Finite-Time Analysis of Temporal Difference Learning with Experience Replay

Anonymous authors
Paper under double-blind review

Abstract

Temporal-difference (TD) learning is widely regarded as one of the most popular algorithms in reinforcement learning (RL). Despite its widespread use, it has only been recently that researchers have begun to actively study its finite time behavior, including the finite time bound on mean squared error and sample complexity. On the empirical side, experience replay has been a key ingredient in the success of deep RL algorithms, but its theoretical effects on RL have yet to be fully understood. In this paper, we present a simple decomposition of the Markovian noise terms and provide finite-time error bounds for tabular on-policy TD-learning with experience replay. Specifically, under the Markovian observation model, we demonstrate that for both the averaged iterate and final iterate cases, the error term induced by a constant step-size can be effectively controlled by the size of the replay buffer and the mini-batch sampled from the experience replay buffer.

1 Introduction

The pioneering Deep Q-network (DQN) (Mnih et al., 2015) has demonstrated the vast potential of reinforcement learning (RL) algorithms, having achieved human-level performances in numerous Atari games (Bellemare et al., 2013). Such successes have fueled extensive research efforts in the development of RL algorithms, e.g., Sewak (2019); Mnih et al. (2016) to name just a few. Beyond video games, RL has showcased notable performances in various fields, including robotics (Singh et al., 2020) and finance (Liu et al., 2020).

On the other hand, temporal-difference (TD) learning (Sutton, 1988) is considered one of the most fundamental and well-known RL algorithms. Its objective is to learn the value function, which represents the expected sum of discounted rewards following a particular policy. While asymptotic convergence of TD-learning (Jaakkola et al., 1993; Bertsekas & Tsitsiklis, 1996) has been extensively studied and is now well-understood, such asymptotic analysis cannot measure how efficiently the estimation progresses towards a solution. Recently, the convergence rate of TD-learning has gained much attention and has been actively investigated (Bhandari et al., 2018; Srikant & Ying, 2019; Lee & Kim, 2022). These studies aim to understand the efficiency of the estimation process, and provide theoretical guarantees on the rate of convergence.

First appeared in Lin (1992), experience replay memory can be viewed as a first-in-first-out queue and is one of the principal pillars of DQN (Mnih et al., 2015). Learning through uniformly random samplings from the experience replay memory, the strategy is known to reduce correlations among experience samplings, and improve the efficiency of the learning. Despite its empirical successes (Mnih et al., 2015; Fedus et al., 2020; Hong et al., 2022), theoretical side of the experience replay memory techniques remains largely an open yet challenging question. Only recently, Di-Castro et al. (2022; 2021) studied asymptotic convergence of actor-critic algorithm (Konda & Tsitsiklis, 1999) with experience replay memory. To the authors' knowledge, its non-asymptotic analysis has not been thoroughly investigated in sense of TD-learning so far.

The aim of this paper is to investigate the impact of experience replay memory on TD-learning (Sutton, 1988), with the goal of shedding light on the question at hand. Our primary theoretical contribution is the derivation of the convergence rate of tabular on-policy TD-learning with experience replay memory under Markovian observation models. This analysis reveals connections between the sizes of the mini-batch and experience replay memory, and the convergence rate. Specifically, we show that the error term, resulting from correlations among the samples from the Markovian observation models, can be effectively controlled

by the sizes of the mini-batch and experience replay memory, for both the averaged and final iterate cases. We expect that our findings can provide further insights into the use of experience replay memory in RL algorithms.

Lastly, although our analysis only considers the standard TD-learning case, our presented arguments can be extended to more general scenarios, such as TD-learning with linear function approximation (Srikant & Ying, 2019), standard Q-learning (Watkins & Dayan, 1992), periodic Q-learning (Lee & He, 2019).

1.1 Related works

Experience replay. We begin by providing an overview of the existing theoretical results on experience replay. The recent work by Nagaraj et al. (2020) leverages the experience replay memory to address the least-squares problem under the Gaussian auto-regressive model. However, there are several notable differences between their approach and the proposed TD-learning with experience replay memory:

1. The approach in Nagaraj et al. (2020) assumes i.i.d. Gaussian noises, whereas the proposed TD-learning with experience replay memory covers Markovian and specific non-Gaussian noises.
2. The overall algorithmic structures are significantly different. Nagaraj et al. (2020) considered an offline learning problem, while the proposed TD-learning framework is an online learning approach.
3. When operating the experience replay buffer, they maintain a sufficiently large gap between the separate samples inside the buffer to ensure the samples are almost identical and independently distributed.
4. The approach by Nagaraj et al. (2020) uses an inner loop to iterate over the samples in the buffer, whereas the proposed framework considers mini-batch style updates, which is widely used in practice.

Kowshik et al. (2021) presented an algorithm that employs the reverse experience memory approach proposed in Rotinov (2019) to tackle linear system identification problems. This algorithm guarantees non-asymptotic convergence, and the reverse experience replay technique (Rotinov, 2019) involves using transitions in reverse order, rather than uniformly and randomly, without introducing any stochasticity, which distinguishes it from the original experience replay method introduced in Mnih et al. (2015). Kowshik et al. (2021) assumes auto-regressive model with i.i.d. noise. Following the spirit of Nagaraj et al. (2020), Kowshik et al. (2021) also maintains a gap between separate samples to enforce independence between the samples.

Using the super-martingale structure of the reverse experience replay memory, Agarwal et al. (2021) derives sample complexity of Q-learning, which also maintains a gap between samples. Moreover, the works, Kowshik et al. (2021) and Agarwal et al. (2021) both exploit full samples from each experience replay memory rather than applying uniformly sampled mini-batch to maintain the reverse order property.

Lazic et al. (2021) presented a regret bound analysis for regularized policy iteration with a replay buffer in the context of averaged reward Markov decision processes (MDPs). The authors assume that an accurate estimate of the action-value function is available, which is obtained via Monte Carlo methods (Singh & Sutton, 1996), as opposed to the TD-learning algorithm (Sutton, 1988). The use of experience replay memory is modified from Mnih et al. (2015). One approach suggested in Lazic et al. (2021) involves storing all data from each phase, with a limit on the size of the replay buffer. When the buffer size exceeds the limit, a subset of the data is eliminated uniformly at random.

Fan et al. (2020) established a non-asymptotic convergence of fitted Q-learning (Ernst et al., 2005), which assumes i.i.d. sampling of transitions from a fixed distribution. This assumption is stronger compared to the Markovian observation models used with the time-varying replay buffer. Markovian observation model is a more natural and realistic scenario than the i.i.d. observation model, which improves practicality of the analysis. Furthermore, The replay buffer technique is different from the sampling i.i.d. transitions because in the mini-batch approach, the samples are from the single episode trajectory in a moving window, which are highly correlated, and the correlation increases proportionally to the size of the window. Lastly, Fan et al. (2020) imposed a condition that the fixed distribution satisfies particular condition such that it is similar to the true underlying distribution over the state-action space, which may not hold in our setting. Moreover, the fitted Q-iteration is similar to stochastic gradient descent algorithm in supervised-learning

problem, whereas TD-learning an online learning algorithm which does not use any (stochastic) gradient of an objective function.

A closely related approach to our work is presented in Di-Castro et al. (2022), which establishes the asymptotic convergence of the actor-critic algorithm (Konda & Tsitsiklis, 1999) using a mini-batch that is uniformly and randomly sampled from the replay buffer. The proof relies on treating the replay buffer and mini-batch indices as random variables, which together form an irreducible and aperiodic Markov chain. Following the O.D.E. approach outlined in Borkar & Meyn (2000) and the natural actor-critic algorithm described in Bhatnagar et al. (2009), Di-Castro et al. (2022) establishs the convergence of the actor-critic algorithm using the replay buffer. While Di-Castro et al. (2022) demonstrates a decrease in auto-correlation and covariance among samples in the experience replay, their proof of asymptotic convergence does not explicitly address the impact of these factors on the convergence behavior. However, the study provides important insights into the convergence properties of actor-critic algorithms using the replay buffer and highlights the potential benefits of using mini-batches to improve the convergence rate.

Non-asymptotic analysis of TD-learning We highlight several recent studies on finite time behavior of TD-learning (Lee & Kim, 2022; Bhandari et al., 2018; Chen et al., 2020; Srikant & Ying, 2019; Dalal et al., 2018; Hu & Syed, 2019). Under i.i.d. observation model and linear function approximation setting, Dalal et al. (2018) derived $O\left(\frac{1}{k^\sigma}\right)$ bound on the mean squared error with diminishing step-size $\frac{1}{(k+1)^\sigma}$ where $\sigma \in (0,1)$ and k is the number of iterations. Under i.i.d. observation model and tabular setup, from the discrete-time stochastic linear system perspective, Lee & Kim (2022) provided geometric convergence rate with constant error at the order of $O(\alpha)$ of TD-learning for both averaged iterate and final iterate using constant step-size α. Bhandari et al. (2018) provided convergence rate of TD-learning with linear function approximation under the Markovian noise following the spirit of convex optimization literature (Nemirovski et al., 2009). With the help of Moreau envelope (Parikh et al., 2014), Chen et al. (2020) derived convergence rate of TD-learning with respect to an arbitrary norm. Mainly based on Lyapunov approach (Khalil, 2015) for continuous time O.D.E. counterpart of TD-learning, Srikant & Ying (2019) derived a finite time bound on the mean squared error of TD-learning under Markovian noise with linear function approximation.

In contrast to the existing studies that have focused on the finite time behavior of TD-learning (Lee & Kim, 2022; Bhandari et al., 2018; Chen et al., 2020; Srikant & Ying, 2019; Dalal et al., 2018; Hu & Syed, 2019), our work investigates the finite time behavior of TD-learning with experience replay, which has not been thoroughly studied to date. Specifically, we demonstrate that the use of experience replay can be an effective means of reducing the constant error term that arises from employing a constant step-size. Our findings may provide valuable insights into the effectiveness of experience replay in RL, shedding new light on the benefits of this widely-used technique. Further research in this area could yield important research topics, with implications for the development of more efficient and effective reinforcement learning algorithms.

2 Preliminaries

2.1 Markov chain

In this section, basic concepts of Markov chain are briefly introduced. To begin with, the so-called total variation distance defines distance between two probability measures as follows.

Definition 2.1 (Total variation distance (Levin & Peres, 2017))**.** *The total variation distance between two probability distributions, μ_1 and μ_2, on $\mathcal{S}$ is given by*

$$d_{\mathrm{TV}}(\mu_1, \mu_2) = \sup_{A \subseteq \mathcal{S}} |\mu_1(A) - \mu_2(A)|.$$

Let us consider a Markov chain with the set of states $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$ and the state transition probability $\mathcal{P}$. For instance, a state $s \in \mathcal{S}$ transits to the next state s' with probability $\mathcal{P}(s, s')$. A stationary distribution of the Markov chain is defined as a distribution $\mu \in \mathbb{R}^{|\mathcal{S}|}$ on $\mathcal{S}$ such that $\mu^\top P = \mu^\top$ where $P \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the transition matrix of Markov chain, i.e., $[P]_{ij} = \mathcal{P}(i, j)$ for $i, j \in \mathcal{S}$.

Let $\{S_k\}_{k \geq 0}$ be a trajectory of a Markov chain. Then, an irreducible and aperiodic Markov chain is known to admit a unique stationary distribution μ such that the total variation distance between the stationary

distribution and the current state distribution decreases exponentially (Levin & Peres, 2017) as follows:

$$\sup_{s\in\mathcal{S}} d_{\mathrm{TV}}(\mathbb{P}[S_k = \cdot \mid S_0 = s], \mu) \leq m\rho^k$$

for some $\rho \in (0,1)$ and positive $m \in \mathbb{R}$. The mixing time of a Markov chain is defined as

$$t^{\mathrm{mix}}(\epsilon) := \min\left\{k \in \mathbb{Z}_+ : \sup_{s\in\mathcal{S}} d_{\mathrm{TV}}(\mathbb{P}[S_k = \cdot \mid S_0 = s], \mu) \leq \epsilon\right\} \tag{1}$$

for any $\epsilon \in \mathbb{R}_+$. Throughout the paper, we will use t^{mix} to denote $t^{\mathrm{mix}}\left(\frac{1}{4}\right)$ for simplicity.

2.2 Markov decision process

A Markov decision process is described by the tuple $(\mathcal{S}, \mathcal{A}, \gamma, \mathcal{P}, r)$, where $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$ is the set of states, $\mathcal{A} := \{1, 2, \ldots, |\mathcal{A}|\}$ is the set of actions, $\gamma \in (0,1)$ is the discount factor, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the state transition probability, i.e., $\mathcal{P}(s,a,s')$ means the probability of the next state $s' \in \mathcal{S}$ when taking action $a \in \mathcal{A}$ at the current state $s \in \mathcal{S}$. For example, at state $s_k \in \mathcal{S}$ at time k, if an agent selects an action a_k, then the state transits to the next state s_{k+1} with probability $\mathcal{P}(s_k, a_k, s_{k+1})$, and incurs the reward $r(s_k, a_k, s_{k+1})$, where the reward generated by the action at time k, $r(s_k, a_k, s_{k+1})$, will be denoted by $r_{k+1} := r(s_k, a_k, s_{k+1})$. In this paper, we adopt the standard assumption on the boundedness of the reward function.

