\pi^*}\boldsymbol{Z})(\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi^*})^{-\top} \in \mathbb{R}^{S\times S}, \tag{10}$$

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. Noticing that $\bar{\boldsymbol{Q}}_T$ is uniformly bounded from Assumption 3.1 (i.e., $\|\bar{\boldsymbol{Q}}_T\|_\infty \leq \frac{1}{1-\gamma}$ for any $T \geq 0$), the bounded convergence theorem yields that as T goes to infinity,

$$\sqrt{T}\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty \to \mathbb{E}\|\mathcal{Z}\|_\infty \approx \sqrt{\ln D}\sqrt{\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty} \text{ where } \mathcal{Z} \sim \mathcal{N}(\mathbf{0}, \mathsf{Var}_{\boldsymbol{Q}}). \tag{11}$$

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

3.2 Functional Central Limit Theorem (FCLT)

We can further establish a functional version of the 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:

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

For simplicity, we use $\boldsymbol{\phi}_T := \{\boldsymbol{\phi}_T(r)\}_{r\in[0,1]}$ or $\boldsymbol{\phi}_T(\cdot)$ to denote the whole function.

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

$$\boldsymbol{\phi}_T(\cdot) \xrightarrow{w} \mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D(\cdot), \tag{12}$$

where $\mathsf{Var}_{\boldsymbol{Q}}$ *is defined in* (9) *and* $\boldsymbol{B}_D(\cdot)$ *is the standard* D*-dimensional Brownian motion on* $[0,1]$.

The CLT in Theorem 3.1 asserts that $\boldsymbol{\phi}_T(1)$ converges in distribution to a rescaled Gaussian random variable $\mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D(1)$ as $T \to \infty$. The FCLT in Theorem 3.2 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 $\boldsymbol{\phi}_T$ converges in distribution. That is, for any given integer $n \geq 1$ and any $0 \leq t_1 < t_2 < \cdots < t_n \leq 1$,

$$(\boldsymbol{\phi}_T(t_1), \boldsymbol{\phi}_T(t_2), \cdots, \boldsymbol{\phi}_T(t_n)) \xrightarrow{d} \mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}(\boldsymbol{B}_D(t_1), \boldsymbol{B}_D(t_2), \cdots, \boldsymbol{B}_D(t_n)) \text{ as } T \text{ goes to infinity.} \tag{13}$$

The convergence $\xrightarrow{w}$ in (12) also corresponds to the weak convergence of measures in the D-dimensional Skorokhod spaces $\mathsf{D}([0,1], \mathbb{R}^D)$ (see Appendix B.1.1 for a short introduction). Here $\mathsf{D}([0,1], \mathbb{R}^D) = \{\text{right continuous with left limits } \omega(r) \in \mathbb{R}^D, r \in [0,1]\}$. Eq. (12) is equivalent to the convergence of finite-dimensional projections in (13). One can prove Theorem 3.1 by applying the continuous mapping theorem to Theorem 3.2 with the functional $f : \mathsf{D}([0,1], \mathbb{R}^D) \to \mathbb{R}^D, f(w) = w(1)$. We provide a full proof of Theorem 3.2 in Appendix B.

Insights on statistical inference. The FCLT 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 FCLT allows us to construct an asymptotically pivotal statistic using the whole function $\boldsymbol{\phi}_T$.

Proposition 3.1. *The continuous mapping theorem together with Theorem 3.2 yields that with probability approaching to one, $\int_0^1 \boldsymbol{\phi}_T(r)\boldsymbol{\phi}_T(r)^\top dr$ is invertible. It additionally follows that*

$$\boldsymbol{\phi}_T(1)^\top \left(\int_0^1 \boldsymbol{\phi}_T(r)\boldsymbol{\phi}_T(r)^\top dr\right)^{-1} \boldsymbol{\phi}_T(1) \xrightarrow{d} \boldsymbol{B}_D(1)^\top \left(\int_0^1 \boldsymbol{B}_D(r)\boldsymbol{B}_D(r)^\top dr\right)^{-1} \boldsymbol{B}_D(1). \tag{14}$$

The left-hand side of (14) is a pivotal quantity involving samples and the unobservable parameter of interest $\boldsymbol{Q}^*$. The pivotal quantity can be constructed in a fully online fashion and thus is computational efficient [Lee et al., 2021, Li et al., 2021c]. The right-hand side of (14) 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 $\boldsymbol{Q}^*$. The construction of an asymptotic pivotal quantity is not unique. Proposition 3.1 provides a particular example. It is of interest to identify other possible constructions and to compare their advantages in an appropriate sense, which we leave for future work.

3.3 Information Theoretical Lower Bound

Theorem 3.1 shows $\bar{\boldsymbol{Q}}_T$ is a $\sqrt{T}$-consistent estimate for $\boldsymbol{Q}^*$ with asymptotic variance $\mathsf{Var}_{\boldsymbol{Q}}$. 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 (CRLB) 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 CRLB is viewed as optimal and efficient. The CRLB concept 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 $\Delta(\mathcal{S})$,[4] 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 CRLB 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 $\theta = (P, R)$. The parameters P and R are variationally independent. Our parameter of interest is $\beta(\theta) = \boldsymbol{Q}^*$. At iteration t, $\boldsymbol{r}_t \in \mathbb{R}^D$ collects all random rewards generated at each (s,a) and $\boldsymbol{P}_t \in \mathbb{R}^{D\times S}$ gathers all empirical transitions following the probability $\boldsymbol{P}_{s,a} := P(\cdot|s,a)$ starting from each (s,a). Specifically, the (s,a)-th entry of $\boldsymbol{r}_t$ is an independent copy of $R(s,a)$, while the (s,a)-th row of $\boldsymbol{P}_t$ is a one-hot random vector with a single nonzero entry. The distribution of $\boldsymbol{P}_t$ is determined by its expectation $\boldsymbol{P} = \mathbb{E}\boldsymbol{P}_t$ which belongs to

$$\mathcal{P}_P := \left\{ \boldsymbol{P} \in \mathbb{R}^{D\times S} : P(s'|s,a) \geq 0, \forall (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S} \text{ and } \sum_{s'\in\mathcal{S}} P(s'|s,a) = 1 \right\}, \tag{15}$$

while R is nonparametric belonging to

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

[4] To determine a distribution with a finite support (say, S discrete points), we only need to specify $S-1$ parameters $\{p_s\}_{s\in[S-1]}$ which satisfies $0 \leq p_s \leq 1$ for all $s \in [S-1]$ and $0 \leq \sum_{s\in[S-1]} p_s \leq 1$.

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 estimators provide a good trade off between expressivity and tractability.

Definition 3.1 (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*

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

where $\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t) = \boldsymbol{0}$ and $\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)^\top$ is finite and nonsingular. Here $\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)$ is the influence function.

Remark 3.1. *Informally speaking, an estimator is called 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 whcih perform poorly in the neighborhood of points of super-efficiency. We suggest interested readers refer to Section 3.1 in Tsiatis [2006] for a detailed introduction. There exists a straightforward criterion on influence functions to check regularity of an asymptotically linear estimator (see Theorem 2.2 in Newey [1990]; we employ this criterion in our proof of Theorem 3.4).*

Theorem 3.3. *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*

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

where $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite and $\mathsf{Var}_{\boldsymbol{Q}}$ is given in (8).

By Definition 3.1, any influence function determines an asymptotic linear estimator for $\boldsymbol{Q}^*$. The semiparametric efficiency bound in Theorem 3.3 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 3.4 implies that $\bar{\boldsymbol{Q}}_T$ is the most efficient estimator among all RAL estimators with the efficient influence function $(\boldsymbol{I} - \gamma\boldsymbol{P}^*)^{-1}\boldsymbol{Z}_t$. Theorem 3.4 is stronger than Theorem 3.1 because it not only implies asymptotic normality, but also shows the regularity of $\bar{\boldsymbol{Q}}_T$. Proofs are provided in Appendix E.

Theorem 3.4. *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*

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

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

4 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 $(1-\gamma)^{-1}$. We first provide a nonasymptotic convergence result to validate (11); the proof is in Appendix D.

Theorem 4.1. *Under Assumptions 3.1 and 3.2, when D is larger than a universal constant,*

- *If $\eta_t = t^{-\alpha}$ with $\alpha \in (0.5, 1)$ for $t \geq 1$ and $\eta_0 = 1$, it follows that for all $T \geq 1$,*

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

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

$$\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty = \mathcal{O}\left(\sqrt{\frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}}\sqrt{\frac{\ln D}{T}}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^6}\frac{1}{T}\right).$$

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

Instance-dependent behavior. To the best of our knowledge, Theorem 4.1 is the first finite-sample analysis of averaged Q-learning in the ℓ_∞-norm. For the polynomial step size, this shows that the instance-dependent term $\mathcal{O}(\sqrt{\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}})$ dominates the ℓ_∞ error, which matches the instance-dependent lower bound established in Khamaru et al. [2021b] given a sufficiently large T. However, for the linearly 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

$$\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty \overset{(a)}{\leq} \|(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\|_\infty^2 \|\mathrm{Var}(\boldsymbol{Z})\|_\infty \overset{(b)}{\leq} \frac{1}{(1-\gamma)^2}\|\mathrm{Var}(\boldsymbol{Z})\|_\infty, \tag{16}$$

where (a) uses $\|\mathrm{diag}(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^\top)\|_\infty \leq \|\boldsymbol{V}\|_\infty \|\boldsymbol{A}\|_\infty^2$ for any diagonal matrix $\boldsymbol{V}$ (see Lemma D.2) and (b) uses $\|(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\|_\infty \leq (1-\gamma)^{-1}$. Therefore, the linearly rescaled step size doesn't match the instance-dependent lower bound. This makes sense because the linearly rescaled step size doesn't satisfy Assumption 3.3, implying that (11) does not necessarily hold for it.

Khamaru et al. [2021b] analyze a variance-reduced variant of Q-learning that achieves instance-dependent optimality with the following guarantee:

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

which has a better nonleading term than averaged Q-learning. This might somewhat explain the finding in Khamaru et al. [2021a] that averaging might be sub-optimal in the nonasymptotic regime

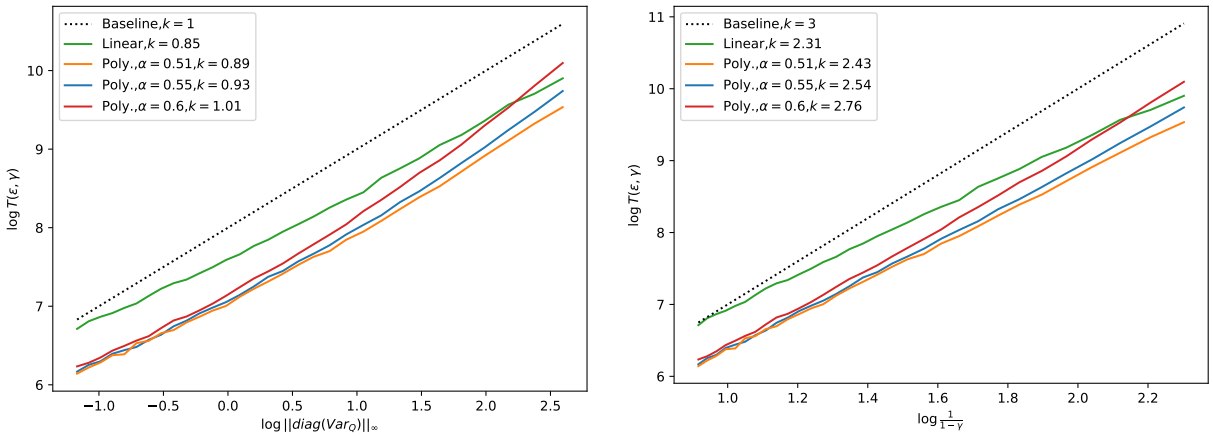

Figure 1: Log-log plots of the sample complexity $T(\varepsilon, \gamma)$ versus the asymptotic variance $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty$ (left) and versus the discount complexity parameter $(1-\gamma)^{-1}$ (right).

with limited samples. However, the dominant terms are equal, implying that averaging is still powerful and efficient in the asymptotic regime. Additionally, instance-dependent convergence with a variance structure in the dominant term has also been found for optimal value estimation [Yin and Wang, 2021], and for policy evaluation in the tabular setting [Pananjady and Wainwright, 2020] or using linear function approximation [Li et al., 2021b], with the specific functional being different in different cases.

Worst-case behavior. Previous works [Azar et al., 2013, Li et al., 2020a] imply the worst-case bound $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{V}})\|_\infty \le \|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty = \mathcal{O}((1-\gamma)^{-3})$.[5] Such a dependence on $(1-\gamma)^{-1}$ is tight, because Khamaru et al. [2021b] construct a a family of MDPs parameterized by $\lambda \ge 0$ where $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty = \Theta((1-\gamma)^{-3+\lambda})$. 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 in 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 (3). 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.

Confirming the theoretical predictions. We provide numerical experiments to illustrate instance-adaptivity as well as the worst-case behavior delineated in Theorem 4.1. We focus on the 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 $T(\varepsilon, \gamma)$ under different values of $\gamma \in \Gamma$ and two step sizes. We then plot the least-squares fits through these points $\{(\log(1-\gamma)^{-1}, \log T(\varepsilon, \gamma))\}_{\gamma\in\Gamma}$ and $\{(\log \|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty, \log T(\varepsilon, \gamma))\}_{\gamma\in\Gamma}$ and provide the slopes k of these lines in the legend. Further

[5] For example, Lemma 8 in Li et al. [2020a] implies that $\|(\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi^*})^{-1}\sqrt{\mathrm{Var}(\boldsymbol{\Pi}^{\pi^*}\boldsymbol{Z})}\|_\infty = \mathcal{O}((1-\gamma)^{-1.5})$. Using the inequality $\|\mathrm{diag}(\boldsymbol{A}\boldsymbol{A}^\top)\|_\infty \le \|\boldsymbol{A}\|_\infty^2$ (see Lemma D.2), we have $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{V}})\|_\infty = \mathcal{O}((1-\gamma)^{-3})$. Lemma 7 in Azar et al. [2013] implies $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty = \mathcal{O}((1-\gamma)^{-3})$. Finally, the relation between $\mathsf{Var}_{\boldsymbol{V}}$ and $\mathsf{Var}_{\boldsymbol{Q}}$ shown in (9) yields $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{V}})\|_\infty \le \|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty$.

details are provided in Appendix G. At a high level, we see that the averaged Q-learning with different step sizes produces sample complexity that is well predicted by our theory: all the slopes are no larger than the theoretical limit k^*.

5 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].

There are many directions to extend our work. For one thing, it is of interest to explore other ways to construct an asymptotically pivotal statistic and compare their advantages theoretically and empirically. In another vein, it is possible to extend our nonasymptotic analysis method to a more general problem, nonlinear stochastic approximation. There are also some open problems. First, it is unclear whether averaged Q-learning with linearly rescaled step sizes can match the instance-dependent lower bound such that $\mathcal{O}\left(\sqrt{\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}}\right)$ would dominate the stochastic error rather than current $\mathcal{O}\left(\sqrt{\frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}}\sqrt{\frac{\ln D}{T}}\right)$. Second, it is also unclear whether the dependence on $(1-\gamma)^{-1}$ of the nonleading terms in Theorem 4.1 can be improved or not.

References

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

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 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.

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.

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

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

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. *Journal of Machine Learning Research*, 21:167–1, 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.

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

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. *arXiv preprint arXiv:2005.12900*, 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 inference via local SGD in federated learning. *arXiv preprint arXiv:2109.01326*, 2021c.

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 *ICLR (Poster)*, 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 *Proceedings of the 25th International Conference on Machine Learning*, pages 664–671, 2008.

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.

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*, volume 3. Cambridge University Press, 2000.

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

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. 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, 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.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, pages 12653–12662. PMLR, 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.

Appendix

A 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 $\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2$. 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.

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

$$\lim_{T\to\infty}\frac{1}{\sqrt{T}}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 = 0. \tag{17}$$

Theorem A.2 provides the specific rates of $\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2$ for two choices of step size. Both step sizes have been analyzed by Wainwright [2019b]. To prove the theorem, we follow the lines of the induction argument in Wainwright [2019b], which is used to provide non-asymptotic convergence rates for $\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty$. In our case, an important immediate result is Theorem A.3 that captures the non-asymptotic convergence rate of $\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2$ for all $0 \le t \le T$. Previous state-of-the-art analysis in Li et al. [2021a], though tight on the dependence of $(1-\gamma)^{-1}$, is more sophisticated and complicated. Moreover, it can neither cover all $0 \le t \le T$, nor provide convergence for polynomial-decaying step sizes. Hence, we still use the technique in Wainwright [2019b] for simplicity and step-size universality, though it is sub-optimal. Hence, it is possible to improve the dependence on $(1-\gamma)^{-1}$ in Theorem A.2 by using tighter bounds for all $\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 (0 \le t \le T)$.

Theorem A.2. *Under Assumption 3.1, 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

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

A.1 A sandwich ℓ_∞ bound

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

Lemma A.1. *For any sequence of step sizes $\{\eta_t\}_{t\geq 0}$ in the interval $(0,1)$, the iterates $\{\boldsymbol{\Delta}_t\}_{t\geq 0}$ satisfies the sandwich relation*

$$-(a_t+b_t)\mathbf{1}+\boldsymbol{N}_t \leq \boldsymbol{\Delta}_t \leq (a_t+b_t)\mathbf{1}+\boldsymbol{N}_t \tag{18}$$

where $\{a_t\}_{t\geq 0}, \{b_t\}_{t\geq 0}$ are non-negative scalars and $\{\boldsymbol{N}_t\}_{t\geq 0}$ 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.

Proof of Lemma A.1. The synchronous Q-learning has the following update rule

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r}_t + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}) \tag{19}$$

in the t-th iteration. We start by decomposing the estimation error $\boldsymbol{\Delta}_t$:

$$\begin{aligned} \boldsymbol{\Delta}_t &= \boldsymbol{Q}_t - \boldsymbol{Q}^* = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r}_t + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}) - \boldsymbol{Q}^* \\ &= (1-\eta_t)(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^*) + \eta_t(\boldsymbol{r}_t + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1} - \boldsymbol{Q}^*) \\ &\overset{(a)}{=} (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\left[(\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^*)\right] \\ &= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\left[(\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^* + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*)\right] \\ &\overset{(b)}{=} (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\left[\boldsymbol{Z}_t + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*)\right], \end{aligned} \tag{20}$$

where (a) uses the equation $\boldsymbol{Q}^* = \boldsymbol{r} + \gamma\boldsymbol{P}\boldsymbol{V}^*$ and (b) uses the shorthand $\boldsymbol{Z}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*$. Further, the term $\boldsymbol{P}_t(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*)$ can be linked with $\boldsymbol{\Delta}_{t-1}$ as following

$$\|\boldsymbol{P}_t(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*)\|_\infty \leq \|\boldsymbol{P}_t\|_\infty\|\boldsymbol{V}_{t-1} - \boldsymbol{V}^*\|_\infty = \|\boldsymbol{V}_{t-1} - \boldsymbol{V}^*\|_\infty \leq \|\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^*\|_\infty. \tag{21}$$

Next we use mathematical induction to prove (18). For iteration $t = 0$, it follows that $-(a_0+b_0)\mathbf{1} + \boldsymbol{N}_0 = -\|\boldsymbol{\Delta}_0\|_\infty\mathbf{1} \leq \boldsymbol{\Delta}_0 \leq \|\boldsymbol{\Delta}_0\|_\infty\mathbf{1} = (a_0+b_0)\mathbf{1} + \boldsymbol{N}_0$ due to the initialization $a_0 = \|\boldsymbol{\Delta}_0\|_\infty, b_0 = 0$ and $\boldsymbol{N}_0 = \mathbf{0}$.

We now assume that the claim holds at iteration $t-1$, and show that it holds for iteration t. For the upper bound, by (20) and (21), we have

$$\begin{aligned} \boldsymbol{\Delta}_t &= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\left[\boldsymbol{Z}_t + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*)\right] \\ &\leq (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\left[\boldsymbol{Z}_t + \gamma\|\boldsymbol{\Delta}_{t-1}\|_\infty\mathbf{1}\right] \\ &\leq (1-\eta_t)\left[(a_{t-1}+b_{t-1})\mathbf{1} + \boldsymbol{N}_{t-1}\right] + \eta_t\left[\boldsymbol{Z}_t + \gamma(a_{t-1}+b_{t-1}+\|\boldsymbol{N}_{t-1}\|_\infty)\mathbf{1}\right] \\ &= (1-\eta_t(1-\gamma))a_{t-1}\mathbf{1} + \left[(1-\eta_t(1-\gamma))b_{t-1} + \eta_t\|\boldsymbol{N}_{t-1}\|_\infty\right]\mathbf{1} + \left((1-\eta_t)\boldsymbol{N}_{t-1} + \eta_t\boldsymbol{Z}_t\right) \\ &= (a_t+b_t)\mathbf{1} + \boldsymbol{N}_t. \end{aligned}$$

Similarly, for the lower bound, we have

$$\boldsymbol{\Delta}_t = (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\left[\boldsymbol{Z}_t + \gamma\boldsymbol{P}_t(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*)\right]$$

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

which completes the proof of the lower bound. □

A.2 Bounding the second moment of a sum of Bellman noise terms

The second step is to bound $\mathbb{E}\|\boldsymbol{N}_T\|_\infty^2$ which is an autoregressive process of independent Bellman noise terms.

Lemma A.2. *Under Assumption 3.1 and assuming $(1-\eta_t)\eta_{t-1} \leq \eta_t$ for any $t \geq 1$, we have*

$$
\mathbb{E}\|\boldsymbol{N}_t\|_\infty^2 \leq \min\left\{\frac{2\eta_t \ln(2eD)}{(1-\gamma)^2}, c\left(\eta_t\|\mathrm{Var}(\boldsymbol{Z})\|_\infty \ln(2eD) + \frac{\eta_t^2 \ln^2(2eD)}{(1-\gamma)^2}\right)\right\}
$$

where $c > 0$ is a universal constant and $\mathrm{Var}(Z) \in \mathbb{R}^{D\times D}$ is the covariance matrix of Bellman noise terms.

Proof of Lemma A.2. Recall that $\{\boldsymbol{N}_t\}$ is recursively defined in Lemma A.1: $\boldsymbol{N}_0 = \mathbf{0}$ and $\boldsymbol{N}_t = (1-\eta_t)\boldsymbol{N}_{t-1} + \eta_t\boldsymbol{Z}_t$ with $\boldsymbol{Z}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*$. Let $\boldsymbol{e}_i \in \mathbb{R}^D$ be the i-th standard basis with only the i-th coordinate non-zero and equal to 1.