Assumption 2.2. *There exists a positive $R_{\max} \in \mathbb{R}$ such that $|r(s,a,s')| \leq R_{\max}$ for all $s, a, s' \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.*

Let us consider a Markov decision process with the policy, $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$. The corresponding state trajectory, $\{S_k\}_{k\geq 0}$, is a Markov chain induced by the policy π, and the state transition probability is given by $\mathcal{P}^\pi : \mathcal{S} \times \mathcal{S} \to [0,1]$, i.e., $\mathcal{P}^\pi(s,s') := \sum_{a\in\mathcal{A}} \mathcal{P}(s,a,s')\pi(a \mid s)$ for $s, s' \in \mathcal{S}$. Throughout the paper, we assume that the induced Markov chain with transition kernel $\mathcal{P}^\pi$ is irreducible and aperiodic so that it admits a unique stationary distribution denoted by $\mu^\pi_{S_\infty}$, and satisfies the exponential convergence property

$$\sup_{s\in\mathcal{S}} d_{\mathrm{TV}}(\mathbb{P}[S_k = \cdot \mid S_0 = s], \mu^\pi_{S_\infty}) \leq m_1\rho_1^k, \quad k \geq 0,$$

for some positive $m_1 \in \mathbb{R}$ and $\rho_1 \in (0,1)$. Let $(S_k, S_{k+1}) \in \mathcal{O}$ be a tuple of states at time step k and its next state $S_{k+1} \sim \mathcal{P}^\pi(S_k, \cdot)$, which will be frequently used in this paper to analyze TD-learning, and $\mathcal{O}$ denotes the realizable set of tuples consisting of state and the next state, i.e., for $(x,y) \in \mathcal{S} \times \mathcal{S}$, we have $(x,y) \in \mathcal{O}$ if and only if $\mathcal{P}^\pi(x,y) > 0$. Then, the tuple forms another induced Markov chain. The transition probability of the induced Markov chain $\{(S_k, S_{k+1})\}_{k\geq 0}$ is $\mathbb{P}[S_{k+2}, S_{k+1} \mid S_{k+1}, S_k] = \mathbb{P}[S_{k+2} \mid S_{k+1}]$, which follows from the Markov property. The next lemma states that the Markov chain $\{(S_k, S_{k+1})\}_{k\geq 0}$ is also irreducible and aperiodic provided that $\{S_k\}_{k\geq 0}$ is irreducible and aperiodic.

Lemma 2.3. *If $\{S_k\}_{k\geq 0}$ is an irreducible and aperiodic Markov chain, then so is $\{(S_k, S_{k+1})\}_{k\geq 0}$.*

The proof is in Appendix Section A.3. Now, let us denote $\mu^\pi_{S_\infty, S'_\infty}$ as the stationary distribution of the Markov chain $\{(S_k, S_{k+1})\}_{k\geq 0}$, which satisfies the relation between $\mu^\pi_{S_\infty}$ and $\mu^\pi_{S_\infty, S'_\infty}$:

$$\sum_{a\in\mathcal{A}} \pi(a \mid s)\mathcal{P}(s,a,s')\mu^\pi_{S_\infty}(s) = \mathcal{P}^\pi(s,s')\mu^\pi_{S_\infty}(s) = \mu^\pi_{S_\infty, S'_\infty}(s,s'). \tag{2}$$

Then, from Lemma 2.3, we have

$$\sup_{(s,s')\in\mathcal{S}\times\mathcal{S}} d_{\mathrm{TV}}\left(\mathbb{P}[(S_k, S_{k+1}) = \cdot \mid (S_0, S_1) = (s,s')], \mu^\pi_{S_\infty, S'_\infty}\right) \leq m_2\rho_2^k$$

for some positive $m_2 \in \mathbb{R}$ and $\rho_2 \in (0,1)$. Similar to the original Markov chain, the mixing time can be defined for this new Markov chain. Throughout this paper, we adopt the notations, t_1^{mix} and t_2^{mix}, to denote the mixing time of the Markov chain $\{S_k\}_{k\geq 0}$ and the mixing time of $\{(S_k, S_{k+1})\}_{k\geq 0}$, respectively. For simplicity of the notation, we will denote $\tau^{\mathrm{mix}} := \max\{t_1^{\mathrm{mix}}, t_2^{\mathrm{mix}}\}$.

2.3 Temporal difference learning

To begin with, several matrix notations are introduced. Let us define

$$D^\pi := \begin{bmatrix} \mu^\pi_{S_\infty}(1) & & & \\ & \mu^\pi_{S_\infty}(2) & & \\ & & \ddots & \\ & & & \mu^\pi_{S_\infty}(|\mathcal{S}|) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, \quad R^\pi = \begin{bmatrix} \mathbb{E}[r(s,a,s')|s=1] \\ \mathbb{E}[r(s,a,s')|s=2] \\ \vdots \\ \mathbb{E}[r(s,a,s')|s=|\mathcal{S}|] \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|},$$

where R^π defined above is a vector of expected rewards when action is taken under π. From Assumption 2.2, one can readily prove that R^π is bounded as well.

Lemma 2.4. *We have* $||R^\pi||_\infty \leq R_{\max}$.

Moreover, in this paper, $P^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denotes the transition matrix under the policy π, i.e., $[P^\pi]_{ij} = \mathcal{P}^\pi(i,j)$ for $i,j \in \mathcal{S}$, where $[P^\pi]_{ij}$ denotes the element in the i-th column and j-th row. The minimum probability in the stationary state distribution is denoted by $\mu^\pi_{\min} := \min_{s\in\mathcal{S}} \mu^\pi_{S_\infty}(s)$.

Proposed in Sutton (1988), TD-learning aims to estimate the value function $V^\pi(s) := \sum_{k=0}^\infty \mathbb{E}\left[\gamma^k r_k \mid S_0 = s, \pi\right]$, $s \in \mathcal{S}$ through the following stochastic recursion for $k \geq 0$:

$$V_{k+1} = V_k + \alpha\delta(O_k, V_k), \tag{3}$$

where $\alpha \in (0,1)$ is a constant step-size, $O_k := (s_k, r_{k+1}, s_{k+1})$, and δ is called the TD-error defined as

$$\delta(O_k, V_k) := e_{s_k} r_k + \gamma e_{s_k} e_{s_{k+1}}^\top V_k - e_{s_k} e_{s_k}^\top V_k, \tag{4}$$

where e_{s_k} is the basis vector in $\mathbb{R}^{|\mathcal{S}|}$ whose s_k-th element is one and all others are zero.

3 Main results

3.1 TD-learning with experience replay

In this subsection, we introduce the proposed TD-learning method employing the original experience replay memory technique from Mnih et al. (2015) with no significant modifications. Experience replay memory, commonly referred to as the replay buffer, is a first-in-first-out (FIFO) queue that facilitates the adoption of mini-batch techniques in machine learning for online learning scenarios. Specifically, the replay buffer stores the state-action-reward transitions in a FIFO manner, serving as the training set. At each step, a mini-batch is uniformly sampled from the replay buffer and utilized to update the learning parameter V_k. This approach presents dual advantages. Firstly, it enables the application of the batch update scheme, thereby reducing variance and accelerating learning. Secondly, through uniform sampling, it may reduce correlations among different samples in the Markovian observation models and consequently reduce extra biases in the value estimation. Despite its empirical benefits, theoretical studies on replay buffer have been limited.

In this paper, we employ the notation B^π_k and M^π_k to represent the replay buffer and mini-batch, respectively, at time step k. They are formally defined as follows:

$$M^\pi_k := \{O^k_1, O^k_2 \ldots, O^k_L\}, \quad B^\pi_k := \{O_{k-N+1}, O_{k-N+2} \ldots, O_k\},$$

where the replay buffer's size is N, the mini-batch's size is L, and $O^k_i, 1 \leq i \leq L$ stands for the i-th sample of the mini-batch M^π_k. For the first N steps, only the observations, O_{-N} from O_{-1} , is collected without updating the V_k. That is, at time-step k, O_k is observed and V_k is updated using a mini-batch M^π_k sampled from batch B^π_k. The overall scheme is depicted in Figure 1.

For each time step $k \geq 0$, the agent selects an action a_k following the target policy π, and observes next state s_{k+1}. The oldest sample $(s_{k-N}, a_{k-N}, s_{k-N+1})$ is dropped from the replay buffer B^π_{k-1}, and the new sample (s_k, a_k, s_{k+1}) is added to the replay buffer, which becomes B^π_k.

Next, a mini-batch of size L, M^π_k, is sampled uniformly from the replay buffer B^π_k at time step k, and the $k+1$-th iterate of TD-learning is updated via

$$V_{k+1} = V_k + \alpha_k \frac{1}{|M^\pi_k|} \sum_{(s,r,s')\in M^\pi_k} (e_s)(r + \gamma e_{s'}^\top V_k - e_s^\top V_k), \tag{5}$$

which is a batch update version of the standard TD-learning. Note that when $|B_k^\pi| = |M_k^\pi| = 1$ for all $k \geq 0$, the update in (5) matches that of the standard TD-learning in (3). In this paper, we consider a constant step-size $\alpha \in (0, 1)$, and the Markovian observation model, which means that transition samples are obtained from a single trajectory of the underlying Markov decision process.

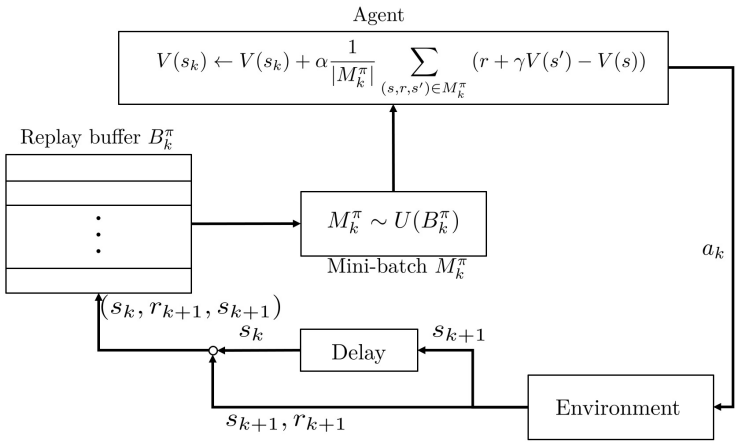

Figure 1: Diagram of TD-learning using experience replay

Algorithm 1 TD-learning with replay buffer

1: Initialize $V_0 \in \mathbb{R}^{|\mathcal{S}|}$ such that $||V_0||_\infty \leq \frac{R_{\max}}{1-\gamma}$.
2: Collect N samples : $B_{-1} := \{O_{-N}, O_{-N+1}, \ldots, O_{-1}\}$.
3: **for** $k \leq T$ **do**
4: Select action $a_k \sim \pi(\cdot|s_k)$.
5: Observe $s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$ and recieve reward $r_{k+1} := r(s_k, a_k, s_{k+1})$.
6: Update replay buffer : $B_k^\pi \leftarrow B_{k-1}^\pi \setminus \{(s_{k-N}, r_{k-N+1}, s_{k-N+1})\} \cup \{(s_k, r_{k+1}, s_{k+1})\}$.
7: Uniformly sample $M_k^\pi \sim B_k^\pi$.
8: Update $V_{k+1} = V_k + \alpha_k \frac{1}{|M_k^\pi|} \sum_{(s,r,s') \in M_k^\pi} (e_s)(r + \gamma e_{s'}^\top V_k - e_s^\top V_k)$.
9: **end for**

To proceed with our analysis, we should establish the boundedness of the iterate resulting from the update in (5), assuming that $||V_0||_\infty \leq \frac{R_{\max}}{1-\gamma}$ and that $\alpha_k \in (0, 1)$. This assumption is crucial to our main developments, and thus requires rigorous proof.

Lemma 3.1. *Under the recursion in (5), V_k remains bounded :* $||V_k||_\infty \leq \frac{R_{\max}}{1-\gamma}$.

The proof is given in Appendix Section A.4.

3.2 Analysis framework

In the previous subsection, the algorithm underlying our analysis was introduced. In this subsection, we provide the preliminary frameworks for our main analysis. Specifically, we analyze TD-learning based partially on the linear dynamical system viewpoint, as presented in the recent work by Lee & Kim (2022). In particular, using the Bellman equation (Bertsekas & Tsitsiklis, 1996), $D^\pi V^\pi = \gamma D^\pi P^\pi V^\pi + D^\pi R^\pi$, we can express the TD-learning update (5) as follows:

$$V_{k+1} - V^\pi := A(V_k - V^\pi) + \alpha w(M_k^\pi, V_k), \tag{6}$$

where

$$A := I - \alpha D^\pi + \alpha \gamma D^\pi P^\pi \tag{7}$$

is the system matrix, and

$$w(M_k^\pi, V_k) := \frac{1}{|M_k^\pi|} \sum_{(s,r,s')\in M_k^\pi} (e_s)(r + \gamma e_{s'}^\top V_k - e_s^\top V_k) - D^\pi(R^\pi + \gamma P^\pi V_k - V_k) \tag{8}$$

is the noise, which is the difference between the empirical mean of the TD-error via samples in mini-batch and the expected TD-error with respect to the stationary distribution $\mu_{S_\infty}^\pi$.

Some useful properties of the system matrix A (boundedness of A, and related Lyapunov theory (Chen, 1984)) are introduced in the next lemma, which play central roles in establishing the convergence rate.

Lemma 3.2 (Properties of matrix A (Lee & Kim, 2022)). *1) $||A||_\infty \leq 1 - \alpha(1-\gamma)\mu_{\min}^\pi$ holds.*

2) There exists a positive definite matrix $M \succ 0$ and $||M||_2 \leq \frac{2|\mathcal{S}|}{\alpha(1-\gamma)\mu_{\min}^\pi}$ such that

$$A^\top M A - M = -I. \tag{9}$$

For completeness of presentation, the proof is given in Appendix A.5. To proceed, let us define the empirical distributions for $i, j \in \mathcal{S}$,

$$\mu_S^{B_k^\pi}(i) := \frac{\sum_{(s,r,s')\in B_k^\pi} 1\{s = i\}}{|B_k^\pi|}, \quad \mu_{S,S'}^{B_k^\pi}(i,j) := \frac{\sum_{(s,r,s')\in B_k^\pi} 1\{(s,s') = (i,j)\}}{|B_k^\pi|},$$

which denote the empirical distributions of the state s and the tuple (s, s') in the replay buffer B_k^π, respectively. Moreover, for an event A, the notation $1\{A\}$ denotes an indicator function that returns one if the event A occurs and otherwise zero. Furthermore, let us introduce the following matrix notations for $i, j \in \mathcal{S}$:

$$[D^{B_k^\pi}]_{ij} := \begin{cases} \mu_S^{B_k^\pi}(i) & \text{if} \quad i = j \\ 0 & \text{otherwise} \end{cases}, \quad [P^{B_k^\pi}]_{ij} := \begin{cases} \dfrac{\sum_{(s,r,s')\in B_k^\pi} 1\{(s,s')=(i,j)\}}{\sum_{(s,r,s')\in B_k^\pi} 1\{s=i\}} & \text{if} \quad |B^\pi(i)| \geq 1 \\ 0 & \text{otherwise} \end{cases},$$

which define the empirical distribution of the state in the replay buffer and $B_k^\pi(s) := \{(\tilde{s}, \tilde{r}, \tilde{s}') \in B_k^\pi : \tilde{s} = s\} \subseteq B^\pi$. With above definitions, we can readily verify the relation $[D^{B_k^\pi} P^{B_k^\pi}]_{ij} = \mu_{S,S'}^{B_k^\pi}(i,j)$ for $i, j \in \mathcal{S}$.