- It is clear that $|\boldsymbol{e}_i^\top \boldsymbol{Z}_t| \leq (1-\gamma)^{-1}$ (see Lemma 4 in Li et al. [2021a] for the reason). Hence, by Hoeffding's lemma, we have

$$
\sup_{i\in[D]} \mathbb{E}\exp(\lambda \boldsymbol{e}_i^\top \boldsymbol{Z}_t) \leq \exp\left(\frac{1}{(1-\gamma)^2}\frac{\lambda^2}{2}\right) \text{ for all } \lambda \in \mathbb{R}. \tag{22}
$$

  By induction, we have the following lemma.

  Lemma A.3. *If $(1-\eta_t)\eta_{t-1} \leq \eta_t$ for any $t \geq 1$, we have for any $t \geq 0$,*

$$
\mathbb{E}\exp\left(\lambda\|\boldsymbol{N}_t\|_\infty\right) \leq 2D\exp\left(\frac{1}{(1-\gamma)^2}\frac{\lambda^2\eta_t}{2}\right).
$$

  As a result, the tail distribution bound of $\|\boldsymbol{N}_t\|_\infty$ is

$$
\mathbb{P}(\|\boldsymbol{N}_t\|_\infty \geq \tau) \leq 2D\exp\left(-\frac{(1-\gamma)^2\tau^2}{2\eta_t}\right).
$$

  Let $\tau_0 = \frac{2\eta_T \ln(2D)}{(1-\gamma)^2}$ such that $2D\exp\left(-\frac{(1-\gamma)^2\tau_0^2}{2\eta_t}\right) = 1$. Then,

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{N}_t\|_\infty^2 &= \int_0^\infty \mathbb{P}(\|\boldsymbol{N}_t\|_\infty^2 \geq \tau)d\tau = \int_0^\infty \mathbb{P}(\|\boldsymbol{N}_t\|_\infty \geq \sqrt{\tau})d\tau \\
&= \int_0^{\tau_0} \mathbb{P}(\|\boldsymbol{N}_t\|_\infty \geq \sqrt{\tau})d\tau + \int_{\tau_0}^\infty \mathbb{P}(\|\boldsymbol{N}_t\|_\infty \geq \sqrt{\tau})d\tau \\
&\leq \tau_0 + 2D\int_{\tau_0}^\infty \exp\left(-\frac{(1-\gamma)^2}{2\eta_t}\tau\right)d\tau \\
&= \tau_0 + \frac{4D\eta_t}{(1-\gamma)^2}\exp\left(-\frac{(1-\gamma)^2\tau_0}{2\eta_t}\right) \leq \frac{2\eta_t \ln(2eD)}{(1-\gamma)^2}.
\end{aligned}
$$

- To apply Lemma 2 in Wainwright [2019b], we verify the conditions therein: we have $(1-\eta_t)\eta_{t-1} \le \eta_t$ for any $t \ge 1$, $|\boldsymbol{e}_i^\top \boldsymbol{Z}_t| \le (1-\gamma)^{-1}$ and $\mathbb{E}|\boldsymbol{e}_i^\top \boldsymbol{Z}_t|^2 \le \|\mathrm{Var}(\boldsymbol{Z})\|_\infty$. Hence, its Lemma 2 yields

$$\sup_{i\in[D]} \ln \mathbb{E} \exp\left(\lambda \boldsymbol{e}_i^\top \boldsymbol{N}_t\right) \le \frac{\lambda^2 \eta_t \|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{1-\eta_t(1-\gamma)^{-1}|\lambda|} \text{ for all } |\lambda| \le \frac{1-\gamma}{\eta_t}.$$

As a result, for all $|\lambda| \le \frac{1-\gamma}{\eta_t}$, we have

$$\mathbb{E}\exp\left(\lambda\|\boldsymbol{N}_t\|_\infty\right) \le \sum_{i\in[D]} \left[\mathbb{E}\exp\left(\lambda \boldsymbol{e}_i^\top \boldsymbol{N}_t\right) + \mathbb{E}\exp\left(-\lambda \boldsymbol{e}_i^\top \boldsymbol{N}_t\right)\right] \le 2D\exp\left(\frac{\lambda^2\eta_t\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{1-\eta_t(1-\gamma)^{-1}|\lambda|}\right).$$

By Proposition 2.10 of Wainwright [2019a], it follows that

$$\mathbb{P}\left(\|\boldsymbol{N}_t\|_\infty \ge \tau\right) \le 2D\exp\left(-\frac{\tau^2}{2\eta_t\|\mathrm{Var}(\boldsymbol{Z})\|_\infty + \eta_t(1-\gamma)^{-1}\tau}\right).$$

Using the last inequality and Lemma A.4, we complete the proof.

Lemma A.4. *Let X be a non-negative random variable satisfying $\mathbb{P}(X \ge \tau) \le C\exp\left(-\frac{\tau^2}{\sigma^2+b\tau}\right)$ for any $\tau \ge 0$, then*

$$\mathbb{E}X^2 \le 12\left[b^2\ln^2(eC) + \sigma^2\ln(eC)\right].$$

□

We conclude this subsection by presenting the proof of Lemma A.3 and Lemma A.4.

Proof of Lemma A.3. We first prove the following inequality

$$\sup_{i\in[D]} \mathbb{E}\exp\left(\lambda \boldsymbol{e}_i^\top \boldsymbol{N}_t\right) \le \exp\left(\frac{1}{(1-\gamma)^2}\frac{\lambda^2\eta_t}{2}\right) \text{ for any } \lambda \in \mathbb{R} \text{ and } t \ge 0.$$

We will prove it by induction on t. Fix any $i \in [D]$. The statement is vacuous for $t = 0$ since $\boldsymbol{N}_0 = \boldsymbol{0}$. We now assume that the claim holds at iteration t, and then verify that it holds at iteration $t+1$. We have

$$\begin{aligned}
\mathbb{E}\exp(\lambda \boldsymbol{e}_i^\top \boldsymbol{N}_{t+1}) &= \mathbb{E}\exp(\lambda(1-\eta_{t+1})\boldsymbol{e}_i^\top \boldsymbol{N}_t + \lambda\gamma\eta_{t+1}\boldsymbol{e}_i^\top \boldsymbol{Z}_{t+1}) \\
&= \mathbb{E}\exp(\lambda(1-\eta_{t+1})\boldsymbol{e}_i^\top \boldsymbol{N}_t)\mathbb{E}\exp(\lambda\gamma\eta_{t+1}\boldsymbol{e}_i^\top \boldsymbol{Z}_{t+1}) \\
&\overset{(a)}{\le} \mathbb{E}\exp(\lambda(1-\eta_{t+1})\boldsymbol{e}_i^\top \boldsymbol{N}_t)\cdot\exp\left(\frac{\lambda^2\gamma^2\eta_{t+1}^2}{2(1-\gamma)^2}\right) \\
&\overset{(b)}{\le} \exp\left(\frac{\lambda^2(1-\eta_{t+1})^2}{2(1-\gamma)^2}\eta_t\right)\exp\left(\frac{\lambda^2\eta_{t+1}^2}{2(1-\gamma)^2}\right) \\
&\overset{(c)}{\le} \exp\left(\frac{\lambda^2\eta_{t+1}}{2(1-\gamma)^2}\right),
\end{aligned}$$

where (a) follows from (22), (b) follows from the induction hypothesis, and (c) follows from $(1-\eta_{t+1})\eta_t \le \eta_{t+1}$. Finally, since $\|\boldsymbol{N}_T\|_\infty = \max_{i\in[D]}\{\pm\boldsymbol{e}_i^\top \boldsymbol{N}_T\}$, we have

$$\mathbb{E}\exp(\lambda\|\boldsymbol{N}_T\|_\infty) \le \sum_{i\in[D]}\left[\mathbb{E}\exp(\lambda\boldsymbol{e}_i^\top \boldsymbol{N}_T) + \mathbb{E}\exp(-\lambda\boldsymbol{e}_i^\top \boldsymbol{N}_T)\right] \le 2D\exp\left(\frac{\lambda^2\eta_T}{(1-\gamma)^2 2}\right).$$

□

Proof of Lemma A.4. Let τ_0 be the positive root of $C\exp\left(-\frac{\tau_0^2}{2(\sigma^2+b\tau_0)}\right)=1$. Hence, it follows that

$$\tau_0 = b\ln C + \sqrt{(b\ln C)^2 + 2\sigma^2\ln C} \le 2b\ln C + \sqrt{2\ln C}\sigma.$$

Then, we have

$$\begin{aligned}
\mathbb{E}X^2 &= \int_0^\infty \mathbb{P}(X^2\ge\tau)d\tau = \int_0^\infty \mathbb{P}(X\ge\sqrt{\tau})d\tau \\
&= \int_0^{\tau_0^2} \mathbb{P}(X\ge\sqrt{\tau})d\tau + \int_{\tau_0^2}^\infty \mathbb{P}(X\ge\sqrt{\tau})d\tau \\
&\le \tau_0^2 + C\int_{\tau_0^2}^\infty \exp\left(-\frac{\tau}{\sigma^2+b\sqrt{\tau}}\right)d\tau \\
&\overset{(a)}{\le} \tau_0^2 + \int_{\tau_0^2}^\infty \exp\left(-\frac{\tau}{2(\sigma^2+b\sqrt{\tau})}\right)d\tau \\
&\overset{(b)}{\le} \tau_0^2 + 12\int_{\ln C}^\infty (b^2y+\sigma^2)\exp(-y)\,dy \\
&\le \tau_0^2 + 12(b^2+\sigma^2) \le 12\left(\ln^2(eC)b^2 + \sigma^2\ln(eC)\right),
\end{aligned}$$

where (a) uses $C\le\exp\left(\frac{\tau}{2(\sigma^2+b\sqrt{\tau})}\right)$ for all $\tau\ge\tau_0^2$ and (b) uses the change of variable $y=\frac{\tau}{2(\sigma^2+b\sqrt{\tau})}$ (or equivalently $\tau=(by+\sqrt{(by)^2+2\sigma^2y})^2$) and

$$\frac{d\tau}{dy} = 2(by+\sqrt{(by)^2+2\sigma^2y})\left(b+\frac{b^2y+\sigma^2}{\sqrt{(by)^2+2\sigma^2y}}\right) \le 12(b^2y+\sigma^2).$$

□

A.3 Capturing the step size dependence

The final step is to establish the dependence of $\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2$ on $\{\eta_t\}_{t\ge0}$. 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 define

$$\widetilde{\eta}_{(t,T)} = \begin{cases} \prod_{j=1}^T(1-\widetilde{\eta}_j), & \text{if } t=0 \\ \widetilde{\eta}_t\prod_{j=t+1}^T(1-\widetilde{\eta}_j), & \text{if } 0<t<T \\ \widetilde{\eta}_T, & \text{if } t=T. \end{cases} \tag{23}$$

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

Lemma A.5. *Under Assumption 3.1, if $(1-\eta_t)\eta_{t-1}\le\eta_t$ for any $t\ge1$, then we have*

$$\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 \le 3\widetilde{\eta}_{(0,T)}^2\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{6\gamma^2\ln(2eD)}{(1-\gamma)^4}\sum_{t=1}^T\widetilde{\eta}_{(t,T)}\eta_{t-1} + \frac{6\ln(2eD)}{(1-\gamma)^2}\eta_T, \tag{24}$$

where $\{\widetilde{\eta}_{(t,T)}\}_{T\ge t\ge0}$ defined in (23) and $\{\boldsymbol{N}_t\}_{t\ge0}$ is defined in Lemma A.1.

Proof of Lemma A.5. By the recursion of $\{a_t\}_{t\geq 0}$ and $\{b_t\}_{t\geq 0}$ in Lemma A.1, 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,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 &\leq 3(a_T^2 + \mathbb{E}b_T^2 + \mathbb{E}\|\boldsymbol{N}_T\|_\infty^2)\\ &\leq 3\widetilde{\eta}_{(0,T)}^2\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{3\gamma^2}{(1-\gamma)^2}\sum_{t=1}^{T}\widetilde{\eta}_{(t,T)}\mathbb{E}\|\boldsymbol{N}_{t-1}\|_\infty^2 + 3\mathbb{E}\|\boldsymbol{N}_T\|_\infty^2. \qquad (25)\end{aligned}$$

Given the condition $(1-\eta_t)\eta_{t-1} \leq \eta_t$, we can apply Lemma A.2 which implies

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

Plugging these bounds into (25) yields (24). □

A.4 Convergence under general step sizes

Proof of Theorem A.1. Recall that Assumption 3.3 requires the step size satisfies

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

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

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

The condition (C2) implies that $(1-\eta_t)\eta_{t-1} \leq \eta_t$ for all $t \geq 1$. Hence, Lemma A.5 yields

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{\Delta}_t\|_\infty^2 \leq \frac{2}{\sqrt{T}}\sum_{t=1}^{T}\left[\widetilde{\eta}_{(0,t)}^2\|\boldsymbol{\Delta}_0\|_\infty^2 + \frac{2\gamma^2\ln(2eD)}{(1-\gamma)^4}\sum_{s=1}^{t}\widetilde{\eta}_{(s,t)}\eta_{s-1} + \frac{2\ln(2eD)}{(1-\gamma)^2}\eta_t\right].$$

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

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

which, together with the Stolz–Cesaro theorem, implies

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\widetilde{\eta}_{(0,t)}^{2}\to 0.$$

Finally, by Lemma A.6 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 A.6. *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 (23) and $\{\widetilde{\eta}_t\}_{t\ge 0}$ satisfies Assumption 3.3.*

Putting all pieces together, we have established (17). □

Proof of Lemma A.6. 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 $l\ge t\ge t_0$,

$$\frac{1}{\widetilde{\eta}_l}\le\frac{l}{t\widetilde{\eta}_t}\le\frac{1}{\widetilde{\eta}_t}\exp\left(\frac{\widetilde{m}_{t,l}+\frac{\widetilde{\eta}_t}{2}}{t\widetilde{\eta}_t}\right)\le\frac{\sqrt{e}}{\widetilde{\eta}_t}\exp\left(\frac{\widetilde{m}_{t,l}}{2}\right).\tag{26}$$

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 (26); (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, $\widetilde{\eta}_{(t,l)}=\widetilde{\eta}_{(t,t_0)}\widetilde{\eta}_{(t_0,l)}/\widetilde{\eta}_{t_0}\le\widetilde{\eta}_{(t_0,l)}$. Here we use $\widetilde{\eta}_{(t,t_0)}\le\widetilde{\eta}_{t_0}$ which follows due to $(1-\widetilde{\eta}_t)\widetilde{\eta}_{t-1}\le\widetilde{\eta}_t$ for all $t\ge 1$ from Assumption 3.3 (ii). Then we have $\sum_{l=t}^{T}\widetilde{\eta}_{(t,l)}=\sum_{l=t}^{t_0}\widetilde{\eta}_{(t,l)}+\sum_{l=t_0}^{T}\widetilde{\eta}_{(t,l)}\le t_0+\sum_{l=t_0}^{T}\widetilde{\eta}_{(t_0,l)}\le t_0+\frac{2\sqrt{e}}{\gamma}$.

- 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 $c=t_0+2\sqrt{e}/\gamma$ which ensures that $\sum_{l=t}^{T}\widetilde{\eta}_{(t,l)}\le c$ for any $T\ge t\ge 1$. □

A.5 Specific rates under two step sizes

We are ready to prove the following theorem.

Theorem A.3. *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 $t \geq 1$ and $\eta_0 = 1$, it follows that for all $T \geq 1$,*

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

, 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 Theorem A.3. In the following, we derive specific rates by plugging the specific step sizes into the result of Lemma A.5.

(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) $1-\eta_t \leq 1-\widetilde{\eta}_t = \frac{1+(1-\gamma)(t-1)}{1+(1-\gamma)t} = \widetilde{\eta}_t/\widetilde{\eta}_{t-1} = \eta_t/\eta_{t-1}$ for $t \geq 1$ and (ii) $\widetilde{\eta}_{(t,T)} \leq \widetilde{\eta}_T$. It implies Lemma A.5 is applicable. Notice that $\sum_{t=1}^T \widetilde{\eta}_{t-1} \leq 1 + \sum_{t=1}^{T-1}\frac{1}{t} \leq 1+\ln(T-1) \leq \ln(eT)$ and $\ln\frac{(1-\gamma)(T+1)}{2} \leq \ln\frac{1+(1-\gamma)(T+1)}{1+(1-\gamma)} = \int_1^{T+1}\frac{1-\gamma}{1+(1-\gamma)t}dt \leq \sum_{t=1}^T\frac{1-\gamma}{1+(1-\gamma)t} = \sum_{t=1}^T\widetilde{\eta}_t$. Hence,

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

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

Finally, plugging these inequalities into (24), we have

$$\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 \leq \frac{12\|\boldsymbol{\Delta}_0\|_\infty^2}{(1-\gamma)^2}\frac{1}{(1+T)^2} + \frac{12\gamma^2\ln(2eD)}{(1-\gamma)^5}\frac{\ln(eT)}{T}. \tag{27}$$

(II) Polynomial step size. If we choose a polynomial step size, i.e., $\eta_t = t^{-\alpha}$ with $\alpha \in (0,1)$ for $t \geq 1$ and $\eta_0 = 1$, then we again have $1-\eta_t = 1-\frac{1}{t^\alpha} \leq \left(\frac{t-1}{t}\right)^\alpha = \eta_t/\eta_{t-1}$ for $t \geq 1$, which implies Lemma A.2 is applicable. Note that

$$\frac{(T+1)^{1-\alpha}-(t+1)^{1-\alpha}}{1-\alpha} = \int_{t+1}^{T+1} j^{-\alpha}dj \leq \sum_{j=t+1}^T j^{-\alpha} \leq \int_t^T j^{-\alpha}dj = \frac{T^{1-\alpha}-t^{1-\alpha}}{1-\alpha}, \tag{28}$$

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

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

Additionally, using $\eta_{t-1} \leq 2\eta_t$ for all $t \geq 1$ and (28), we have,

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

which implies

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

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

Lemma A.7. *For any $\alpha \in (0,1)$ and $\beta > 0$, it follows that*

$$\sum_{t=1}^{T}\frac{\exp\left(\frac{1-\gamma}{1-\alpha}t^{1-\alpha}\right)}{t^{\beta}} \leq \left(\frac{\beta}{1-\gamma}\right)^{\frac{1}{1-\alpha}}\exp\left(\frac{1-\gamma}{1-\alpha}\right) + \frac{\beta}{(1-\gamma)\alpha}\frac{\exp\left(\frac{1-\gamma}{1-\alpha}(1+T)^{1-\alpha}\right)}{(1+T)^{\beta-\alpha}}. \tag{29}$$

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

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

Therefore,

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

Putting together the pieces, we can safely conclude that

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

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 A.7. 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,

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

Using integrating by parts, it follows that

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

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

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

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. □

A.6 Proof of Theorem A.2

Proof of Theorem A.2. The result directly follows from Theorem A.3.

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

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

- For the second item, we have $\mathbb{E}\|\boldsymbol{\Delta}_T\|_\infty^2 \le \Delta_0\exp\left(-\frac{1-\gamma}{1-\alpha}\left((1+T)^{1-\alpha}-1\right)\right) + \frac{114\ln(2eD)}{(1-\gamma)^4}\frac{1}{T^\alpha}$ 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

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

and $\sum_{t=1}^T t^{-\alpha} \le \int_0^T t^{-\alpha}dt = \frac{T^{1-\alpha}}{1-\alpha}$. 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

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

which implies that

$$\exp\left(\frac{1-\gamma}{1-\alpha}\right)(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right) \le \frac{\sqrt{2\pi e}}{\sqrt{1-\alpha}}. \tag{30}$$

Therefore,

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

□

B Proof of Theorem 3.2

B.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.

B.1.1 Weak convergence of measures in Polish spaces

We will introduce some basic knowledge of weak convergence in metric spaces. See Chapter VI of 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 $w_1, w_2 \in \mathsf{D}([0,1],\mathbb{R}^d)$,

$$d_0(w_1,w_2) = \inf_{\lambda\in\Lambda} \left\{ \sup_{0\le s<t\le 1} \left| \ln \frac{\lambda(t)-\lambda(s)}{t-s} \right| + \sup_{t\in[0,1]} \|w_1(\lambda(t)) - w_2(t)\|_\infty \right\}, \tag{31}$$

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

$$\|\omega\|_{\sup} := \sup_{r\in[0,1]} \|\omega(r)\|_\infty. \tag{32}$$

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 $\{X_t\}_{t\ge 0} \subseteq \mathsf{D}([0,1],\mathbb{R}^d)$ 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 $\{X_t\}_{t\ge 0} \subseteq \mathsf{D}([0,1],\mathbb{R}^d)$ weakly converges to X, if for any bounded continuous function $f : \mathsf{D}([0,1],\mathbb{R}^d) \to \mathbb{R}$, we have

$$\mathbb{E}f(X_T) \to \mathbb{E}f(X) \text{ as } T \text{ goes to infinity.} \tag{33}$$

The condition is equivalent to (13) in the text. We denote the weak convergence by $X_T \xrightarrow{w} X$.

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

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

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

B.1.2 Proof idea

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

$$\boldsymbol{\phi}_T \xrightarrow{w} \mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D,$$

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 $\boldsymbol{\phi}_T$ weakly converges to the measure introduced by $\mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D$ 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 B.1. The proof of Lemma B.1 can be found in Appendix B.4.1.