Likewise, let us define the empirical estimation of the expected return, which is averaged over the action, a, and next state, s'. It is calculated from the samples of the replay buffer as follows for $i \in \mathcal{S}$

$$[R^{B_k^\pi}]_i = \begin{cases} \dfrac{\sum_{(s,r,s')\in B_k^\pi} 1\{s=i\} r}{\sum_{(s,r,s')\in B_k^\pi} 1\{s=i\}} & \text{if} \quad |B_k^\pi(i)| \geq 1 \\ 0 & \text{otherwise} \end{cases}.$$

3.3 Bounds on noise

Our aim in this subsection is to bound the first and second moment of $\|w(M_k^\pi, V_k)\|_2$ where $w(M_k^\pi, V_k)$ is defined in (8), which will play important role in deriving the convergence rate. For simplicity, let us further define the functions, $\Delta_k : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ and $\Delta_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$, as follows:

$$\Delta_k(V) = D^{B_k^\pi} R^{B_k^\pi} + \gamma D^{B_k^\pi} P^{B_k^\pi} V - D^{B_k^\pi} V, \tag{10}$$

$$\Delta_\pi(V) = D^\pi R^\pi + \gamma D^\pi P^\pi V - D^\pi V. \tag{11}$$

The functions, Δ_k and Δ_π, can be viewed as expected TD-errors, respectively, in terms of the distribution of the replay buffer and the stationary distribution of the Markov chain. Moreover, note that $\Delta_\pi(V^\pi) = 0$. Based on the notations introduced, the noise term, $w(M_k^\pi, V_k)$, can be decomposed into the two parts

$$w(M_k^\pi, V_k) = \frac{1}{|M_k^\pi|} \sum_{i=1}^{|M_k^\pi|} \delta(O_i^k, V_k) - \Delta_k(V_k) - (\Delta_\pi(V_k) - \Delta_k(V_k)), \tag{12}$$

where

1. $\frac{1}{|M_k^\pi|}\sum_{i=1}^{|M_k^\pi|}\delta(O_i^k, V_k) - \Delta_k(V_k)$: the difference between the empirically expected TD-error with respect to the distribution of mini-batch and empirically expected TD-error with respect to the replay buffer.

2. $\Delta_\pi(V_k) - \Delta_k(V_k)$: the difference between the expected TD-error with respect to the stationary distribution and the empirically expected TD-error with respect to the distribution of replay buffer.

The high level idea for bounding the first and second moment of $\|w(M_k^\pi, V_k)\|_2$ is summarized below.

1. $\frac{1}{|M_k^\pi|}\sum_{i=1}^{|M_k^\pi|}\delta(O_i^k, V_k) - \Delta_k(V_k)$, the error between the empirical distribution of the mini-batch and the replay buffer, can be bounded by Bernstein inequality (Tropp et al., 2015) because the mini-batch samples are independently sampled from the replay buffer with the uniform distribution.

2. $\Delta_\pi(V_k) - \Delta_k(V_k)$, the error between the stationary distribution and distribution of replay buffer, can be bounded using the property of irreducible and aperiodic Markov chain (Levin & Peres, 2017).

Next, we introduce several lemmas to derive a bound on the the first moment of $\|w(M_k^\pi, V_k)\|_2$, which can be bounded at the order of $O\left(\sqrt{\frac{1}{|M_k^\pi|}}\right)$. To bound $\Delta_\pi(V_k) - \Delta_k(V_k)$, we introduce the coupled process $\{\tilde{S}_k\}_{k\geq -N}$, which starts from the stationary distribution of the Markov chain with transition kernel $\mathcal{P}^\pi$, i.e., $\tilde{S}_{-N} \sim \mu_{S_\infty}^\pi$. Let $\tilde{B}_k^\pi$ be the corresponding replay buffer of such a Markov chain. We first derive the expected error bounds in terms of the coupled process $\{\tilde{S}_k\}_{k\geq -N}$, and the desired result will be obtained using the total variation distance between the distribution of $\tilde{S}_k$ and S_k. The detailed proof is given in Appendix A.6. Now, combining with the bound on $\frac{1}{|M_k^\pi|}\sum_{i=1}^{|M_k^\pi|}\delta(O_i^k, V_k) - \Delta_k(V_k)$, which follows from concentration inequalities, we get the following result:

Lemma 3.3. *For $k \geq 0$, $\mathbb{E}[||w(M_k^\pi, V_k)||_2]$ can be bounded as follows:*

$$\mathbb{E}[||w(M_k^\pi, V_k)||_2] \leq \frac{4\sqrt{|\mathcal{S}|}R_{\max}}{1-\gamma}\left(2\sqrt{\frac{2\log(2|\mathcal{S}|)}{|M_k^\pi|}} + 3|\mathcal{S}|^2|\mathcal{A}|\sqrt{\frac{\tau^{\text{mix}}}{|B_k^\pi|}} + 16|\mathcal{S}|d_{\text{TV}}(\mu_{S_\infty}^\pi, \mu_{S_{k-N}}^\pi)\right).$$

The detailed proof is given in Appendix Section A.7. Using similar arguments to bound the first moment of $\|w(M_k^\pi, V_k)\|_2$, we can bound the second moment, which is given in the following lemma.

Lemma 3.4 (Second moment of $\|w(M_k^\pi, V_k)\|_2$)**.** *Let us consider the noise term, $w(M_k, V_k)$, is defined in (8). For $k \geq 0$, the second moment of $||w(M_k, V_k)||_2$ is bounded as follows:*

$$\mathbb{E}\left[\|w(M_k^\pi, V_k)\|_2^2\right] \leq \frac{4|\mathcal{S}|(R_{\max}+1)^2}{(1-\gamma)^2}\left(\frac{120(\log(2|\mathcal{S}|)^2}{|M_k^\pi|} + 4|\mathcal{S}|^4|\mathcal{A}|^2\frac{\tau^{\text{mix}}}{|B_k^\pi|} + 8|\mathcal{S}|d_{\text{TV}}(\mu_{S_\infty}^\pi, \mu_{S_{k-N}}^\pi)\right).$$

The proof is deferred to Appendix Section A.8 for compactness of the presentation.

3.4 Averaged iterate convergence

In the previous subsection, we derived a bound on the noise term. Based on it, in this subsection, we analyze the convergence of the averaged iterate of TD-learning with experience replay. In the next theorem, we present the main result for the convergence rate on the average iterate.

Theorem 3.5 (Convergence rate on average iterate of TD-learning)**.** *Suppose $N > \tau^{mix}$.*

1) For $T \geq 0$, the following inequality holds:

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[||V_k - V^\pi||_2^2] \leq \frac{1}{T}\frac{2|\mathcal{S}|}{\alpha(1-\gamma)}\mathbb{E}[||V_0 - V^\pi||_2^2] + \underbrace{\frac{32|\mathcal{S}|^2R_{\max}^2}{(1-\gamma)^3\mu_{\min}^\pi}\sqrt{\frac{8\log(2|\mathcal{S}|)}{L}}}_{E_1^{\text{avg}}\text{: Concentration error in the first moment}}$$

$$+\underbrace{\frac{32|\mathcal{S}|^2R_{\max}^2}{(1-\gamma)^3\mu_{\min}^{\pi}}\left(2|\mathcal{S}|^{\frac{3}{2}}|\mathcal{A}|\sqrt{\frac{\tau^{\text{mix}}}{N}}+\frac{64t_1^{\text{mix}}}{T}\right)}_{E_2^{\text{avg}}\text{: Markovian noise in the first moment}} \tag{13}$$

$$+\underbrace{\alpha\frac{4|\mathcal{S}|^2(R_{\max}+1)^2}{(1-\gamma)^3\mu_{\min}^{\pi}}\left(\frac{120(\log 2(|\mathcal{S}|))^2}{L}\right)}_{E_3^{\text{avg}}\text{: Concentration error in the second moment}}$$

$$+\underbrace{\alpha\frac{4|\mathcal{S}|^2(R_{\max}+1)^2}{(1-\gamma)^3\mu_{\min}^{\pi}}\left(4|\mathcal{S}|^4|\mathcal{A}|^2\frac{\tau^{\text{mix}}}{N}+\frac{32t_1^{\text{mix}}}{T}\right)}_{E_4^{\text{avg}}\text{: Markovian noise in the second moment}}.$$

2) For $T \geq 0$, the following inequality holds:

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{T}\sum_{k=0}^{T}V_k-V^{\pi}\right\|_2\right] \leq & \frac{1}{\sqrt{T}}\sqrt{\frac{2|\mathcal{S}|}{\alpha(1-\gamma)}}||V_0-V^{\pi}||_2+\sqrt{\frac{32|\mathcal{S}|^2R_{\max}^2}{(1-\gamma)^3\mu_{\min}^{\pi}}\sqrt{\frac{8\log(2|\mathcal{S}|)}{L}}}\\ &+\sqrt{\frac{32|\mathcal{S}|^2R_{\max}^2}{(1-\gamma)^3\mu_{\min}^{\pi}}\left(2|\mathcal{S}|^{\frac{3}{2}}|\mathcal{A}|\sqrt{\frac{\tau^{\text{mix}}}{N}}+\frac{64t_1^{\text{mix}}}{T}\right)}\\ &+\sqrt{\alpha\frac{4|\mathcal{S}|^2(R_{\max}+1)^2}{(1-\gamma)^3\mu_{\min}^{\pi}}\frac{120(\log(2|\mathcal{S}|))^2}{L}}\\ &+\sqrt{\alpha\frac{4|\mathcal{S}|^2(R_{\max}+1)^2}{(1-\gamma)^3\mu_{\min}^{\pi}}\left(4|\mathcal{S}|^4|\mathcal{A}|^2\frac{\tau^{\text{mix}}}{N}+\frac{32t_1^{\text{mix}}}{T}\right)}.\end{aligned}$$

The proof is given in Appendix A.9. Several comments can be made about the results in Theorem 3.5. The term (13) arises because $\mathbb{E}[(V_k-V^{\pi})^{\top}A^{\top}Mw(M_k^{\pi},V_k)]$ is not zero in comparison to the i.i.d. sampling update in (3). This is due to the Markovian noise and correlation between V_k and the samples in the replay buffer. As discussed in Section 3.3, the bound on the term $\mathbb{E}[(V_k-V^{\pi})^{\top}A^{\top}Mw(M_k^{\pi},V_k)]$ can be decomposed into two parts: the concentration error corresponding to the mini-batch uniformly sampled from the replay buffer, denoted as E_1^{avg}, and the Markovian noise term, denoted as E_2^{avg}. The first part, E_1^{avg}, can be controlled by the size of the mini-batch, decreasing at the order of $O\left(\frac{1}{\sqrt{L}}\right)$, as shown in Lemma 3.3. On the other hand, the Markovian noise term, E_2^{avg}, can be controlled by the size of the replay buffer, decreasing at the order of $O\left(\frac{1}{\sqrt{N}}\right)$. The terms, E_3^{avg} and E_4^{avg}, arise from the second moment of $\mathbb{E}[||w(M_k,V_k)||_2^2]$. These terms are non-zero in both i.i.d. sampling and Markovian noise cases under the standard TD-learning update (3). However, the errors can be controlled by the size of the mini-batch and replay buffer, as shown in Lemma 3.4. Specifically, E_3^{avg} and E_4^{avg} can be decreased at the order of $O\left(\frac{1}{L}\right)$ and $O\left(\frac{1}{N}\right)$, respectively.

Furthermore, the condition $N > \tau^{\text{mix}}$ is standard in the literature in sense that the analysis of Markovian observation model holds after a quantity related to mixing time (Srikant & Ying, 2019; Chen et al., 2022).

3.5 Comparative analysis

Table 1 presents a comprehensive comparison of the finite-time analysis of TD-learning. In the context of on-policy linear function approximation, and under the assumption of starting from the stationary distribution, the work presented in Bhandari et al. (2018) derives the following convergence rate of $\frac{1}{\sqrt{T}}$ for the averaged iterate of TD-learning with a constant step-size, where $T \in \mathbb{N}$ represents the final time of the iterate:

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{k=0}^{T}V_k-V^{\pi}\right\|_2\right] \leq O\left(\sqrt{\frac{\log T}{2(1-\gamma)^3\mu_{\min}^{\pi}\sqrt{T}}}\right).$$

Notably, our result does not impose any condition on the step-size $\alpha \in (0,1)$, indicating that the use of experience replay memory can ease the strict requirements for selecting a constant step-size. Moreover, we

Table 1: Comparitive analysis on results of root mean squared error of averaged iterate convergence using constant step-size.

Method	Experience replay	Observation model	Step-size	Feature	Initial distribution
Ours	✓	Markovian	$\alpha \in (0,1)$	Tabular	Arbitrary
Bhandari et al. (2018)	✗	Markovian	$\frac{1}{\sqrt{T}}$	Linear	Stationary
Lee & Kim (2022)	✗	i.i.d.	$\alpha \in (0,1)$	Tabular	Arbitrary
Lakshmi-narayanan & Szepesvari (2018)	✗	i.i.d.	Universal	Linear	Arbitrary

The paper (Bhandari et al., 2018) adopted the assumption that the initial state distribution is already the stationary distribution for simplicity of the proof. Moreover, Lakshminarayanan & Szepesvari (2018) derives a convergence rate of $O(\frac{1}{T})$ for general linear stochastic approximation problems. The universal step-size means that the step-size is dependent on the general linear stochastic approximation problems.

can achieve fast convergence rate while maintaining small constant error by controlling the size of mini-batch and replay buffer instead of controlling the step-size. The result in Lee & Kim (2022) also holds for general step-size condition, but to obtain smaller bias, it requires small step-size yielding slower convergence rate.

Under i.i.d. observation model and linear function approximation, Theorem 1 in Lakshminarayanan & Szepesvari (2018) provides $O\left(\frac{1}{\sqrt{T}}\right)$ convergence rate for the root mean squared error using a specific step-size depending on the model parameters, which cannot be known beforehand in TD-learning. The work assumes a more general setup. However, the step-size condition depends on not only the feature space but also $\mu^{\pi}_{\min}$, the minimum state-visitation distribution even in the tabular case.