Lemma B.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 $t \geq 1$*

$$\boldsymbol{\Delta}_t^1 = \boldsymbol{A}_t \boldsymbol{\Delta}_{t-1}^1 + \eta_t \left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{t-1}\right] \tag{34}$$

$$\boldsymbol{\Delta}_t^2 = \boldsymbol{A}_t \boldsymbol{\Delta}_{t-1}^2 + \eta_t \boldsymbol{W}_t \ . \tag{35}$$

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

$$\boldsymbol{\Delta}_t^2 \leq \boldsymbol{\Delta}_t \leq \boldsymbol{\Delta}_t^1. \tag{36}$$

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

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

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

$$\boldsymbol{\phi}_T^2(r) \leq \boldsymbol{\phi}_T(r) \leq \boldsymbol{\phi}_T^1(r). \tag{37}$$

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

$$\mathcal{Z} \xrightarrow{w} \mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D. \tag{38}$$

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

$$\lim_{T\to\infty} \mathbb{E}\|\boldsymbol{\phi}_T^k - \mathcal{Z}\|_{\sup} = 0 \text{ for } k = 1, 2. \tag{39}$$

By the sandwich inequality (37), we have

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

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

B.2 FCLT for $\boldsymbol{\phi}_T^1$

We first establish the FLCT of $\boldsymbol{\phi}_T^1(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{\Delta}_t^1$, i.e., $\lim_{T\to\infty}\mathbb{E}\|\boldsymbol{\phi}_T^1 - \mathcal{Z}\|_{\sup} = 0$ for some $\mathcal{L}(\mathcal{Z}) = \mathcal{L}(\mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D)$. The FCLT of $\boldsymbol{\phi}_T^2(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{\Delta}_t^2$ can be validated in an almost identical way. We start by rewriting (34) as

$$\boldsymbol{\Delta}_t^1 = \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^1 + \eta_t\left(\boldsymbol{Z}_t + \gamma\boldsymbol{D}_{t-1}^1\right), \tag{40}$$

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

$$\boldsymbol{D}_{t-1}^1 = (\boldsymbol{P}_t - \boldsymbol{P})(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*) + (\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{t-1}. \tag{41}$$

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 (40) gives

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

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

$$\begin{aligned}
\boldsymbol{\phi}_T^1(r) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{\Delta}_t^1 = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\prod_{j=1}^{t}\boldsymbol{A}_j\boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\sum_{j=1}^{t}\prod_{i=j+1}^{t}\boldsymbol{A}_i\eta_j\left(\boldsymbol{Z}_j + \gamma\boldsymbol{D}_{j-1}^1\right)\\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\prod_{j=1}^{t}\boldsymbol{A}_j\boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\sum_{t=j}^{\lfloor Tr\rfloor}\prod_{i=j+1}^{t}\boldsymbol{A}_i\eta_j\left(\boldsymbol{Z}_j + \gamma\boldsymbol{D}_{j-1}^1\right)\\
&= \frac{1}{\eta_0\sqrt{T}}(\boldsymbol{A}_0^{\lfloor Tr\rfloor} - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^{\lfloor Tr\rfloor}\left(\boldsymbol{Z}_j + \gamma\boldsymbol{D}_{j-1}^1\right),
\end{aligned} \tag{42}$$

where the last line uses the following notation:

$$\boldsymbol{A}_j^T = \eta_j\sum_{t=j}^{T}\prod_{i=j+1}^{t}\boldsymbol{A}_i \text{ for any } T \geq j \geq 0. \tag{43}$$

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 B.4). By the observation, we further expand (42) and decompose $\boldsymbol{\phi}_T^1(r)$ into six terms $\{\boldsymbol{\psi}_i\}_{i=0}^5$ which will be analyzed respectively in the following:

$$\begin{aligned}
\boldsymbol{\phi}_T^1(r) &= \frac{1}{\eta_0\sqrt{T}}(\boldsymbol{A}_0^{\lfloor Tr\rfloor} - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}}\sum_{j=1}^{[Tr]}\boldsymbol{A}_j^{\lfloor Tr\rfloor}\left(\boldsymbol{Z}_j + \gamma\boldsymbol{D}_{j-1}^1\right)\\
&= \frac{1}{\eta_0\sqrt{T}}(\boldsymbol{A}_0^{\lfloor Tr\rfloor} - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{G}^{-1}\boldsymbol{Z}_j + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}(\boldsymbol{A}_j^{\lfloor Tr\rfloor} - \boldsymbol{G}^{-1})\boldsymbol{Z}_j
\end{aligned}$$

$$
\begin{aligned}
&+\frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^{\lfloor Tr\rfloor}(\boldsymbol{P}_j-\boldsymbol{P})(\boldsymbol{V}_{j-1}-\boldsymbol{V}^*)+\frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^{\lfloor Tr\rfloor}(\boldsymbol{P}^{\pi_{j-1}}-\boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{j-1}\\
=&\frac{1}{\eta_0\sqrt{T}}(\boldsymbol{A}_0^{\lfloor Tr\rfloor}-\eta_0\boldsymbol{I})\boldsymbol{\Delta}_0+\frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{G}^{-1}\boldsymbol{Z}_j+\frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}(\boldsymbol{A}_j^T-\boldsymbol{G}^{-1})\boldsymbol{Z}_j\\
&+\frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^T(\boldsymbol{P}_j-\boldsymbol{P})(\boldsymbol{V}_{j-1}-\boldsymbol{V}^*)+\frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}(\boldsymbol{A}_j^{\lfloor Tr\rfloor}-\boldsymbol{A}_j^T)\left[\boldsymbol{Z}_j+\gamma(\boldsymbol{P}_j-\boldsymbol{P})(\boldsymbol{V}_{j-1}-\boldsymbol{V}^*)\right]\\
&+\frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^{\lfloor Tr\rfloor}(\boldsymbol{P}^{\pi_{j-1}}-\boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{j-1}\\
:=&\boldsymbol{\psi}_0(r)+\boldsymbol{\psi}_1(r)+\boldsymbol{\psi}_2(r)+\boldsymbol{\psi}_3(r)+\boldsymbol{\psi}_4(r)+\boldsymbol{\psi}_5(r). \qquad (44)
\end{aligned}
$$

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 (38) is true by setting $\mathcal{Z}=\boldsymbol{\psi}_1$. In order to establish (39), we will show that $\mathbb{E}\|\boldsymbol{\psi}_i\|_{\sup}=o(1)$ for $i=0,2,3,4,5$. In this way, based on (44), we have

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

and validate (39). To that end, we first study the properties of $\{\boldsymbol{A}_j^T\}_{0\le j\le T}$ since it appears in many $\boldsymbol{\psi}_i$'s.

B.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 $\{\eta_t\}_{t\ge0}$ satisfying Assumption 3.3 and establishes the following lemma.

Lemma B.2 (Lemma 1 in Polyak and Juditsky [1992]). *For $\{\eta_t\}_{t\ge0}$ satisfying Assumption 3.3,*

- *Uniform boundedness: $\|\boldsymbol{A}_j^T\|_\infty\le C_0$ uniformly for all $T\ge j\ge0$ for some constant $C_0\ge1$;*
- *Uniform approximation: $\lim_{T\to\infty}\frac{1}{T}\sum_{j=1}^T\|\boldsymbol{A}_j^T-\boldsymbol{G}^{-1}\|_2=0$.*

Lemma B.2 shows that when the step size η_t decreases at a slow rate, $\boldsymbol{A}_j^T$ is uniformly bouned (that is $\sup_{T\ge j\ge1}\|\boldsymbol{A}_j^T\|_\infty<\infty$) and is a good surrogate of $\boldsymbol{G}^{-1}:=(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^*})^{-1}$ in the asymptotic sense: $\lim_{T\to\infty}\frac{1}{T}\sum_{j=1}^T\|\boldsymbol{A}_j^T-\boldsymbol{G}^{-1}\|_2=0$.[6] 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 $t\ge1$ and $\eta_0=1$.

[6] 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 B.2.

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

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

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

The second is the uniform approximation. The proof is deferred in Appendix B.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 B.2 since (S1) doesn't satisfy Assumption 3.3.

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

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

B.2.2 Establishing the FCLT

Uniform negligibility of $\boldsymbol{\psi}_0$. It is clear that $\boldsymbol{\psi}_0$ is a deterministic function. Using the uniform boundedness of $\boldsymbol{A}_j^T (T \ge j \ge 0)$ in Lemma B.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 $\|\boldsymbol{\Delta}_0\|_\infty \le \frac{1}{1-\gamma}$.

Partial-sum asymptotic behavior of $\boldsymbol{\psi}_1$. 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 B.5. *For any $r \in [0,1]$,*

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

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

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

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),

$$
\begin{aligned}
\mathbb{E} \sup_{r \in [0,1]} \|\psi_2(r)\|_2^2 &= \mathbb{E} \sup_{0 \le t \le T} \|\boldsymbol{X}_t\|_2^2 \overset{(*)}{\le} 4\mathbb{E}\|\boldsymbol{X}_T\|_2^2 \\
&= 4\mathbb{E}\|\mathcal{T}_2(1)\|_2^2 = 4\mathbb{E} \left\| \frac{1}{\sqrt{T}} \sum_{j=1}^{T} (\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}) \boldsymbol{Z}_j \right\|_2^2 \\
&= \frac{4}{T} \sum_{j=1}^{T} \mathbb{E}\|(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}) \boldsymbol{Z}_j\|_2^2 \le \frac{4}{T} \sum_{j=1}^{T} \|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_2^2 \mathbb{E}\|\boldsymbol{Z}_j\|_2^2 \\
&\le c_1 \cdot \frac{1}{T} \sum_{j=1}^{T} \|\boldsymbol{A}_j^T - \boldsymbol{G}^{-1}\|_2.
\end{aligned}
$$

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 $c_1 = (\frac{1}{1-\gamma} + \sup_{T \ge j} \|\boldsymbol{A}_j^T\|_2) \mathrm{tr}(\mathsf{Var}_{\boldsymbol{Q}})$ thanks to Lemma B.2. In addition, Lemma B.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, $\mathbb{E}\|\psi_2\|_{\sup} = \mathbb{E} \sup_{r \in [0,1]} \|\psi_2(r)\|_\infty \le \mathbb{E} \sup_{r \in [0,1]} \|\psi_2(r)\|_2 \le \sqrt{\mathbb{E} \sup_{r \in [0,1]} \|\psi_2(r)\|_2^2} = o(1)$.

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 ψ_2, we have $\mathbb{E} \sup_{r \in [0,1]} \|\psi_3(r)\|_2^2 \le 4\mathbb{E}\|\psi_3(1)\|_2^2$ by Doob's maximum inequality. Therefore,

$$
\begin{aligned}
\mathbb{E} \sup_{r \in [0,1]} \|\psi_3(r)\|_2^2 \le 4\mathbb{E}\|\psi_3(1)\|_2^2 &\overset{(a)}{=} \frac{4}{T} \sum_{j=1}^{T} \mathbb{E}\|\boldsymbol{A}_j^T (\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)\|_2^2 \\
&\le \frac{4}{T} \sum_{j=1}^{T} \|\boldsymbol{A}_j^T\|_2^2 \mathbb{E}\|\boldsymbol{P}_j - \boldsymbol{P}\|_2^2 \|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_2^2 \\
&\overset{(b)}{\le} c_2 \cdot \frac{1}{T} \sum_{j=1}^{T} \mathbb{E}\|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_2^2,
\end{aligned}
$$

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 B.2 and $\|\boldsymbol{P}_j - \boldsymbol{P}\|_2^2 \le D\|\boldsymbol{P}_j - \boldsymbol{P}\|_\infty^2 = 4D$. By Theorem A.3, 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, $\mathbb{E}\|\psi_3(r)\|_{\sup} = \mathbb{E} \sup_{r \in [0,1]} \|\psi_3(r)\|_\infty \le \mathbb{E} \sup_{r \in [0,1]} \|\psi_3(r)\|_2 \le \sqrt{\mathbb{E} \sup_{r \in [0,1]} \|\psi_3(r)\|_2^2} = o(1)$.

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 $\boldsymbol{\varepsilon}_j = \boldsymbol{Z}_j + \gamma(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)$. 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.

Recall $\mathcal{F}_t = \sigma(\{\boldsymbol{r}_j, \boldsymbol{P}_j\}_{0\le j\le t})$ 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 (43), it follows that for any $0 \le k \le T$,

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

On one hand, $\|\boldsymbol{A}_{k+1}^T\boldsymbol{A}_{k+1}\|_2 \le c_3$ is uniformly bounded with $c_3 = (\sup_{T\ge j}\|\boldsymbol{A}_j^T\|_2)(1+\|\boldsymbol{G}\|_2)$ for any $T \ge k+1$ from Lemma B.2. On the other hand, we define an auxiliary sequence $\{\boldsymbol{Y}_k\}_{k\ge1}$ 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 $k \ge 1$. 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 $k \ge 0$. These results imply we can apply Lemma A.8 of Li et al. [2021c]. Putting these pieces together, we have that

$$\begin{aligned}\|\psi_4\|_{\sup} \le \sup_{r\in[0,1]}\|\psi_4(r)\|_2 &\le c_3\sup_{0\le k\le T}\left\|\frac{1}{\sqrt{T}\eta_{k+1}}\sum_{j=1}^{k}\left(\prod_{i=j+1}^{k}\boldsymbol{A}_i\right)\eta_j\boldsymbol{\varepsilon}_j\right\|_2\\ &= c_3\sup_{0\le k\le T}\frac{1}{\sqrt{T}}\frac{\|\boldsymbol{Y}_{k+1}\|_2}{\eta_{k+1}} \overset{(*)}{=} o_{\mathbb{P}}(1),\end{aligned}$$

where $(*)$ follows from Lemma A.8 of Li et al. [2021c].

Uniform negligibility of ψ_5. In the following, we will prove $\|\psi_5\|_{\sup} = o_{\mathbb{P}}(1)$ by showing $\mathbb{E}\|\psi_5\|_{\sup} = o(1)$. It is worth mentioning that ψ_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 ψ_5 would disappear in the case. Notice that $\psi_5(r) = \frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^{\lfloor Tr\rfloor}(\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{j-1}$ is a sum of correlated random variables (which are even not mean-zero). We need a high-order residual condition (45) to bound $\mathbb{E}\|\psi_5\|_{\sup}$, which is ensured by a positive optimality gap as Lemma B.6 shows. With such a Lipschitz condition, Lemma B.7 shows $\mathbb{E}\|\psi_5\|_{\sup}$ is dominated by $\frac{1}{\sqrt{T}}\sum_{j=1}^{T}\mathbb{E}\,\|\boldsymbol{\Delta}_{j-1}\|_\infty^2$, which is $o(1)$ for the general step size as suggested by Theorem A.1. The proof of Lemma B.6 is deferred in Appendix B.4.4 and that of Lemma B.7 is in Appendix B.4.5.

Lemma B.6. *If π^* is unique, then we have a positive optimality gap* $\mathrm{gap} > 0$. *Here* $\mathrm{gap} := \min_s\min_{a\ne\pi^*(s)}|V^*(s) - Q^*(s,a)|$ *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*

$$\|(\boldsymbol{P}^{\pi_{\boldsymbol{Q}}} - \boldsymbol{P}^{\pi^*})(\boldsymbol{Q} - \boldsymbol{Q}^*)\|_\infty \le L\|\boldsymbol{Q} - \boldsymbol{Q}^*\|_\infty^2 \text{ with } L = \frac{4}{\text{gap}}, \tag{45}$$

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.*

Lemma B.7. *With a positive optimality gap defined in Lemma B.6, define* $L = \frac{4}{\text{gap}}$*. It follows that*

$$\mathbb{E}\|\boldsymbol{\psi}_5\|_{\sup} = \mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_5(r)\|_\infty \le \gamma L C_0 \cdot \frac{1}{\sqrt{T}}\sum_{j=1}^{T}\mathbb{E}\,\|\boldsymbol{\Delta}_{j-1}\|_\infty^2 .$$

Putting the pieces together. From (42), $\boldsymbol{\phi}_T^1 = \sum_{i=0}^5 \boldsymbol{\psi}_i$. We have shown $\boldsymbol{\psi}_1 \xrightarrow{w} \mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D$ in the sense of $(\mathsf{D}([0,1],\mathbb{R}^D), d_0)$ and $\|\boldsymbol{\psi}_i\|_{\sup} = o_{\mathbb{P}}(1)$ for $i \neq 1$. Using $\|\boldsymbol{\phi}_T^1 - \boldsymbol{\psi}_1\|_{\sup} \le \sum_{i\neq 1}\|\boldsymbol{\psi}_i\|_{\sup}$, we know that $\|\boldsymbol{\phi}_T^1 - \boldsymbol{\psi}_1\|_{\sup} = o_{\mathbb{P}}(1)$. Proposition B.1 implies $\boldsymbol{\phi}_T^1 \xrightarrow{w} \mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D$. We then establish the FCLT for $\boldsymbol{\phi}_T^1(r)$.

B.3 FCLT for $\boldsymbol{\phi}_T^2$

We can repeat the above analysis for $\boldsymbol{\phi}_T^2$. We rewrite (35) as

$$\boldsymbol{\Delta}_t^2 = \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^2 + \eta_t\left(\boldsymbol{Z}_t + \gamma\boldsymbol{D}_{t-1}^2\right), \tag{46}$$

where $\boldsymbol{A}_t = \boldsymbol{I} - \eta_t(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})$ and $\boldsymbol{Z}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*$ are the same as those defined in (40) except that $\boldsymbol{D}_{t-1}^1$ (defined in (41)) is replaced by

$$\boldsymbol{D}_{t-1}^2 = (\boldsymbol{P}_t - \boldsymbol{P})(\boldsymbol{V}_{t-1} - \boldsymbol{V}^*). \tag{47}$$

Since $\boldsymbol{D}_{t-1}^2$ is much simpler than $\boldsymbol{D}_{t-1}^1$, the analysis for $\boldsymbol{\phi}_T^2(r)$ should be easier than $\boldsymbol{\phi}_T^1(r)$. Using the notation $\boldsymbol{A}_j^T$ (see(43)), we decompose $\boldsymbol{\phi}_T^2(r)$ into five terms:

$$\begin{aligned}
\boldsymbol{\phi}_T^2(r) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{\Delta}_t^2 = \frac{1}{\eta_0\sqrt{T}}(\boldsymbol{A}_0^{\lfloor Tr\rfloor} - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^{\lfloor Tr\rfloor}\left(\boldsymbol{Z}_j + \gamma\boldsymbol{D}_{j-1}^2\right)\\
&= \frac{1}{\eta_0\sqrt{T}}(\boldsymbol{A}_0^{\lfloor Tr\rfloor} - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{G}^{-1}\boldsymbol{Z}_j + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j\\
&\quad + \frac{\gamma}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}\boldsymbol{A}_j^T(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)\\
&\quad + \frac{1}{\sqrt{T}}\sum_{j=1}^{\lfloor Tr\rfloor}(\boldsymbol{A}_j^{\lfloor Tr\rfloor} - \boldsymbol{A}_j^T)\left[\boldsymbol{Z}_j + \gamma(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*)\right]\\
&:= \boldsymbol{\psi}_0(r) + \boldsymbol{\psi}_1(r) + \boldsymbol{\psi}_2(r) + \boldsymbol{\psi}_3(r) + \boldsymbol{\psi}_4(r).
\end{aligned} \tag{48}$$

Here $\{\boldsymbol{\psi}_i\}_{i=0}^4$ are exactly the same as those in (44). Our previous analysis provides us a low-hanging fruit result: $\boldsymbol{\psi}_1 \xrightarrow{w} \mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D$ in the sense of $(\mathsf{D}([0,1],\mathbb{R}^D), d_0)$ and $\|\boldsymbol{\psi}_i\|_{\sup} = o_{\mathbb{P}}(1)$ for $i \neq 1$. Then we know that $\|\boldsymbol{\phi}_T^2 - \mathcal{T}_1\|_{\sup} = o(1)$ and $\boldsymbol{\phi}_T^2 \xrightarrow{w} \mathsf{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D$ due to Proposition B.1. We thus establish the FCLT for $\boldsymbol{\phi}_T^2$.

B.4 Proofs of lemmas

B.4.1 Proof of Lemma B.1

Proof of Lemma B.1. We use mathematical induction to prove the statement. When $t = 0$, the inequality (36) holds by initialization. Assume (36) holds at $t-1$, i.e., $\boldsymbol{\Delta}_{t-1}^2 \leq \boldsymbol{\Delta}_{t-1} \leq \boldsymbol{\Delta}_{t-1}^1$. Let us analyze the case of t. By the Q-learning update rule (e.g., see (20)), it follows that

$$\begin{aligned}
\boldsymbol{\Delta}_t &= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\left[(\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^*)\right] \\
&\overset{(a)}{=} (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^*)\right] \\
&\overset{(b)}{=} (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \eta_t\left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \boldsymbol{P}^{\pi^*}\boldsymbol{Q}^*)\right] \\
&\overset{(c)}{=} \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1} + \eta_t\left[\boldsymbol{W}_t + \gamma(\boldsymbol{P}^{\pi_{t-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{Q}_{t-1}\right],
\end{aligned} \tag{49}$$

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 $\sup_t \eta_t \leq 1$), then $\boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^2 \leq \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1} \leq \boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}^1$.

For one hand, based on (49), we have

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

where the last inequality uses $\boldsymbol{P}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} \geq \boldsymbol{P}^{\pi^*}\boldsymbol{Q}_{t-1}$ 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

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

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

B.4.2 Proof of Lemma B.3

Proof of Lemma B.3. By the definition of (43), we have $\|\boldsymbol{A}_j^T\|_\infty \leq \eta_j \sum_{t=j}^T \prod_{i=j+1}^t (1-\widetilde{\eta}_i)$. Plugging the specific form of $\{\eta_t\}$, we have for (S1)

$$\begin{aligned}
\|\boldsymbol{A}_j^T\|_\infty &\leq \eta_j \sum_{t=j}^T \prod_{i=j+1}^t \frac{1+(1-\gamma)(i-1)}{1+(1-\gamma)i} = \eta_j \sum_{t=j}^T \frac{1+(1-\gamma)j}{1+(1-\gamma)t} \\
&\leq \frac{1}{1-\gamma} \ln \frac{1+(1-\gamma)T}{1+(1-\gamma)(j-1)} \leq \frac{\ln(1+(1-\gamma)T)}{1-\gamma}
\end{aligned} \tag{50}$$

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 $\sum_{i=j}^{t} i^{-\alpha} \ge \frac{1}{1-\alpha}((t+1)^{1-\alpha} - j^{1-\alpha})$ and $\exp((1-\gamma)j^{-\alpha}) \le e$, (b) uses the change of variable $y = \frac{1-\gamma}{1-\alpha}(t^{1-\alpha} - j^{1-\alpha})$, (c) uses $(a+b)^p \le \max\{2^{p-1}, 1\}(a^p + b^p)$ for any $p > 0$, and (d) uses $(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right) \le \frac{\sqrt{2\pi}e^{1/2}}{\sqrt{1-\alpha}}$ from (30) and $\max\left\{2^{\frac{\alpha}{1-\alpha}}, 1\right\} \le 2^{\frac{1}{1-\alpha}}$. □

B.4.3 Proof of Lemma B.4

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

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

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

For (S2), based on (43) and $\boldsymbol{G} = \eta_j^{-1}(\boldsymbol{I} - (\boldsymbol{I} - \eta_j \boldsymbol{G}))$, we have

$$\begin{aligned}
\boldsymbol{A}_j^T - \boldsymbol{G}^{-1} &= (\boldsymbol{A}_j^T \boldsymbol{G} - \boldsymbol{I})\boldsymbol{G}^{-1} = \sum_{t=j}^{T} \left(\prod_{i=j+1}^{t} (\boldsymbol{I} - \eta_i \boldsymbol{G}) - \prod_{i=j}^{t} (\boldsymbol{I} - \eta_i \boldsymbol{G}) \right) \boldsymbol{G}^{-1} - \boldsymbol{G}^{-1} \\
&= \sum_{t=j+1}^{T} \left(\prod_{i=j+1}^{t} (\boldsymbol{I} - \eta_i \boldsymbol{G}) - \prod_{i=j}^{t-1} (\boldsymbol{I} - \eta_i \boldsymbol{G}) \right) \boldsymbol{G}^{-1} - \prod_{t=j}^{T} (\boldsymbol{I} - \eta_t \boldsymbol{G}) \boldsymbol{G}^{-1} \\
&= \sum_{t=j+1}^{T} (\eta_j - \eta_t) \prod_{i=j+1}^{t-1} (\boldsymbol{I} - \eta_i \boldsymbol{G}) - \prod_{t=j}^{T} (\boldsymbol{I} - \eta_t \boldsymbol{G}) \boldsymbol{G}^{-1} \\
&:= \boldsymbol{M}_{T,j}^{(1)} + \boldsymbol{M}_{T,j}^{(2)}. \qquad (51)
\end{aligned}$$