3.6 Final iterate convergence

In the preceding subsection, we presented a finite-time analysis of the averaged iterate. In this subsection, we extend our analysis to investigate the convergence of the final iterate in TD-learning with a replay buffer, following a similar approach to the one used in the previous section. Instead of using Lyapunov arguments, we utilize the recursive formulas and the fact that $||A||_\infty < 1$ as given in Lemma 3.2. In contrast to the averaged iterate analysis, we assume that the initial distribution corresponds to the stationary distribution of the Markov chain, a common assumption in the literature (Bhandari et al., 2018; Nagaraj et al., 2020; Jain et al., 2021). With the assumption, we are able to derive the convergence rate of the final iterate.

Theorem 3.6. *Suppose $S_{-N} \sim \mu^{\pi}_{S_\infty}$.*

1) For any $k \geq 0$, we have

$$\begin{aligned}\mathbb{E}\left[\|V_k - V^\pi\|_2^2\right] \leq & ||V_0 - V^\pi||_2^2 \|\mathcal{S}|(1-\alpha(1-\gamma)\mu^\pi_{\min})^{2k+2} \\ & + \underbrace{\frac{64|\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^3\mu^\pi_{\min}}\sqrt{\frac{8\log(2|\mathcal{S}|)}{L}}}_{E_1^{\text{final}}:\ \textit{Concentration error in the first moment}} + \underbrace{\frac{64|\mathcal{S}|^{\frac{7}{2}}|\mathcal{A}|R_{\max}^2}{(1-\gamma)^3\mu^\pi_{\min}}\sqrt{\frac{\tau^{\text{mix}}}{N}}}_{E_2^{\text{final}}:\ \textit{Markovian noise in the first moment}} \\ & + \underbrace{\alpha\frac{4|\mathcal{S}|(R_{\max}+1)^2}{(1-\gamma)^3\mu^\pi_{\min}}\frac{120(\log 2|\mathcal{S}|)^2}{L}}_{E_3^{\text{final}}:\ \textit{Concentration error in the second moment}} + \underbrace{\alpha\frac{16|\mathcal{S}|^5|\mathcal{A}|^2(R_{\max}+1)^2}{(1-\gamma)^3\mu^\pi_{\min}}\frac{\tau^{\text{mix}}}{N}}_{E_4^{\text{final}}:\ \textit{Markovian noise in the second moment}}.\end{aligned} \tag{14}$$

2) For any $k \geq 0$, we have

$$\mathbb{E}\left[\|V_k - V^\pi\|_2\right] \leq \sqrt{|\mathcal{S}|}\,\|V_0 - V^\pi\|_2\,(1-\alpha(1-\gamma)\mu^\pi_{\min})^{k+1}$$

$$+ \frac{8|\mathcal{S}|R_{\max}}{(1-\gamma)^{\frac{3}{2}}(\mu^{\pi}_{\min})^{\frac{1}{2}}}\sqrt{\sqrt{\frac{8\log(2|\mathcal{S}|)}{L}} + 2|\mathcal{S}|^{\frac{3}{2}}|\mathcal{A}|\sqrt{\frac{\tau^{\text{mix}}}{N}}} \tag{15}$$
$$+ \sqrt{\alpha}\frac{2|\mathcal{S}|^{\frac{1}{2}}(R_{\max}+1)}{(1-\gamma)^{\frac{3}{2}}(\mu^{\pi}_{\min})^{\frac{1}{2}}}\sqrt{\frac{120(\log(2|\mathcal{S}|)^2}{L} + 4|\mathcal{S}|^4|\mathcal{A}|^2\frac{\tau^{\text{mix}}}{N}}.$$

The proof is given in Appendix Section A.10. In a manner similar to the case of the averaged iterate, the term (14) arises due to the non-zero value of $\mathbb{E}[(V_k - V^{\pi})^{\top}A^{\top}Mw(M_k^{\pi}, V_k)]$ as compared to the i.i.d. sampling update with (3). This non-zero value is caused by Markovian noise and correlation between V_k and the samples in the replay buffer. As explained in Section 3.3, the bound on the term $\mathbb{E}[(V_k - V^{\pi})^{\top}A^{\top}Mw(M_k^{\pi}, V_k)]$ can be decomposed into two parts: E_1^{final}, which is the concentration error of uniform random sampling, and E_2^{final}, which is the Markovian noise term. As can be seen from Lemma 3.3, E_1^{final} can be controlled by the size of mini-batch, where it can be decreased at the order of $O\left(\frac{1}{\sqrt{L}}\right)$. Moreover, the Markovian noise E_2^{final}can be controlled by the replay buffer size, decreasing at the order of $O\left(\frac{1}{\sqrt{N}}\right)$. The error bounds provided in above are new in the sense that the it shows dependency on the size of the replay buffer which can be controlled by adjusting the buffer size, L and N.

Furthermore, the assumption on the initial distribution is just for simplicity of the proof, and it was also used in Bhandari et al. (2018). The assumption can be relaxed to the more practical case that the initial state distribution is arbitrary with some more efforts. To this end, we can bound the error caused by the initial distribution using the geometric mixing property of the uniformly ergodic Markov chain.

3.7 Comparative analysis

The overall comparative analysis is given in Table 2. Under the Markovian assumption and on-policy linear function approximation, Bhandari et al. (2018) provided a final iterate convergence under the constant step-size α smaller than $\frac{1}{\mu^{\pi}_{\min}(1-\gamma)}$, which is

$$\mathbb{E}\left[\|V_T - V^{\pi}\|_2\right] \leq \left(e^{-\alpha(1-\gamma)\mu^{\pi}_{\min}T}\right)\|V^{\pi} - V_0\|_2 + \underbrace{\sqrt{\alpha\left(\frac{(9+12t^{\text{mix}}(\alpha))}{2(1-\gamma)^2\mu^{\pi}_{\min}}\right)}}_{\text{Error from constant step-size}}. \tag{16}$$

Although the constant step-size is a widely used method, it introduces constant error terms. However, the use of experience replay can reduce these errors. Specifically, as stated in the second statement of Theorem 3.6, the term (15) decreases at the order of $O\left(\frac{1}{L^{\frac{1}{4}}} + \frac{1}{N^{\frac{1}{4}}}\right)$. Furthermore, as in the averaged iterate case, since we do not impose any condition on the step-size, we can achieve fast convergence rate while maintaining small constant error by controlling the size of mini-batch and replay buffer instead of the step-size.

The work in Lee & Kim (2022) provided a final iterate convergence of tabular TD-learning with a constant step-size and i.i.d. observation models, where the step-size induces constant errors proportional to $O(\sqrt{\alpha})$.

The approach in Srikant & Ying (2019) provided a mean squared bound of TD-learning with linear function approximation and Markovian observation models, which is given by $\mathbb{E}\left[\|V_T - V^{\pi}\|_2^2\right] \leq O\left((1-\alpha c_1)^{T-\tau} + c_2\alpha\tau\right)$ where c_1, c_2 are model dependent parameters, and τ is the mixing time such that $\forall k \geq \tau$,

$$\left\|\gamma D^{\pi}P^{\pi} - D^{\pi} - (\mathbb{E}\left[\gamma e_{s_k}^{\top}e_{s_{k+1}} - e_{s_k}^{\top}e_{s_k}\right])\right\|_2 \leq \delta, \quad \left\|D^{\pi}R^{\pi} - \mathbb{E}\left[e_{s_k}^{\top}R^{\pi}\right]\right\|_2 \leq \delta.$$

However, the choice of the step-size depends on the mixing time and the model parameters.

Under the i.i.d. assumption and using linear function approximation, Dalal et al. (2018) derived $O\left(\frac{1}{T^{\sigma}}\right)$, $\sigma \in (0,1)$ bounds on the mean squared error bound, which is worse than $O\left(\frac{1}{T}\right)$ convergence rate. Since we used constant step-size, the result is not directly comparable.

Table 2: Comparitive analysis on results of root mean squared error of final iterate convergence using constant step-size

Method	Experience replay	Observation model	Step-size	Feature	Constant error term
Ours	✓	Markovian	$\alpha \in (0,1)$	Tabular	$O\left(\frac{1}{L^{\frac{1}{4}}} + \frac{\tau^{\text{mix}}}{N^{\frac{1}{4}}}\right)$
Bhandari et al. (2018)	✗	Markovian	Model dependent	Linear	$O\left(\sqrt{\alpha \log(\alpha)}\right)$
Lee & Kim (2022)	✗	i.i.d.	$\alpha \in (0,1)$	Tabular	$O(\sqrt{\alpha})$
Srikant & Ying (2019)	✗	Markovian	Model dependent	Linear	$O(\sqrt{\alpha})$
Dalal et al. (2018)	✗	i.i.d.	$\frac{1}{(k+1)^{\sigma}},\ \sigma \in (0,1)$	Linear	$O(\frac{1}{T^{\sigma}}),\ \sigma \in (0,1)$

Model dependent step size implies that the step size depends on model parameters, e.g., maximum eignevalue of matrix A, discount factor γ, mixing time.

Overall, setting L and N instead of α will be useful in the following scenario : Suppose we are already given a sufficient amount of data as in offline RL setting in Agarwal et al. (2020), and then interaction with the environment is allowed. Then, the convergence rate, $\mathcal{O}(\exp(-\alpha k))$, becomes faster as we choose large α, which is possible thanks to $\alpha \in (0,1)$. This will only require $\tilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)\mu^{\pi}_{\min}} \frac{\ln\left(\frac{1}{\epsilon}\right)}{\alpha}\right)$ number of samples for $\mathbb{E}\left[||V_k - V^{\pi}||\right] \leq \epsilon$ to hold for $\epsilon > 0$, which is the so-called sample complexity. In other words, if we exclude N in the sample complexity, permitting large α yields better sample complexity result. However, as discussed in Appendix Section A.11, if we take into account N in the sample complexity, the overall sample complexity result $\tilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^7|\mathcal{A}|^2\tau^{\text{mix}}}{(1-\gamma)^3\mu^{\pi}_{\min}} \frac{1}{\alpha} \frac{1}{\epsilon^4}\right)$ can be sub-optimal compared to the result of $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2} \frac{t^{\text{mix}}}{(1-\gamma)^3(\mu^{\pi}_{\min})^2}\right)$ by Bhandari et al. (2018), which leaves room for improvement in terms of $\frac{1}{1-\gamma}$ an $\mu^{\pi}_{\min}$ for our analysis.

Furthremore, setting L and N would be better than α is because most of the related works require condition on α for the convergence statement to hold. However, we do not require such conditions, while the error can be controlled by L and N. Moreover, the convergence statement holds for arbitrary $\alpha \in (0,1)$, L, and N.

Lastly, we would like to emphasize that our main contributions are as providing the first analysis of reinforcement learning in a setting that closely resembles the deep Q-learning approach described in Mnih et al. (2015) within the TD-learning framework, and providing a rigorous analysis on the relation between the error bound and the mini-batch size (L), and the replay buffer size (N) within the TD-learning framework. We hope these contributions will provide valuable insights to the community. In particular, the error bounds given in our paper are new in the sense that the it depends on the size of the replay buffer which can be controlled by adjusting the buffer size. For instance, the constant error ball caused by the constant step-size can become arbitrarily small by increasing the buffer size.

4 Conclusion

In this work, we have undertaken an analysis of the behavior of TD-learning utilizing experience replay memory under a Markovian observation model, which has thus far been unexplored despite the prevalence of experience replay memory in RL algorithms. By leveraging a simple matrix concentration inequality and the geometric mixing property of irreducible and aperiodic Markov chains, we have demonstrated that the expected root mean squared error of the averaged iterate of TD-learning can be reduced at the order of $O\left(\frac{1}{L^{\frac{1}{4}}} + \frac{1}{N^{\frac{1}{4}}}\right)$ under the constant step-size. Similarly, for the final iterate case, we have established that the root mean squared error can be effectively reduced at the order of $O\left(\frac{1}{L^{\frac{1}{4}}} + \frac{1}{N^{\frac{1}{4}}}\right)$. Potential avenues for future research include extending the proposed analysis frameworks to off-policy and linear function approximation settings, as well as investigating the applicability of these results to Q-learning.

References

Naman Agarwal, Syomantak Chaudhuri, Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps. *arXiv preprint arXiv:2110.08440*, 2021.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

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

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

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

George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

Chi-Tsong Chen. *Linear system theory and design*. Saunders college publishing, 1984.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874*, 2020.

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. *Automatica*, 146:110623, 2022.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Shirli Di-Castro, Dotan Di Castro, and Shie Mannor. Sim and real: Better together. *Advances in Neural Information Processing Systems*, 34:6868–6880, 2021.

Shirli Di-Castro, Shie Mannor, and Dotan Di Castro. Analysis of stochastic processes through replay buffers. In *International Conference on Machine Learning*, pp. 5039–5060. PMLR, 2022.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

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

William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, and Will Dabney. Revisiting fundamentals of experience replay. In *International Conference on Machine Learning*, pp. 3061–3071. PMLR, 2020.

Zhang-Wei Hong, Tao Chen, Yen-Chen Lin, Joni Pajarinen, and Pulkit Agrawal. Topological experience replay. *arXiv preprint arXiv:2203.15845*, 2022.

Bin Hu and Usman Syed. Characterizing the exact behaviors of temporal difference learning algorithms using markov jump linear system theory. *Advances in neural information processing systems*, 32, 2019.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

Prateek Jain, Suhas S Kowshik, Dheeraj Nagaraj, and Praneeth Netrapalli. Streaming linear system identification with reverse experience replay. *arXiv preprint arXiv:2103.05896*, 2021.

Hassan K Khalil. *Nonlinear control*, volume 406. Pearson New York, 2015.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Suhas Kowshik, Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Streaming linear system identification with reverse experience replay. *Advances in Neural Information Processing Systems*, 34:30140–30152, 2021.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355. PMLR, 2018.

Nevena Lazic, Dong Yin, Yasin Abbasi-Yadkori, and Csaba Szepesvari. Improved regret bound and experience replay in regularized policy iteration. In *International Conference on Machine Learning*, pp. 6032–6042. PMLR, 2021.

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

Donghwan Lee and Do Wan Kim. Analysis of temporal difference learning: Linear system approach. *arXiv preprint arXiv:2204.10479*, 2022.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3):293–321, 1992.

Xiao-Yang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Dan Wang. Finrl: A deep reinforcement learning library for automated stock trading in quantitative finance. *arXiv preprint arXiv:2011.09607*, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3): 127–239, 2014.

Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2015.

Egor Rotinov. Reverse experience replay. *arXiv preprint arXiv:1910.08780*, 2019.