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,

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

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

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

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

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

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

Therefore,

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

□

Proof of Lemma B.8. Clearly we have

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

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

$$\begin{aligned}
\sum_{t=j}^{T} \widetilde{m}_{j,t} \exp\left(-\widetilde{m}_{j,t}\right) &= \sum_{t=j}^{T} \widetilde{m}_{j,t} \exp(-\widetilde{m}_{j-1,t-1}) \exp(\widetilde{\eta}_{j-1} - \widetilde{\eta}_t) \\
&= e\sum_{t=j}^{T} \frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right) \exp\left(-\frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right)\right) \\
&\leq 2e\int_{j-1}^{\infty} \frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right) \exp\left(-\frac{1-\gamma}{1-\alpha}\left(t^{1-\alpha} - (j-1)^{1-\alpha}\right)\right) dt \\
&\overset{(a)}{=} \frac{2e}{1-\gamma}\int_0^{\infty} y\exp\left(-y\right)\left(\frac{1-\alpha}{1-\gamma}y + (j-1)^{1-\alpha}\right)^{\frac{\alpha}{1-\alpha}} dt \\
&\overset{(b)}{\leq} \frac{e\max\{2^{\frac{\alpha}{1-\alpha}}, 2\}}{1-\gamma}\int_0^{\infty} y\exp\left(-y\right)\left[\left(\frac{1-\alpha}{1-\gamma}y\right)^{\frac{\alpha}{1-\alpha}} + (j-1)^{\alpha}\right] dt \\
&\overset{(c)}{\leq} \frac{e2^{\frac{1}{1-\alpha}}}{1-\gamma}\left[\left(\frac{1-\alpha}{1-\gamma}\right)^{\frac{\alpha}{1-\alpha}}\Gamma\left(1+\frac{1}{1-\alpha}\right) + (j-1)^{\alpha}\right] \\
&\overset{(d)}{\leq} \frac{\sqrt{2\pi}e^{\frac{3}{2}}2^{\frac{1}{1-\alpha}}}{(1-\alpha)^{\frac{3}{2}}} \frac{1}{(1-\gamma)^{\frac{1}{1-\alpha}}} + \frac{e2^{\frac{1}{1-\alpha}}}{1-\gamma}(j-1)^{\alpha},
\end{aligned}$$

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 (30). □

B.4.4 Proof of Lemma B.6

Proof of Lemma B.6. 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.4.5 Proof of Lemma B.7

Proof of Lemma B.7. By Lemma B.6 and Lemma B.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 B.2. □

C Proof of Auxiliary Results

C.1 Proof of Theorem 3.1

Proof of Theorem 3.1. One can prove Theorem 3.1 by applying continuous mapping theorem to Theorem 3.2 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 (33) would conclude the proof. Recalling the metric (31) 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. □

C.2 Proof of Corollary 3.1

Proof of Corollary 3.1. We first prove

$$\mathsf{Var}_{\boldsymbol{V}} = \boldsymbol{\Pi}^{\pi^*} \mathsf{Var}_{\boldsymbol{Q}} (\boldsymbol{\Pi}^{\pi^*})^\top. \tag{52}$$

Recall the definition

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

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, (52) 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

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

If we can prove

$$\sqrt{T}\left(\boldsymbol{\Pi}^{\bar{\pi}_T} \bar{\boldsymbol{Q}}_T - \boldsymbol{\Pi}^{\pi^*} \bar{\boldsymbol{Q}}_T\right) = o_{\mathbb{P}}(1), \tag{53}$$

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

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

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 $\|\mathbf{\Pi}^{\bar{\pi}_T} - \mathbf{\Pi}^{\pi^*}\|_\infty = 2\cdot 1_{\{\bar{\pi}_T \neq \pi^*\}}$, (c) uses the fact $\{\bar{\pi}_t \neq \pi^*\} \subseteq \{\|\bar{\boldsymbol{Q}}_t - \boldsymbol{Q}^*\|_\infty \ge \frac{\mathrm{gap}}{2}\}$ which we derived in the proof of Lemma B.6, and finally (d) follows from Jensen's inequality.

From Theorem A.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 $\sqrt{T}\mathbb{E}\|\mathbf{\Pi}^{\bar{\pi}_T}\bar{\boldsymbol{Q}}_T - \mathbf{\Pi}^{\pi^*}\bar{\boldsymbol{Q}}_T\|_\infty = o(1)$ which implies (53) is true. □

C.3 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

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

Once we prove g is continuous in $(\mathrm{dom}(g), d_0)$, the continuous mapping theorem together with Theorem 3.2 would complete the proof for Proposition 3.1.

In Appendix C.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 $(\mathrm{dom}(g), d_0)$. 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 3.2 and definition of weak convergence, we know that as T goes to infinity,

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

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. □

D Proof of Theorem 4.1

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 B. The average Q-learning estimator $\bar{\boldsymbol{Q}}_T$ has the error

$$\bar{\boldsymbol{\Delta}}_T := \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Delta}_t = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Q}_t - \boldsymbol{Q}^*). \tag{54}$$

Using two auxiliary sequences $\{\boldsymbol{\Delta}_t^1\}_{t\geq 0}$ and $\{\boldsymbol{\Delta}_t^2\}_{t\geq 0}$ defined in Lemma B.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

$$\bar{\boldsymbol{\Delta}}_T^2 \leq \bar{\boldsymbol{\Delta}}_T \leq \bar{\boldsymbol{\Delta}}_T^1. \tag{55}$$

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 B.1), they share a similar error decomposition.

D.1 Error decomposition

Setting $r = 1$ in (42), 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 (44), we decompose $\bar{\boldsymbol{\Delta}}_T^1$ into five separate terms

$$\begin{aligned}\bar{\boldsymbol{\Delta}}_T^1 &= \frac{1}{\eta_0 T}(\boldsymbol{A}_0^T - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{T}\sum_{j=1}^{T}\boldsymbol{G}^{-1}\boldsymbol{Z}_j + \frac{1}{T}\sum_{j=1}^{T}(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j \\ &\quad + \frac{\gamma}{T}\sum_{j=1}^{T}\boldsymbol{A}_j^T(\boldsymbol{P}_j - \boldsymbol{P})(\boldsymbol{V}_{j-1} - \boldsymbol{V}^*) + \frac{\gamma}{T}\sum_{j=1}^{T}\boldsymbol{A}_j^T(\boldsymbol{P}^{\pi_{j-1}} - \boldsymbol{P}^{\pi^*})\boldsymbol{\Delta}_{j-1} \\ &:= \mathcal{T}_0 + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4. \end{aligned} \tag{56}$$

Here one should distinguish $\mathcal{T}_i$ with $\boldsymbol{\psi}_i$, the former a random variable and the latter a random function. Comparing (41) and (47), 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

$$\begin{aligned}\bar{\boldsymbol{\Delta}}_T^2 &= \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Delta}_t^2 = \frac{1}{\eta_0 T}(\boldsymbol{A}_0^T - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{T}\sum_{j=1}^{T}\boldsymbol{A}_j^T\left(\boldsymbol{Z}_j + \gamma\boldsymbol{D}_{j-1}^2\right) \\ &= \frac{1}{\eta_0 T}(\boldsymbol{A}_0^T - \eta_0\boldsymbol{I})\boldsymbol{\Delta}_0 + \frac{1}{T}\sum_{j=1}^{T}\boldsymbol{G}^{-1}\boldsymbol{Z}_j + \frac{1}{T}\sum_{j=1}^{T}(\boldsymbol{A}_j^T - \boldsymbol{G}^{-1})\boldsymbol{Z}_j\end{aligned}$$

$$+\frac{\gamma}{T}\sum_{j=1}^{T}\boldsymbol{A}_j^T(\boldsymbol{P}_j-\boldsymbol{P})(\boldsymbol{V}_{j-1}-\boldsymbol{V}^*)$$
$$=\mathcal{T}_0+\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3. \tag{57}$$

Here $\{\mathcal{T}_i\}_{i=0}^3$ are exactly the same as in (56). Putting the pieces together, we have

$$\mathbb{E}\|\bar{\boldsymbol{\Delta}}_T\|_\infty \le \mathbb{E}\max\{\|\bar{\boldsymbol{\Delta}}_T^1\|_\infty, \|\bar{\boldsymbol{\Delta}}_T^2\|_\infty\} \le \sum_{i=0}^{4}\mathbb{E}\|\mathcal{T}_i\|_\infty. \tag{58}$$

D.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

$$\|\mathcal{T}_0\|_\infty = \frac{1}{\eta_0 T}\|(\boldsymbol{A}_0^T-\eta_0\boldsymbol{I})\boldsymbol{\Delta}_0\|_\infty \le \frac{1}{\eta_0 T}(\|\boldsymbol{A}_0^T\|_\infty+\eta_0)\|\boldsymbol{\Delta}_0\|_\infty \le \frac{2C_0}{1-\gamma}\frac{1}{T}. \tag{59}$$

For $\|\mathcal{T}_1\|_\infty$. We apply (73) in Lemma F.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 $\|\boldsymbol{W}_T\|_\infty \le \frac{\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty}{T}$ defined therein. Hence,

$$\mathbb{E}\|\mathcal{T}_1\|_\infty \le 6\sqrt{\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln(2D)}{T}} + \frac{4\ln(6D)}{3(1-\gamma)^2 T}. \tag{60}$$

For $\|\mathcal{T}_2\|_\infty$. We also apply (73) in Lemma F.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,

$$\mathbb{E}\|\mathcal{T}_2\|_\infty \le 6\sqrt{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}\sqrt{\frac{\ln(2D)}{T}}\sqrt{\frac{1}{T}\sum_{j=1}^{T}\|\boldsymbol{A}_j^T-\boldsymbol{G}^{-1}\|_\infty^2 + \frac{8C_0\ln(6D)}{3(1-\gamma)T}}. \tag{61}$$

For $\|\mathcal{T}_3\|_\infty$. We apply (74) in Lemma F.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 D.4.2.

Lemma D.1.

$$\mathbb{E}\|\mathcal{T}_3\|_\infty \le 4\gamma C_0\sqrt{\frac{\ln(2DT^2)}{T}}\cdot\sqrt{\frac{1}{T}\sum_{j=1}^{T}\mathbb{E}\,\|\boldsymbol{\Delta}_{j-1}\|_\infty^2 + \frac{32\gamma C_0\ln(3DT^2)}{3(1-\gamma)T}}. \tag{62}$$

where C_0 is the uniform bound given in Lemma B.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 B.7. It follows that

$$\mathbb{E}\|\mathcal{T}_4\|_\infty = \frac{1}{\sqrt{T}}\mathbb{E}\|\boldsymbol{\psi}_5(1)\|_\infty \le \frac{1}{\sqrt{T}}\mathbb{E}\|\boldsymbol{\psi}_5\|_{\sup} \le \gamma L C_0\cdot\frac{1}{T}\sum_{j=1}^{T}\mathbb{E}\,\|\boldsymbol{\Delta}_{j-1}\|_\infty^2 \tag{63}$$

with L defined in Lemma B.6.