Mohit Sewak. Deep q network (dqn), double dqn, and dueling dqn. In *Deep Reinforcement Learning*, pp. 95–108. Springer, 2019.

Avi Singh, Huihan Liu, Gaoyue Zhou, Albert Yu, Nicholas Rhinehart, and Sergey Levine. Parrot: Data-driven behavioral priors for reinforcement learning. *arXiv preprint arXiv:2011.10024*, 2020.

Satinder P Singh and Richard S Sutton. Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1-3):123–158, 1996.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pp. 2803–2830. PMLR, 2019.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

A Appendix

A.1 Notations

Throughout the paper, the following notations will be adopted: $\mathbb{R}$: set of real numbers; $\mathbb{R}^n$: set of all n-dimensional vectors; $\mathbb{R}^{n\times m}$: set of all $n \times m$ real matrices; $\mathbb{Z}_+$: set of all non-negative integers; $\mathbb{R}_+$: set of non-negative real numbers; $\mathbb{R}_+$: set of non-negative real numbers; for matrix $A \in \mathbb{R}^{n\times m}$, $[A]_{ij}$, $1 \leq i \leq n, 1 \leq j \leq m$ denotes i-th row and j-th column element of A; $e_s \in \mathbb{R}^n$ for $1 \leq s \leq n$: s-th basis vector of $\mathbb{R}^{|\mathcal{S}|}$ space, i.e., only the s-th element of e_s is one and other elements are zero; $||A||_\infty$ for $A \in \mathbb{R}^{n\times m}$ denotes the infininty norm $||A||_\infty := \max_{1\leq i\leq n} \sum_{j=1}^m |[A]_{ij}|$; $|S|$: cardinality of a finite set S; $O(\cdot)$ denotes the big O notation.

A.2 Technical Lemmas

Lemma A.1. *[Concentration bound for i.i.d. matrix random variables (Tropp et al. (2015),Corollary 6.2.1)] Let $X \in \mathbb{R}^{d_1\times d_2}$, where d_1 and d_2 are some positive integers. Moreover, assume that the sequence of random matrices $\{X_k\}_{k=1}^n$ are i.i.d. samples from a distribution such that $\mathbb{E}[X_k] = X$ and $||X_k||_2 \leq X_{\max}$ for all $1 \leq k \leq n$. Let $\sigma = ||\mathbb{E}[X_k X_k^\top]||_2$. Then, we get*

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k - X\right\|_2 \geq t\right] \leq (d_1+d_2)\exp\left(-nt^2/(\sigma + 2X_{\max}t/3)\right),$$

and

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k - X\right\|_2\right] \leq \sqrt{\frac{2\sigma\log(d_1+d_2)}{n}} + \frac{2L\log(d_1+d_2)}{3n}.$$

With the above result, a bound on the second moment $\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k - X\right\|_2^2\right]$ can be obtained as follows.

Corollary A.2. *Let $X \in \mathbb{R}^{d_1\times d_2}$. Assume that the sequence of random matrices $\{X_k\}_{k=1}^n$ are i.i.d. samples from a distribution such that $\mathbb{E}[X_k] = X$ and $||X_k||_2 \leq X_{\max}$ for all $1 \leq k \leq n$. Letting $||\mathbb{E}[X_k X_k^\top]||_2 \leq \sigma$, the corresponding second moment can be bounded as follows:*

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k - X\right\|_2^2\right]\\ \leq&\frac{2\sigma}{n}\log(d_1+d_2) + \frac{2\sigma}{n} + \frac{16X_{\max}^2}{9n^2}(\log(d_1+d_2))^2 + \frac{32X_{\max}^2}{9n^2}\log(d_1+d_2) + \frac{8X_{\max}}{3n}.\end{aligned}$$

Proof. The proof is completed by the inequalities

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^{n}X_k - X\right\|_2^2\right] \\
&= \int_0^\infty \mathbb{P}\left[\left\|\frac{1}{n}\sum_{k=1}^{n}X_k - X\right\|_2 \geq \sqrt{t}\right] dt \\
&\leq \int_0^\infty \min\left\{(d_1+d_2)\exp(-nt/(\sigma + 2X_{\max}\sqrt{t}/3)), 1\right\} dt \\
&\leq \int_0^\infty \min\left\{1, (d_1+d_2)\exp(-nt/(2\sigma))\right\} dt + \int_0^\infty \min\left\{1, (d_1+d_2)\exp(-3n\sqrt{t}/(4X_{\max}))\right\} dt \\
&\leq \frac{2\sigma}{n}\log(d_1+d_2) + \frac{2\sigma}{n} + \frac{16X_{\max}^2}{9n^2}(\log(d_1+d_2))^2 + \int_{\frac{16X_{\max}^2}{9n^2}(\log(d_1+d_2))^2}^\infty (d_1+d_2)\exp(-3n\sqrt{t}/(4X_{\max}))dt \\
&\leq \frac{2\sigma}{n}\log(d_1+d_2) + \frac{2\sigma}{n} + \frac{16X_{\max}^2}{9n^2}(\log(d_1+d_2))^2 + \int_{\frac{4X_{\max}}{3n}\log(d_1+d_2)}^\infty 2(d_1+d_2)u\exp(-3nu/(4X_{\max}))du \\
&\leq \frac{2\sigma}{n}\log(d_1+d_2) + \frac{2\sigma}{n} + \frac{16X_{\max}^2}{9n^2}(\log(d_1+d_2))^2 + \frac{32X_{\max}^2}{9n^2}\log(d_1+d_2) + \frac{8X_{\max}}{3n},
\end{aligned}
\tag{17}
$$

where the first equality follows from the inequality, $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}[Y \geq t]\, dt$ for non-negative random variable Y, which can be found in Exercise 2.14 in Casella & Berger (2021).

Moreover, the first inequality follows from applying Lemma A.1, and the last inequality follows from applying change of variables $\sqrt{t} = x$ and integration by parts. □

Lemma A.3 (Variance of empirical distribution, Paulin (2015), Proposition 3.21)**.** *Let $\{S_k\}_{k=0}^n$ be a Markov chain following transition kernel $\mathcal{P}^\pi$ with the state space $\mathcal{S}$ and starting from its stationary distribution $\mu_{S_\infty}^\pi$, i.e., $S_0 \sim \mu_{S_\infty}^\pi$. The empirical distribution is defined as $\mu^{\mathrm{em}}(s) := \sum_{k=1}^n 1\{S_k = s\}/n,\ s \in \mathcal{S}$. Then,*

1) $\sum_{s\in\mathcal{S}} \mathbb{E}\left[(\mu_{S_\infty}^\pi(s) - \mu^{\mathrm{em}}(s))^2\right] \leq |\mathcal{S}|\frac{t^{\mathrm{mix}}}{n}$,

2) $\mathbb{E}\left[d_{\mathrm{TV}}(\mu_{S_\infty}^\pi, \mu^{\mathrm{em}}) \leq\right] \leq |\mathcal{S}|\sqrt{\frac{t^{\mathrm{mix}}}{n}}$.

Lemma A.4 (Bound on total variation distance, Chapter 4.5 and Exercise 4.2 in Levin & Peres (2017))**.** *Let $\{S_k\}_{k=0}^n$ be a Markov chain following transition kernel $\mathcal{P}^\pi$ with the state space $\mathcal{S}$. Denote its stationary distribution $\mu_{S_\infty}^\pi$ and mixing time as t^{mix}. Then, the following holds:*

1) For a non-negative integer l, we have $\sup_{s\in\mathcal{S}} d_{\mathrm{TV}}(\mathbb{P}[S_{lt^{\mathrm{mix}}} \mid S_0 = s], \mu_{S_\infty}^\pi) \leq 2^{-l}$.

2) $\sup_{s\in\mathcal{S}} d_{\mathrm{TV}}(\mathbb{P}[S_k \mid S_0 = s], \mu_{S_\infty}^\pi)$ *is non-increasing in terms of k.*

Lemma A.5 (Proposition 3.4 in Paulin (2015))**.** *For uniformly ergodic Markov chain, we have, for $\epsilon > 0$*

$$
t^{mix}(\epsilon) \leq t^{\mathrm{mix}}\left(\frac{1}{4}\right)\left(1 + 2\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{d_{\min}}\right)\right).
$$

A.3 Proof of Lemma 2.3

Before the main proof, the formal definition of irreducible and aperiodic Markov chain is reviewed first.

Definition A.6 (Irreducible and aperiodic Markov chain (Levin & Peres, 2017))**.** *A Markov chain is called irreducible if for any two states $x, y \in \mathcal{X}$, $e_x^\top P^k e_y > 0$ for some $k \geq 0$. Let $\mathcal{T}(x) := \{k \geq 1 \mid e_x^\top P^k e_x > 0\}$ be the set of time instances such that the probability of returning back to the initial state x is positive. The period of state x is the greatest common divisor of $\mathcal{T}(x)$. A Markov chain is said to be aperiodic if all states have period one.*

The proof of Lemma 2.3 is given below.

Proof. First of all, we prove that the extended Markov chain is irreducible. By hypothesis, the Markov chain, $(S_k)_{k=0}^{\infty}$, with transition kernel $\mathcal{P}^{\pi}$ is irreducible. Then, one can prove that starting from any $(S_k, S_{k+1}) = (x, y) \in \mathcal{O}$, there exists a positive probability that any $(S_{k+T}, S_{k+1+T}) = (w, z) \in \mathcal{O}$ can be reached for some $T \geq 0$. This is because by hypothesis, $\mathbb{P}[S_{k+T} = w \mid S_0 = x] > 0$ for some $T \geq 0$, and since $(w, z) \in \mathcal{O}$ is a realizable pair, $(S_{k+T}, S_{k+1+T}) = (w, z) \in \mathcal{O}$ is visited with a positive probability. This proves the irreducibility.

Next, we prove the aperiodicity. By hypothesis, the Markov chain, $(S_k)_{k=0}^{\infty}$, is aperiodic. To prove that the induced Markov chain $(S_k, S_{k+1})_{k=0}^{\infty}$ is aperiodic by contradiction, let us suppose that $(S_k, S_{k+1})_{k=0}^{\infty}$ is periodic with some period $T > 1$. Then, there exists a a state $s \in \mathcal{S}$ such that its period is larger than one, which contradicts the aperiodicity assumption of $(S_k)_{k=0}^{\infty}$. This completes the proof.

□

A.4 Proof of Lemma 3.1

Proof. The proof proceeds by induction. Suppose $||V_k||_{\infty} \leq \frac{R_{\max}}{1-\gamma}$ holds for $k \geq 0$. Rewriting (5), we have

$$V_{k+1} = \left(I + \alpha \frac{1}{|M_k^{\pi}|} \sum_{(s,r,s') \in M_k^{\pi}} (\gamma e_s e_{s'}^{\top} - e_s e_s^{\top}) \right) V_k + \alpha \frac{1}{|M_k^{\pi}|} \sum_{(s,r,s') \in M_k^{\pi}} e_s r.$$

We can prove that for $s \in \mathcal{S}$, if $(s, r, s') \notin M_k^{\pi}$, then $e_s^{\top} V_{k+1}$ is identical to $e_s^{\top} V_k$. If $(s, r, s') \in M_k^{\pi}$ for some $s \in \mathcal{S}$, then the update of $e_s^{\top} V_{k+1}$ can be expressed as follows:

$$e_s^{\top} V_{k+1} = e_s^{\top} \left\{ \left(I + \alpha \frac{1}{|M_k^{\pi}|} \sum_{(s,r,s') \in M_k^{\pi}} (\gamma e_s e_{s'}^{\top} - e_s e_s^{\top}) \right) V_k + \alpha \frac{1}{|M_k^{\pi}|} \sum_{(s,r,s') \in M_k^{\pi}} e_s r \right\}.$$

Consider a subset of the mini-batch M_k^{π} such that the first element of the tuple is s, i.e., $M_k^{\pi}(s) := \{(\tilde{s}, \tilde{r}, \tilde{s}') \in M_k^{\pi} : \tilde{s} = s\} \subseteq M_k^{\pi}$. Then, we can bound $e_s^{\top} V_{k+1}$ as follows:

$$\begin{aligned}
|e_s^{\top} V_{k+1}| &\leq \left| e_s^{\top} \left(I + \alpha \frac{1}{|M_k^{\pi}|} \sum_{(s,r,s') \in M_k^{\pi}(s)} (\gamma e_s e_{s'}^{\top} - e_s e_s^{\top}) \right) V_k \right| + \alpha \frac{|M_k^{\pi}(s)|}{|M_k^{\pi}|} R_{\max} \\
&\leq \left\| \left(I + \alpha \frac{1}{|M_k^{\pi}|} \sum_{(s,r,s') \in M_k^{\pi}(s)} (\gamma e_s e_{s'}^{\top} - e_s e_s^{\top}) \right)^{\top} e_s \right\|_1 ||V_k||_{\infty} + \alpha \frac{|M_k^{\pi}(s)|}{|M_k^{\pi}|} R_{\max} \\
&\leq \left(1 + \alpha \frac{|M_k^{\pi}(s)|}{|M_k^{\pi}|} (\gamma - 1) \right) ||V_k||_{\infty} + \alpha \frac{|M_k^{\pi}(s)|}{|M_k^{\pi}|} R_{\max} \\
&\leq \frac{R_{\max}}{1-\gamma},
\end{aligned}$$

where the first inequality follows from the application of the triangle inequality and the assumption that the reward is bounded, as given in Assumption 2.4. The second inequality is due to the Cauchy-Schwartz inequality, and the last inequality comes from the induction hypothesis. This completes the proof. □

A.5 Proof of Lemma 3.2

Proof. First of all, to prove the first statement, $||A||_{\infty}$ is bounded as

$$||A||_{\infty} = \max_{1 \leq i \leq |\mathcal{S}|} \left(1 - \alpha [D^{\pi}]_{ii} + \alpha \gamma [D^{\pi}] \sum_{j=1}^{|\mathcal{S}|} [P^{\pi}]_{ij} \right) = \min_{s \in \mathcal{S}} (1 - \alpha(1-\gamma)\mu^{\pi}(s)),$$

where the last equality is obtained by expanding A according to the definition (7) from the fact that P^π is a stochastic matrix, i.e., the row sum of the matrix equals one.

Next, for the second statement, one can readily prove that $M := \sum_{k=0}^{\infty}(A^k)^\top(A)^k$ is a solution of (9) by simply plugging it into M in (9). It remains to prove that M is bounded. In particular, $||M||_2$ is bounded as

$$\begin{aligned}
||M||_2 &\leq \|I\|_2^2 + \|A\|_2^2 + \|A\|_2^4 + \cdots \\
&\leq 1 + |\mathcal{S}|||A||_\infty^2 + |\mathcal{S}|\,\|A\|_\infty^4 + \cdots \\
&\leq 1 + |\mathcal{S}|(1-\alpha(1-\gamma)\mu_{\min}^\pi)(1 + (1-\alpha(1-\gamma)\mu_{\min}^\pi)^2 + (1-\alpha(1-\gamma)\mu_{\min}^\pi)^4 + \cdots \\
&\leq 1 + \frac{|\mathcal{S}|}{\alpha(1-\gamma)\mu_{\min}^\pi},
\end{aligned}$$

where the first inequality follows from applying triangle inequality and the fact that $||A^\top||_2 = ||A||_2$. The second inequality is due to the relation $||A||_2 \leq \sqrt{|\mathcal{S}|}||A||_\infty$, and the third inequality comes from Lemma 3.2. □

A.6 Auxiliary lemmas to prove Lemma 3.3

Let $\mu_{\tilde{S}}^{\tilde{B}_k^\pi}$ and $\mu_{\tilde{S},\tilde{S}'}^{\tilde{B}_k^\pi}$ be the empirical distributions of the state s and the consecutive state pair (s, s') in the replay buffer, $\tilde{B}_k^\pi$, respectively, which are defined in the same manner as $\mu_S^{B_k^\pi}$ and $\mu_{S,S'}^{B_k^\pi}$, respectively.

Lemma A.7. *For $k \geq 0$, we have the following bounds:*

1) $\mathbb{E}\left[\left\|D^\pi - D^{\tilde{B}_k^\pi}\right\|_2\right] \leq \sqrt{|\mathcal{S}|}\sqrt{\frac{t_1^{\text{mix}}}{|\tilde{B}_k^\pi|}}$,

2) $\mathbb{E}\left[\left\|D^\pi P^\pi - D^{\tilde{B}_k^\pi} P^{\tilde{B}_k^\pi}\right\|_2\right] \leq |\mathcal{S}|\sqrt{\frac{t_2^{\text{mix}}}{|\tilde{B}_k^\pi|}}$,

3) $\mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2\right] \leq |\mathcal{S}|^{\frac{5}{2}}|\mathcal{A}|R_{\max}\sqrt{\frac{t_2^{\text{mix}}}{|\tilde{B}_k^\pi|}}$.

Proof. First of all, the expected value of $D^\pi - D^{\tilde{B}_k^\pi}$ can be bounded as follows:

$$\mathbb{E}\left[\left\|D^\pi - D^{\tilde{B}_k}\right\|_2\right] \leq \mathbb{E}\left[\sqrt{\sum_{s\in\mathcal{S}}\left(\mu_{S_\infty}^\pi(s) - \mu_{\tilde{S}}^{\tilde{B}_k^\pi}(s)\right)^2}\right] \leq \sqrt{|\mathcal{S}|}\sqrt{\frac{t_1^{\text{mix}}}{|\tilde{B}_k^\pi|}},$$

which proves the first statement. In the above inequalities, the first inequality follows from the fact that $||A||_2 \leq ||A||_F$, where $A \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ and $||\cdot||_F$ stands for Frobenius norm. The last inequality follows from Jensen's inequality and Lemma A.3 in Appendix.

The second statement can be proved using the following similar arguments:

$$\mathbb{E}\left[\left\|D^\pi P^\pi - D^{\tilde{B}_k^\pi} P^{\tilde{B}_k^\pi}\right\|_2\right] \leq \mathbb{E}\left[\sqrt{\sum_{(s,s')\in\mathcal{S}\times\mathcal{S}}\left(\mu_{S_\infty,S'_\infty}^\pi(s,s') - \mu_{\tilde{S},\tilde{S}'}^{\tilde{B}_k^\pi}(s,s')\right)^2}\right] \leq |\mathcal{S}|\sqrt{\frac{t_2^{\text{mix}}}{|\tilde{B}_k^\pi|}},$$

where the first inequality follows from the fact that $||A||_2 \leq ||A||_F$, where $A \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ and $||\cdot||_F$, the second inequality is due to Lemma A.3 in Appendix, and follows from applying Jensen's inequality.

For the third statement, we can bound $\mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2\right]$ in a similar sense as follows:

$$\mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2\right] \leq \mathbb{E}\left[\sqrt{\sum_{s\in\mathcal{S}}\left(\mathbb{E}\left[r(S,A,S') \mid S=s,\pi\right]\mu_{S_\infty}^\pi(s) - \mu_{\tilde{S}}^{\tilde{B}_k^\pi}(s)[R^{\tilde{B}_k^\pi}]_s\right)^2}\right]$$

$$
\begin{aligned}
&\leq |\mathcal{S}||\mathcal{A}|R_{\max}\mathbb{E}\left[\sqrt{\sum_{s\in\mathcal{S}}\left(\sum_{s'\in\mathcal{S}}\mu^{\pi}_{S_\infty,S'_\infty}(s,s')-\mu^{\tilde{B}^{\pi}_k}_{\tilde{S},\tilde{S}'}(s,s')\right)^2}\right]\\
&\leq |\mathcal{S}|^{\frac{5}{2}}|\mathcal{A}|R_{\max}\sqrt{\frac{t_2^{\text{mix}}}{|\tilde{B}^{\pi}_k|}},
\end{aligned}
$$

where the second inequality applies the inequality $(\sum_{k=0}^{n} a_k b_k) \leq (\sum_{k=0}^{n} a_k)(\sum_{k=0}^{n} b_k)$ for $a_k, b_k \in \mathbb{R},\ k \geq 0$, and the relation between $\mu^{\pi}_{S_\infty}$ and $\mu^{\pi}_{S_\infty,S'_\infty}$ in (2), and

$$
\sum_{a\in\mathcal{A}}\frac{\sum\limits_{l=k-N+1}^{N}1\{s_l,a_l,s_{l+1}=(s,a,s')\}}{\sum\limits_{l=k-N+l}^{N}1\{s_l=i\}}\mu^{\tilde{B}^{\pi}_k}_{\tilde{S}}(s)=\mu^{\tilde{B}^{\pi}_k}_{\tilde{S},\tilde{S}'}(s,s').
$$

Moreover, the last inequality follows the same logic as that used for bounding $\mathbb{E}\left[\left\|D^{\pi}P^{\pi}-D^{\tilde{B}^{\pi}_k}P^{\tilde{B}^{\pi}_k}\right\|_2\right]$. This completes the proof. □

Lemma A.8. *For $k \geq 0$, we have*

1) $\mathbb{E}\left[\left\|D^{\pi}-D^{B^{\pi}_k}\right\|_2\right] \leq \sqrt{|\mathcal{S}|}\sqrt{\frac{t_1^{\text{mix}}}{|B^{\pi}_k|}}+4d_{\text{TV}}(\mu^{\pi}_{S_\infty},\mu^{\pi}_{S_{k-N}})$,

2) $\mathbb{E}\left[\left\|D^{\pi}P^{\pi}-D^{B^{\pi}_k}P^{B^{\pi}_k}\right\|_2\right] \leq |\mathcal{S}|\sqrt{\frac{t_2^{\text{mix}}}{|B^{\pi}_k|}}+4\sqrt{|\mathcal{S}|}d_{\text{TV}}(\mu^{\pi}_{S_\infty},\mu^{\pi}_{S_{k-N}})$,

3) $\mathbb{E}\left[\left\|D^{\pi}R^{\pi}-D^{B^{\pi}_k}R^{B^{\pi}_k}\right\|_2\right] \leq |\mathcal{S}|^{\frac{5}{2}}|\mathcal{A}|R_{\max}\sqrt{\frac{t_2^{\text{mix}}}{|B^{\pi}_k|}}+2\sqrt{|\mathcal{S}|}R_{\max}d_{\text{TV}}(\mu^{\pi}_{S_\infty},\mu^{\pi}_{S_{k-N}})$.

Proof. Let $\{\tilde{S}_k\}_{k\geq -N}$ be the stationary Markov chain defined in Section 3.3. The expected value of $||D^{\pi}-D^{B^{\pi}_k}||_2$ can be bounded using irreducible and aperiodic property of the Markov chain as follows:

$$
\begin{aligned}
\mathbb{E}\left[\left\|D^{\pi}-D^{B^{\pi}_k}\right\|_2\right] &= \sum_{s\in\mathcal{S}}\mathbb{E}\left[\left\|D^{\pi}-D^{B^{\pi}_k}\right\|_2\Big|S_{k-N}=s\right]\mathbb{P}[S_{k-N}=s]\\
&= \sum_{s\in\mathcal{S}}\mathbb{E}\left[\left\|D^{\pi}-D^{\tilde{B}^{\pi}_k}\right\|_2\Big|\tilde{S}_{k-N}=s\right]\mathbb{P}[\tilde{S}_{k-N}=s]\\
&\quad+\sum_{s\in\mathcal{S}}\mathbb{E}\left[\left\|D^{\pi}-D^{\tilde{B}^{\pi}_k}\right\|_2\Big|\tilde{S}_{k-N}=s\right]\left(\mathbb{P}[S_{k-N}=s]-\mathbb{P}[\tilde{S}_{k-N}=s]\right)\\
&\leq \mathbb{E}\left[\left\|D^{\pi}-D^{\tilde{B}^{\pi}_k}\right\|_2\right]+2\sum_{s\in\mathcal{S}}|\mathbb{P}[S_{k-N}=s]-\mathbb{P}[\tilde{S}_{k-N}=s]|\\
&\leq \sqrt{|\mathcal{S}|}\sqrt{\frac{t_1^{\text{mix}}}{|B^{\pi}_k|}}+4d_{\text{TV}}(\mu^{\pi}_{S_\infty},\mu^{\pi}_{S_{k-N}}).
\end{aligned}
$$

In the above inequalities, the first equality follows from the law of total expectation, and the second equality is due to the fact that $\mathbb{P}[S_{k-N+1},S_{k-N+2},\ldots,S_k \mid S_{k-N}] = \mathbb{P}[\tilde{S}_{k-N+1},\tilde{S}_{k-N+2},\ldots,\tilde{S}_k \mid \tilde{S}_{k-N}]$ because the transition probabilities of both Markov chains are identical. Next, the first inequality follows from the fact that $||D^{\pi}||_2$ and $||D^{\tilde{B}^{\pi}_k}||_2$ are both smaller than one, and the last equality follows from the definition of the total variation distance in (2.1), and applying Lemma A.7.

In the second statement, the expected value of $\left\|D^{\pi}P^{\pi}-D^{B^{\pi}_k}P^{B^{\pi}_k}\right\|_2$ can be bounded using the same argument as in the first statement. In particular, we have

$$
\mathbb{E}\left[\left\|D^{\pi}P^{\pi}-D^{B^{\pi}_k}P^{B^{\pi}_k}\right\|_2\right]
$$

$$
\begin{aligned}
&= \sum_{s\in\mathcal{S}} \mathbb{E}\left[\left\|D^\pi P^\pi - D^{\tilde{B}_k^\pi} P^{\tilde{B}_k^\pi}\right\|_2 \middle| \tilde{S}_{k-N} = s\right] \mathbb{P}[\tilde{S}_{k-N} = s] \\
&\quad + \sum_{s\in\mathcal{S}} \mathbb{E}\left[\left\|D^\pi P^\pi - D^{\tilde{B}_k^\pi} P^{\tilde{B}_k^\pi}\right\|_2 \middle| \tilde{S}_k = s\right] \left(\mathbb{P}[S_{k-N} = s] - \mathbb{P}[\tilde{S}_{k-N} = s]\right) \\
&\leq \sqrt{|\mathcal{S}|}\sqrt{\frac{t_2^{\text{mix}}}{N}} + 4\sqrt{|\mathcal{S}|} d_{\text{TV}}(\mu^\pi_{S_\infty}, \mu^\pi_{S_{k-N}}).
\end{aligned}
$$

The same logic holds for $\mathbb{E}\left[\left\|D^\pi R^\pi - D^{B_k^\pi} R^{B_k^\pi}\right\|_2\right]$. In particular, it follows from

$$
\begin{aligned}
&\mathbb{E}\left[\left\|D^\pi R^\pi - D^{B_k^\pi} R^{B_k^\pi}\right\|_2\right] \\
&= \sum_{s\in\mathcal{S}} \mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2 \middle| \tilde{S}_{k-N} = s\right] \mathbb{P}[\tilde{S}_{k-N} = s] \\
&\quad + \sum_{s\in\mathcal{S}} \mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2 \middle| \tilde{S}_{k-N} = s\right] \left(\mathbb{P}[S_{k-N} = s] - \mathbb{P}[\tilde{S}_{k-N} = s]\right) \\
&\leq |\mathcal{S}|^{\frac{5}{2}} |\mathcal{A}| R_{\max} \sqrt{\frac{t_2^{\text{mix}}}{|\tilde{B}_k^\pi|}} + 2\sqrt{|\mathcal{S}|} R_{\max} d_{\text{TV}}(\mu^\pi_{S_\infty}, \mu^\pi_{S_{k-N}}).
\end{aligned}
$$

This completes the proof. □

A.7 Proof of Lemma 3.3

Proof. First of all, the triangle inequality applied to the noise term in (12) leads to

$$
\begin{aligned}
\mathbb{E}[||w(M_k^\pi, V_k)||_2] &= \mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{i=1}^{|M_k^\pi|} \delta(O_i^k; V_k) - \Delta_k(V_k) + \Delta_k(V_k) - \Delta_\pi(V_k)\right\|_2\right] \\
&\leq \mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{i=1}^{|M_k^\pi|} \delta(O_i^k; V_k) - \Delta_k(V_k)\right\|_2\right] + \mathbb{E}\left[\|\Delta_k(V_k) - \Delta_\pi(V_k)\|_2\right]. \qquad (18)
\end{aligned}
$$

The first term in (18) is the difference of the empirically expected TD-errors in terms of replay buffer and mini-batch, which is bounded as

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{i=1}^{|M_k^\pi|} \delta(O_i^k; V_k) - \Delta_k(V_k)\right\|_2\right] \leq& \frac{2\sqrt{|\mathcal{S}|}R_{\max}}{1-\gamma}\left(\mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{(s,r,s')\in M_k^\pi}(e_s e_s^\top - D^{B_k^\pi})\right\|_2\right]\right. \\
&\left. + \mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{(s,s')\in M_k^\pi}\gamma(e_s e_{s'}^\top - D^{B_k^\pi} P^{B_k^\pi})\right\|_2\right]\right) \\
&+ \mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{(s,r,s')\in M_k^\pi}(e_s r - D^{B_k^\pi} R^{B_k^\pi})\right\|_2\right], \qquad (19)
\end{aligned}
$$

where the inequality is obtained simply by expanding the terms in (10) and (4) and applying triangle inequality, together with the boundedness of V_k and reward function in Assumption 2.4 and Lemma 3.1, respectively.