Remark D.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 B.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 A.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.*

D.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 B.3 and Lemma B.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 A.2 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 (59), (60), (61), (62), and (63), we have

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

where $\widetilde{\mathcal{O}}(\cdot)$ hides polynomial dependence on logarithmic terms namely $\ln D$ and $\ln T$. Here we use $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty \le \frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}$ 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 B.3 and Lemma B.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 A.2 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 (59), (60), (61), (62), and (63), we have

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

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.

D.4 Proofs of lemmas

D.4.1 A useful inequality

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

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

$$\|\mathrm{diag}(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^\top)\|_\infty \le \|\boldsymbol{V}\|_{\max}\|\boldsymbol{A}\|_\infty^2, \tag{64}$$

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

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

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

□

D.4.2 Proof of Lemma D.1

Proof of Lemma D.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 F.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, $\{\boldsymbol{X}_j\}_{j\geq 0}$ 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 $\boldsymbol{U}_j = \mathrm{Var}[\boldsymbol{X}_j|\mathcal{F}_{j-1}]$.[7]

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

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

Here the last inequality uses (64). 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,

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

where (a) again uses (64) and (b) uses $\|\boldsymbol{P}_j - \boldsymbol{P}\|_\infty \leq \|\boldsymbol{P}_j\|_\infty + \|\boldsymbol{P}\|_\infty = 2$.

Putting the pieces together, we have

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

where we use $\sup_j\|\boldsymbol{B}_j\|_\infty \leq B = \frac{C_0}{1-\gamma}$. The rest follows from (74) in Lemma F.1 by plugging the corresponding B, X, D and σ^2 and the inequality $\|\boldsymbol{V}_{j-1} - \boldsymbol{V}^*\|_\infty \leq \|\boldsymbol{Q}_{j-1} - \boldsymbol{Q}^*\|_\infty = \|\boldsymbol{\Delta}_{j-1}\|_\infty$. □

E Proof of Information-Theoretic Lower Bound

E.1 Proof of Theorem 3.3

The semiparametric model $\mathcal{P}_\theta \in \mathcal{P}_P \times \mathcal{P}_R$ described in Section 3.3 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

[7] 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.

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

$$\sup_{\mathcal{P}_\gamma \subset \mathcal{P}} \boldsymbol{\Gamma}(\gamma_0)\boldsymbol{I}(\gamma_0)^{-1}\boldsymbol{\Gamma}^\top(\gamma_0), \tag{65}$$

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, (65) 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 $\boldsymbol{P}_j$.

E.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

$$h_{\tilde{s},\tilde{a}}(\boldsymbol{P}) = \sum_{s,a,s'} P(s'|s,a)1_{\{(s,a)=(\tilde{s},\tilde{a})\}} - 1. \tag{66}$$

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

$$\boldsymbol{\Gamma}(\boldsymbol{P})\boldsymbol{I}(\boldsymbol{P})^{-1}\boldsymbol{\Gamma}^\top(\boldsymbol{P}) = \left(\frac{\partial \boldsymbol{Q}^*}{\partial \boldsymbol{P}}\right)^\top \boldsymbol{C}_T(\boldsymbol{P}) \frac{\partial \boldsymbol{Q}^*}{\partial \boldsymbol{P}} = \frac{1}{T} \cdot \left(\frac{\partial \boldsymbol{Q}^*}{\partial \boldsymbol{P}}\right)^\top \boldsymbol{C}_1(\boldsymbol{P}) \frac{\partial \boldsymbol{Q}^*}{\partial \boldsymbol{P}}, \tag{67}$$

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 (67), we first compute $\frac{\partial \boldsymbol{Q}^*}{\partial \boldsymbol{P}}$.

Lemma E.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 E.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*

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

with $\boldsymbol{h}$ defined in (66). 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 E.1 and E.2, we have

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

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 E.1 and E.2.

Proof of Lemma E.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 E.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

$$\frac{\partial \log f_{\boldsymbol{P}}(\boldsymbol{P}_t)}{\partial P(s'|s,a)} = \frac{1_{\{s_t(s,a)=s'\}}}{P(s'|s,a)}. \tag{68}$$

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 $\boldsymbol{H}(\boldsymbol{P}) = \mathrm{diag}(\{\mathbf{1}_S\}_{(s,a)})$, where $\mathbf{1}_S$ is an all-1 vector with dimension S, and $\boldsymbol{U}(\boldsymbol{P}) = \mathrm{diag}(\{\boldsymbol{U}_{(s,a)}\}_{(s,a)})$, 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

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

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 $\mathrm{diag}(\boldsymbol{P}_{s,a}) - \boldsymbol{P}_{s,a}\boldsymbol{P}_{s,a}^\top$. Therefore,

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

□

E.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 E.3 shows it is still true in our case.

Lemma E.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

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

Proof of Lemma E.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 $\mathcal{P}_{R_\varepsilon} = \{R_\varepsilon : \varepsilon \in [-\delta,\delta]$ 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),

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

$$= \frac{\partial \mathbb{E} r_t(s,a)}{\partial \varepsilon}\bigg|_{\varepsilon=0} + \gamma \sum_{s'} P(s'|s,a) \frac{\partial Q^*(s', \pi^*(s'))}{\partial \varepsilon}\bigg|_{\varepsilon=0}.$$

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

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

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

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

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,

$$\frac{\partial Q^*(\tilde{s}, \tilde{a})}{\partial \varepsilon}\bigg|_{\varepsilon=0} = \int \phi(\tilde{s}, \tilde{a}) \frac{\partial}{\partial \varepsilon} \log dF_\varepsilon \bigg|_{\varepsilon=0} dF_0, \tag{69}$$

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). □

E.2 Proof of Theorem 3.4

Proof of Theorem 3.4. Recall that $\bar{\boldsymbol{\Delta}}_T = \frac{1}{T} \sum_{t=1}^T (\boldsymbol{Q}_T - \boldsymbol{Q}^*)$. Combining (55), (56) and (57), we have

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

where the inequality holds coordinate-wise. In Appendix D.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 D.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

$$\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t) S_\gamma^\top(\gamma_0) = \left. \frac{\partial \boldsymbol{Q}^*}{\partial \gamma} \right|_{\gamma = \gamma_0}, \tag{70}$$

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 (68) 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 E.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

$$\gamma (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}(\cdot, (\tilde{s}, \tilde{a})) \left[V^*(s') - V^*(s_0) \right]. \tag{71}$$

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 (71). 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

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

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

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

$$= \left(V^*(s') - \sum_{i \neq s_0} P(i|s,a) V^*(i) \right) + \sum_{i \in \mathcal{S}} P(i|s,a) V^*(i) = V^*(s') - V^*(s_0).$$

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 (70) holds for the $\boldsymbol{P}$ part.

For the random reward R. Using the notation in the proof of Lemma E.3, $S_\varepsilon(\gamma_0) = \frac{\partial}{\partial \varepsilon} \log dF_\varepsilon|_{\varepsilon=0}$. By (69), 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 (70) holds for the ε part.

$\mathcal{P}_{R_\varepsilon}$ can be arbitrary, so (70) 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. □

F 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 F.1. *Assume $\{\boldsymbol{X}_j\} \subseteq \mathbb{R}^d$ are martingale differences adapted to the filtration $\{\mathcal{F}_j\}_{j \geq 0}$ 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 $\{\boldsymbol{X}_j\}_{j \geq 0}$ is uniformly bounded, i.e., $\sup_j \|\boldsymbol{X}_j\|_\infty \leq X$. For any sequence of deterministic matrices $\{\boldsymbol{B}_j\}_{j \geq 0} \subseteq \mathbb{R}^{D \times d}$ satisfying $\sup_j \|\boldsymbol{B}_j\|_\infty \leq B$, we define the weighted sum as*

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

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

$$\mathbb{P}\left(\|\boldsymbol{Y}_T\|_\infty \geq \frac{2BX}{3} \ln \frac{2D}{\delta} + \sqrt{2\sigma^2 \ln \frac{2D}{\delta}} \text{ and } \|\boldsymbol{W}_T\|_\infty \leq \sigma^2 \right) \leq \delta \tag{72}$$

$$\mathbb{E}\|\boldsymbol{Y}_T\|_\infty 1_{\{\|\boldsymbol{W}_T\|_\infty \leq \sigma^2\}} \leq 6\sigma \sqrt{\ln(2D)} + \frac{4BX}{3} \ln(6D). \tag{73}$$

Generally, we have

$$\mathbb{E}\|\boldsymbol{Y}_T\|_\infty \leq \frac{8BX}{3} \ln(3DT^2) + 2\sqrt{\mathbb{E}\|\boldsymbol{W}_T\|_\infty} \sqrt{\ln(2DT^2)}. \tag{74}$$

Proof of Lemma F.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

$$\begin{aligned}
\mathbb{P}(\|\boldsymbol{Y}_T\|_\infty \ge \tau \text{ and } \|\boldsymbol{W}_T\|_\infty \le \sigma^2) &= \mathbb{P}\left(\max_{i\in[D]} |\boldsymbol{Y}_T(i)| \ge \tau \text{ and } \max_{i\in[D]} |\boldsymbol{W}_T(i,i)| \le \sigma^2\right) \\
&\le \sum_{i\in[D]} \mathbb{P}\left(|\boldsymbol{Y}_T(i)| \ge \tau \text{ and } \max_{i\in[D]} |\boldsymbol{W}_T(i,i)| \le \sigma^2\right) \\
&\le \sum_{i\in[D]} \mathbb{P}\left(|\boldsymbol{Y}_T(i)| \ge \tau \text{ and } |\boldsymbol{W}_T(i,i)| \le \sigma^2\right) \\
&\le 2D\exp\left(-\frac{\tau^2/2}{\sigma^2 + BX\tau/3}\right). \qquad (75)
\end{aligned}$$

Solving for τ such that the right-hand side of (75) 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 (75). 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

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

where (a) uses Lemma D.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 (73), 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

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

and claim that we have $\mathbb{P}(\mathcal{H}_K) \leq \delta$. 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 $0 \leq k < K$, $\mathcal{B}_k$ is defined to be

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

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

Invoking (72) with a proper $\sigma^2 = \frac{TB^2X^2}{2^{k-1}}$ and $\delta = \frac{\delta}{K}$, we have $\mathbb{P}(\mathcal{B}_k) \leq \frac{\delta}{K}$ for all $k \in [K]$. Taken this result together with the union bound gives $\mathbb{P}(\mathcal{H}_K) \leq \sum_{k\in[K]} \mathbb{P}(\mathcal{B}_k) \leq \delta$. Then we have

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

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

G Details of Experiment

According to Theorem 4.1, for sufficiently small error $\varepsilon > 0$, we expect the sample complexity $T(\varepsilon, \gamma)$ is always upper bounded by $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty$ 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 1, we choose $\varepsilon = e^{-4}$ and plot the results on a log-log scale. We have also plotted the least-squares fits through these points and the slopes of these lines are also provided in the legend. In particular, we run linear regression of pairs $\left(\log \frac{1}{1-\gamma_i}, \log T(\varepsilon, \gamma_i)\right)$ and obtain the coefficient k on $\log \frac{1}{1-\gamma}$ in the legend of Figure 1.