As the next step, we will upper bound the above terms using Bernstein inequality (Tropp et al., 2015). In particular, note that for $(s, r, s') \in M_k^\pi$, s can be thought as a uniform sample from $D^{B_k^\pi}$. Hence, using the matrix concentration inequality in Lemma A.1 in Appendix, one gets

$$
\mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{(s,r,s')\in M_k^\pi}(e_s e_s^\top - D^{B_k^\pi})\right\|_2\right] \leq \sqrt{\frac{8\log(2|\mathcal{S}|)}{|M_k^\pi|}},
$$

which is obtained by letting $\sigma = 1$ and $X_{\max} = 1$ in Lemma A.10 in the Appendix because $||\mathbb{E}[e_s e_s^\top]||_2 \leq 1$ and $||e_s e_s^\top||_2 \leq 1$.

Moreover, noting that $||\mathbb{E}[e_s e_{s'}^\top]||_2 \leq \mathbb{E}[||e_s e_{s'}^\top||_2] \leq 1$, and $||e_s e_{s'}^\top||_2 \leq 1$, we get the following bound on the second term:

$$\mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{(s,r,s')\in M_k^\pi}\gamma(e_s e_{s'}^\top - D^{B_k^\pi}P^{B_k^\pi})\right\|_2\right] \leq 2\gamma\sqrt{\frac{2\log(2|\mathcal{S}|)}{|M_k^\pi|}}.$$

In a similar sense, the third term can be bounded as

$$\mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{(s,r,s')\in M_k^\pi}(e_s r - D^{B_k^\pi}R^{B_k^\pi})\right\|_2\right] \leq 2R_{\max}\sqrt{\frac{2\log(2|\mathcal{S}|)}{|M_k^\pi|}},$$

which uses the inequalities, $||\mathbb{E}[e_s r]||_2 \leq R_{\max}$ and $||\mathbb{E}[r^2 e_s e_s^\top]||_2 \leq R_{\max}^2$. Collecting the above three inequalities yields the upper bound on (19), which is

$$\mathbb{E}\left[\left\|\frac{1}{|M_k^\pi|}\sum_{i=1}^{|M_k^\pi|}\delta(O_i^k;V_k) - \Delta_k(V_k)\right\|_2\right] \leq \frac{4\sqrt{|\mathcal{S}|}R_{\max}}{1-\gamma}\sqrt{\frac{8\log(2|\mathcal{S}|)}{|M_k^\pi|}}. \tag{20}$$

Now, we turn our attention to he second term, $\mathbb{E}\left[\|\Delta_\pi(V_k) - \Delta_k(V_k)\|_2\right]$, in (18), which is the difference between the expected TD-error with respect to the stationary distribution and the replay buffer distribution. Plugging the definitions in (10) and (11) into (18) yields

$$\begin{aligned}
&\mathbb{E}\left[\|\Delta_\pi(V_k) - \Delta_k(V_k)\|_2\right] \\
\leq &\frac{2\sqrt{|\mathcal{S}|}R_{\max}}{1-\gamma}\mathbb{E}\left[\left(\left\|D^\pi - D^{B_k^\pi}\right\|_2 + \left\|D^\pi P^\pi - D^{B_k^\pi}P^{B_k^\pi}\right\|_2\right)\right] + \mathbb{E}\left[\left\|D^\pi R^\pi - D^{B_k^\pi}R^{B_k^\pi}\right\|_2\right] \\
\leq &\frac{2\sqrt{|\mathcal{S}|}R_{\max}}{1-\gamma}\left(3|\mathcal{S}|^2|\mathcal{A}|\sqrt{\frac{\tau^{\text{mix}}}{|B_k^\pi|}} + 16|\mathcal{S}|d_{\text{TV}}(\mu_{S_\infty}^\pi, \mu_{S_{k-N}}^\pi)\right).
\end{aligned} \tag{21}$$

where the first inequality is due to the boundedness of V_k in Lemma 3.1, and the last inequality follows from Lemma A.8. Collecting the terms in (20) and (21), we obtain the desired result.

□

A.8 Proof of Lemma 3.4 (Bound on second moment of the noise term)

Proof. Expanding the noise term in (8) we get

$$\mathbb{E}[||w(M_k^\pi;V_k)||_2^2] \leq 2\left(\mathbb{E}\left[\left\|\frac{1}{L}\sum_{i=1}^{L}\delta(O_i^k;V_k) - \Delta_k(V_k)\right\|_2^2\right] + \mathbb{E}\left[\|\Delta_\pi(V_k) - \Delta_k(V_k)\|_2^2\right]\right), \tag{22}$$

where the inequality follows from the fact that $||a+b||^2 \leq 2||a||^2 + 2||b||^2$ for $a, b \in \mathbb{R}^n$. In what follows, each terms in the above inequality will be bounded. First of all, the term $\mathbb{E}\left[\left\|\frac{1}{L}\sum_{i=1}^{L}\delta(O_i^k;V_k) - \Delta_k(V_k)\right\|_2^2\right]$ in (22) is bounded as

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{L}\sum_{i=1}^{L}\delta(O_i^k;V_k) - \Delta_k(V_k)\right\|_2^2\right] \\
=&\mathbb{E}\left[\left\|\frac{1}{L}\sum_{(s,r,s')\in M_k}(e_s r - D^{B_k^\pi}R^{B_k^\pi}) + (e_s e_s^\top - D^{B_k})(V_k) + \gamma(e_s e_{s'}^\top - D^{B_k}P^{B_k})V_k\right\|_2^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{3|\mathcal{S}|R_{\max}^2}{(1-\gamma)^2}\left(\mathbb{E}\left[\left\|\frac{1}{L}\sum_{(s,r,s')\in M_k}(e_s e_s^\top - D^{B_k^\pi})\right\|_2^2\right] + \mathbb{E}\left[\left\|\frac{1}{L}\sum_{(s,r,s')\in M_k}\gamma(e_s e_{s'}^\top - D^{B_k^\pi}P^{B_k^\pi})\right\|_2^2\right]\right) \\
&\quad + 3\mathbb{E}\left[\left\|\frac{1}{L}\sum_{(s,r,s')\in M_k}(e_s r - D^{B_k^\pi}R^{B_k^\pi})\right\|_2^2\right] \\
&\leq \frac{|\mathcal{S}|(1+R_{\max})^2}{(1-\gamma)^2}\frac{480}{|M_k^\pi|}(\log(2|\mathcal{S}|)^2,
\end{aligned}
$$

where the first equality is obtained by expanding the TD-error term in (4) and expected TD-error in terms of replay buffer in (10). The first inequality follows from the fact that $||a+b+c||^2 \leq 3(||a||^2 + ||b||^2 + ||c||^2)$ for $a, b, c \in \mathbb{R}^n$, and from the fact that the iterate V_k is bounded in (3.1). We use concentration inequality in Corollary A.2 to bound the last inequality. To bound $\mathbb{E}\left[\left\|\frac{1}{L}\sum_{(s,r,s')\in M_k}(e_s e_s^\top - D^{B_k^\pi})\right\|_2^2\right]$, we let $\sigma = 1$ and $X_{\max} = 1$ in Corollary A.2 since $||\mathbb{E}[e_s e_s^\top]||_2 \leq 1$ and $||e_s e_s^\top||_2 \leq 1$. Moreover, we can bound remaining terms in similar sense, which yields the last inequality.

To proceed further, Lemma A.9 is needed, which is provided after this proof. Applying Lemma A.9, the term $\mathbb{E}\left[\|\Delta_\pi(V_k) - \Delta_k(V_k)\|_2^2\right]$ is bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\|\Delta_\pi(V_k) - \Delta_k(V_k)\|_2^2\right] \\
=&\mathbb{E}\left[\left\|(D^\pi - D^{B_k^\pi})V_k + (D^\pi R^\pi - D^{B_k^\pi}R^{B_k^\pi}) + \gamma(D^\pi P^\pi - D^{B_k^\pi}P^{B_k^\pi})V_k\right\|_2^2\right] \\
\leq&\frac{3|\mathcal{S}|R_{\max}^2}{(1-\gamma)^2}\left(4|\mathcal{S}|^4|\mathcal{A}|^2\frac{\max\{t_1^{\text{mix}}, t_2^{\text{mix}}\}}{|B_k^\pi|} + 8|\mathcal{S}|d_{\text{TV}}(\mu_{S_\infty}^\pi, \mu_{S_{k-N}}^\pi)\right),
\end{aligned}
$$

where the inequality follows from applying the bound on V_k in Lemma 3.1, and collecting the terms in Lemma A.9. This completes the proof. □

The following two lemmas have been used in the proof above, and they are formally introduced in the sequel.

Lemma A.9. *For $k \geq 0$, we have*

1) $\mathbb{E}[||D^\pi - D^{B_k^\pi}||_2^2] \leq |\mathcal{S}|\frac{t_1^{\text{mix}}}{|B_k^\pi|} + 8d_{\text{TV}}(\mu_{S_\infty}^\pi, \mu_{S_{k-N}}^\pi)$,

2) $\mathbb{E}[||D^\pi P^\pi - D^{B_k^\pi}P^{B_k^\pi}||_2^2] \leq |\mathcal{S}|^2\frac{t_2^{\text{mix}}}{|B_k^\pi|} + 2|\mathcal{S}|d_{\text{TV}}(\mu_{S_\infty}^\pi, \mu_{S_{k-N}}^\pi)$,

3) $\mathbb{E}\left[\left\|D^\pi R^\pi - D^{B_k^\pi}R^{B_k^\pi}\right\|_2^2\right] \leq |\mathcal{S}|^5|\mathcal{A}|^2R_{\max}^2\frac{t_2^{\text{mix}}}{|B_k^\pi|} + 2|\mathcal{S}|R_{\max}^2 d_{\text{TV}}(\mu_{S_\infty}^\pi, \mu_{S_{k-N}}^\pi)$.

Proof. Let $\{\tilde{S}_k\}_{k\geq -N}$ be the stationary Markov chain defined in Section 3.3. Then, for the first statement, we get the following bounds:

$$
\begin{aligned}
\mathbb{E}[||D^\pi - D^{B_k^\pi}||_2^2] =& \sum_{s\in\mathcal{S}}\mathbb{E}\left[||D^\pi - D^{\tilde{B}_k^\pi}||_2^2\middle|\tilde{S}_{k-N} = s\right]\mathbb{P}[S_{k-N} = s] \\
=& \sum_{s\in\mathcal{S}}\mathbb{E}[||D^\pi - D^{\tilde{B}_k^\pi}||_2^2 \mid \tilde{S}_{k-N} = s]\mathbb{P}[\tilde{S}_{k-N} = s] \\
&+ \sum_{s\in\mathcal{S}}\mathbb{E}[||D^\pi - D^{\tilde{B}_k^\pi}||_2^2 \mid \tilde{S}_{k-N} = s](\mathbb{P}[S_{k-N} = s] - \mathbb{P}[\tilde{S}_{k-N} = s]) \\
\leq& |\mathcal{S}|\frac{t_1^{\text{mix}}}{|B_k^\pi|} + 8d_{\text{TV}}(\mu_{S_\infty}^\pi, \mu_{S_{k-N}}^\pi),
\end{aligned}
$$

where the first equality is due to the law of total expectation, and the fact that $\mathbb{P}[S_{k-N+1}, S_{k-N+2}, \ldots, S_k \mid S_{k-N}]$ and $\mathbb{P}[\tilde{S}_{k-N+1}, \tilde{S}_{k-N+2}, \ldots, \tilde{S}_k \mid \tilde{S}_{k-N}]$ are identical. Moreover, the second equality follows from simple algebraic decomposition, and the inequality is due to the fact that $||A-B||_2^2 \leq 2(||A||_2^2 + ||B||_2^2)$ for $A, B \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, and applies Lemma A.10 and the definition of total variation distance in Definition 2.1. Note that Lemma A.10 will be given after the proof.

For the second statement, it follows that

$$\begin{aligned}
\mathbb{E}\left[\left\|D^\pi P^\pi - D^{B_k} P^{B_k}\right\|_2^2\right] =& \sum_{s\in\mathcal{S}} \mathbb{E}\left[\left\|D^\pi P^\pi - D^{B_k} P^{B_k}\right\|_2^2 \mid S_{k-N} = s\right] \mathbb{P}[S_{k-N} = s] \\
\leq& \mathbb{E}\left[\left\|D^\pi P^\pi - D^{\tilde{B}_k^\pi} P^{\tilde{B}_k^\pi}\right\|_2^2\right] + 2|\mathcal{S}| d_{\mathrm{TV}}(\mu^\pi_{S_\infty}, \mu^\pi_{S_{k-N}}) \\
\leq& |\mathcal{S}|^2 \frac{t_2^{\mathrm{mix}}}{|\tilde{B}_k^\pi|} + 2|\mathcal{S}| d_{\mathrm{TV}}(\mu^\pi_{S_\infty}, \mu^\pi_{S_{k-N}}),
\end{aligned}$$

where the first inequality is due to the fact that $||P^\pi||_2^2 \leq |\mathcal{S}|$, $||P^{\tilde{B}_k^\pi}||_2^2 \leq |\mathcal{S}|$ since P^π and $P^{\tilde{B}_k^\pi}$ are stochastic matrix, i.e., the row sum equals one. This concludes the proof of the second statement.

Next, similar arguments hold for $\mathbb{E}\left[\left\|D^\pi R^\pi - D^{B_k^\pi} R^{B_k^\pi}\right\|_2^2\right]$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\left\|D^\pi R^\pi - D^{B_k^\pi} R^{B_k^\pi}\right\|_2^2\right] =& \sum_{s\in\mathcal{S}} \mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2^2 \middle| S_{k-N} = s\right] \mathbb{P}[S_{k-N} = s] \\
\leq& \mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2^2\right] + 2|\mathcal{S}| R_{\max}^2 d_{\mathrm{TV}}(\mu^\pi_{S_\infty}, \mu^\pi_{S_{k-N}}) \\
\leq& |\mathcal{S}|^5 |\mathcal{A}|^2 R_{\max}^2 \frac{t_2^{\mathrm{mix}}}{|B_k^\pi|} + 2|\mathcal{S}| R_{\max}^2 d_{\mathrm{TV}}(\mu^\pi_{S_\infty}, \mu^\pi_{S_{k-N}}).
\end{aligned}$$

This completes the proof. □

Lemma A.10. *For $k \geq 0$, we have*

1) $\mathbb{E}\left[\left\|D^\pi - D^{\tilde{B}_k^\pi}\right\|_2^2\right] \leq |\mathcal{S}| \frac{t_1^{\mathrm{mix}}}{|\tilde{B}_k^\pi|}$,

2) $\mathbb{E}\left[\left\|D^\pi P^\pi - D^{\tilde{B}_k^\pi} P^{\tilde{B}_k^\pi}\right\|_2^2\right] \leq |\mathcal{S}|^2 \frac{t_2^{\mathrm{mix}}}{|\tilde{B}_k^\pi|}$,

3) $\mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2^2\right] \leq |\mathcal{S}|^5 |\mathcal{A}|^2 R_{\max}^2 \frac{t_2^{\mathrm{mix}}}{|\tilde{B}_k^\pi|}$.

Proof. The first statement is proved via the following inequalities:

$$\mathbb{E}\left[\left\|D^\pi - D^{\tilde{B}_k^\pi}\right\|_2^2\right] \leq \sum_{s\in\mathcal{S}} \mathbb{E}\left[\left(\mu^\pi_{S_\infty}(s) - \mu^{\tilde{B}_k^\pi}_{\tilde{S}}(s)\right)^2\right] \leq |\mathcal{S}| \frac{t_1^{\mathrm{mix}}}{|\tilde{B}_k^\pi|}.$$

Here, the first inequality follows from the matrix norm inequality $||\cdot||_2 \leq ||\cdot||_F$, where $||\cdot||_F$ stands for Frobenius norm, and the second inequality applies Lemma A.3. The proof of the second statement follows similar lines. In particular, letting $\mu^{\tilde{B}_k^\pi}_{\tilde{S},\tilde{S}'}$ stand for stationary distribution of $\{(\tilde{S}_k, \tilde{S}_{k+1})\}_{k\geq -N}$, one gets

$$\mathbb{E}\left[\left\|D^\pi P^\pi - D^{\tilde{B}_k^\pi} P^{\tilde{B}_k^\pi}\right\|_2^2\right] \leq \sum_{(s,r,s')\in\mathcal{O}} \mathbb{E}\left[(\mu^\pi_{S_\infty, S'_\infty}(s,s') - \mu^{\tilde{B}_k^\pi}_{\tilde{S},\tilde{S}'}(s,s'))^2\right] \leq |\mathcal{S}|^2 \frac{t_2^{\mathrm{mix}}}{|\tilde{B}_k^\pi|},$$

where the first inequality follows from the relation between spectral norm and Frobenius norm, and the second inequality applies Lemma A.3.

For the third statement, one can bound $\mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2^2\right]$ with the same logic as in the proof of Lemma 3.3 as follows:

$$\begin{aligned}\mathbb{E}\left[\left\|D^\pi R^\pi - D^{\tilde{B}_k^\pi} R^{\tilde{B}_k^\pi}\right\|_2^2\right] =& \mathbb{E}\left[\sum_{s\in\mathcal{S}}\left(\mathbb{E}\left[r(S,A,S') \mid S=s,\pi\right]\mu^\pi_{S_\infty}(s) - \mu^{\tilde{B}_k^\pi}_{\tilde{S}}(s)[R^{\tilde{B}_k^\pi}]_s\right)^2\right]\\ \leq& |\mathcal{S}|^3|\mathcal{A}|^2 R_{\max}^2 \mathbb{E}\left[\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\left(\mu^\pi_{S_\infty,S'_\infty}(s,s') - \mu^{\tilde{B}_k^\pi}_{\tilde{S},\tilde{S}'}(s,s')\right)^2\right]\\ \leq& |\mathcal{S}|^5|\mathcal{A}|^2 R_{\max}^2 \frac{t_2^{\text{mix}}}{|\tilde{B}_k^\pi|}.\end{aligned}$$

This completes the proof. □

A.9 Proof of Theorem 3.5

Proof. We prove the first statement. Using (6), $||V_{k+1} - V^\pi||_M^2$ can be expanded as follows:

$$\begin{aligned}&||V_{k+1} - V^\pi||_M^2\\ =&(V_k - V^\pi)^\top A^\top M A (V_k - V^\pi) + 2\alpha (V_k - V^\pi)^\top A^\top M w(M_k^\pi, V_k) + \alpha^2 ||w(M_k^\pi, V_k)||_M^2\\ =&||V_k - V^\pi||_M^2 - ||V_k - V^\pi||_2^2 + 2\alpha (V_k - V^\pi)^\top A^\top M w(M_k^\pi, V_k) + \alpha^2 ||w(M_k^\pi, V_k)||_M^2,\end{aligned}$$

where the first equality follows from (6), and the last equality uses the Lyapunov equation in (9). Taking expectation yields

$$\begin{aligned}&\mathbb{E}[||V_{k+1} - V^\pi||_M^2]\\ =&\mathbb{E}[||V_k - V^\pi||_M^2] - \mathbb{E}[||V_k - V^\pi||_2] + 2\alpha\mathbb{E}[(V_k - V^\pi)^\top A^\top M w(M_k^\pi, V_k)] + \alpha^2\mathbb{E}[|w(M_k^\pi, V_k)||_M^2]\\ \leq&\mathbb{E}[||V_k - V^\pi||_M^2] - \mathbb{E}[||V_k - V^\pi||_2] + \frac{8|\mathcal{S}|^{\frac{3}{2}} R_{\max}}{(1-\gamma)^2\mu^\pi_{\min}}\mathbb{E}[||w(M_k^\pi, V_k)||_2]\\ &+ \frac{2\alpha|\mathcal{S}|}{(1-\gamma)\mu^\pi_{\min}}\mathbb{E}[|w(M_k^\pi, V_k)||_2^2]\\ \leq&\mathbb{E}\left[||V_k - V^\pi||_M^2\right] - \mathbb{E}[||V_k - V^\pi||_2^2]\\ &+ \frac{32|\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^3\mu^\pi_{\min}}\left(\sqrt{\frac{8\log(2|\mathcal{S}|)}{L}} + 2|\mathcal{S}|^{\frac{3}{2}}|\mathcal{A}|\sqrt{\frac{\tau^{\text{mix}}}{N}} + 16|\mathcal{S}| d_{\text{TV}}(\mu^\pi_{S_\infty}, \mu^\pi_{S_{k-N}})\right)\\ &+ \alpha\frac{4|\mathcal{S}|^2(R_{\max}+1)^2}{(1-\gamma)^3\mu^\pi_{\min}}\left(\frac{120(\log(2|\mathcal{S}|))^2}{L} + 4|\mathcal{S}|^4|\mathcal{A}|^2\frac{\tau^{\text{mix}}}{N} + 8|\mathcal{S}| d_{\text{TV}}(\mu^\pi_{S_\infty}, \mu^\pi_{S_{k-N}})\right),\end{aligned}$$

where the first inequality follows from Cauchy-Schwartz inequality, and the boundedness of V_k in Lemma 3.1. The last inequality comes from bounding the first and second moment of $\|w(M_k^\pi, V_k)\|_2$ by Lemma 3.3 and Lemma 3.4, respectively.

Summing the last inequality from $k=0$ to $k=T-1$ and dividing both sides by T, one gets

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[||V_k - V^\pi||_2^2] \leq& \frac{1}{T}\mathbb{E}[||V_0 - V^\pi||_M^2]\\ &+ \frac{32|\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^3\mu^\pi_{\min}}\left(\sqrt{\frac{8\log(2|\mathcal{S}|)}{L}} + 2|\mathcal{S}|^{\frac{3}{2}}|\mathcal{A}|\sqrt{\frac{\tau^{\text{mix}}}{N}}\right)\\ &+ \frac{1}{T}\frac{32|\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^3\mu^\pi_{\min}}\left(\sum_{k=0}^{T-1} 16|\mathcal{S}| d_{\text{TV}}(\mu^\pi_{S_\infty}, \mu^\pi_{S_{k-N}})\right)\\ &+ \alpha\frac{4|\mathcal{S}|^2(R_{\max}+1)^2}{(1-\gamma)^3\mu^\pi_{\min}}\left(\frac{120(\log(2|\mathcal{S}|))^2}{L} + 4|\mathcal{S}|^4|\mathcal{A}|^2\frac{\tau^{\text{mix}}}{N}\right)\end{aligned}$$

$$+ \alpha \frac{1}{T} \frac{4|\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^3 \mu_{\min}^{\pi}} \left(\sum_{k=0}^{T-1} 16|\mathcal{S}| d_{\mathrm{TV}}(\mu_{S_\infty}^{\pi}, \mu_{S_{k-N}}^{\pi}) \right).$$

To bound the sum of total variation terms in the above inequality, the following relation can be used:

$$\sum_{k=0}^{T-1} d_{\mathrm{TV}}(\mu_{S_\infty}^{\pi}, \mu_{S_{k-N}}^{\pi}) \leq t_1^{\mathrm{mix}} \sum_{k=0}^{\lfloor T/t_1^{\mathrm{mix}} \rfloor} d_{\mathrm{TV}}(\mu_{S_\infty}^{\pi}, \mu_{S_{kt_1^{\mathrm{mix}}-N}}^{\pi}) \leq t_1^{\mathrm{mix}} \sum_{k=0}^{\lfloor T/t_1^{\mathrm{mix}} \rfloor} 2^{-k} \leq 2t_1^{\mathrm{mix}},$$

where the first inequality follows from non-increasing property of total variation of irreducible and aperiodic Markov chain in Lemma A.4 in Appendix. Moreover, the second inequality follows from the first item in Lemma A.4 in Appendix. Combining the last two inequalities leads to

$$\begin{aligned}
&\frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E}[||V_k - V^\pi||_2^2] \\
\leq &\frac{1}{T} \frac{2|\mathcal{S}|}{\alpha(1-\gamma)} \mathbb{E}[||V_0 - V^\pi||_2^2] \\
&+ \frac{32|\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^3 \mu_{\min}^{\pi}} \left(\sqrt{\frac{\log(8|\mathcal{S}|)}{L}} + 2|\mathcal{S}|^{\frac{3}{2}} |\mathcal{A}| \sqrt{\frac{\tau^{\mathrm{mix}}}{N}} \right) + \frac{32^2 |\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^3 \mu_{\min}^{\pi}} \frac{2t_1^{\mathrm{mix}}}{T} \\
&+ \alpha \frac{4|\mathcal{S}|^2 (R_{\max}+1)^2}{(1-\gamma)^3 \mu_{\min}^{\pi}} \left(\frac{120(\log(2|\mathcal{S}|))^2}{L} + 4|\mathcal{S}|^4 |\mathcal{A}|^2 \frac{\tau^{\mathrm{mix}}}{N} \right) + \alpha \frac{128|\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^3 \mu_{\min}^{\pi}} \frac{t_1^{\mathrm{mix}}}{T}.
\end{aligned}$$

This completes the proof of first statement. The second statement can be directly obtained from applying Jensen's inequality.

□

A.10 Proof of Theorem 3.6

Proof. First of all, we prove the first statement. The term $\|V_{k+1} - V^\pi\|_2^2$ is expanded according to the update in (6) as follows:

$$\begin{aligned}
&\|V_{k+1} - V^\pi\|_2^2 \\
=&||V_k - V^\pi||_{A^\top A}^2 + 2\alpha(V_k - V^\pi)A^\top w(M_k^\pi, V_k) + \alpha^2 ||w(M_k^\pi, V_k)||_2^2 \\
=&(V_0 - V^\pi)^\top (A^\top)^{k+1} A^{k+1} (V_0 - V^\pi) + \sum_{i=0}^{k} 2\alpha (V_i - V^\pi)^\top A^\top (A^\top)^{k-i} A^{k-i} w(M_i^\pi, V_i) \qquad (23) \\
&+ \alpha^2 \sum_{i=0}^{k} w(M_i^\pi; V_i)^\top (A^\top)^{k-i} A^{k-i} w(M_i^\pi; V_i), \qquad (24)
\end{aligned}$$

where the second equality follows from the fact that for $0 \leq i \leq k-1$, the i-th expansion results to weighted inner product of $A(V_{k-i-1} - V^\pi)$ and $w(M_{k-i-1}^\pi, V_{k-i-1})$ and weighted squared norm of $w(M_{k-i-1}^\pi, V_{k-i-1})$ which are (23) and (24), respectively.

Next, applying Cauchy–Schwartz inequality to (24) yields

$$\begin{aligned}
&\|V_{k+1} - V^\pi\|_2^2 \\
\leq &||V_0 - V^\pi||_2^2 |\mathcal{S}| ||A^{2k+2}||_\infty + 2\alpha|\mathcal{S}| \sum_{i=0}^{k} ||V_i - V^\pi||_2 ||A^{k-i+1}||_\infty ||A^{k-i}||_\infty ||w(M_i^\pi; V_i)||_2 \\
&+ \alpha^2 \sum_{i=0}^{k} ||A^{2k-2i}||_2 ||w(M_i^\pi; V_i)||_2^2
\end{aligned}$$

$$
\begin{aligned}
&\leq ||V_0 - V^\pi||_2^2 |\mathcal{S}| ||A||_\infty^{2k+2} + 2\alpha|\mathcal{S}| \sum_{i=0}^{k} ||V_i - V^\pi||_2 ||A||_\infty^{2k-2i+1} ||w(M_i^\pi; V_i)||_2 \\
&\quad + \alpha^2 \sqrt{|\mathcal{S}|} \sum_{i=0}^{k} ||A||_\infty^{2k-2i} ||w(M_i^\pi; V_i)||_2^2,
\end{aligned}
$$

where the first inequality follows from the matrix norm inequality $||A||_2 \leq \sqrt{|\mathcal{S}|}||A||_\infty$, and the last inequality is due to sub-multiplicativity of the matrix norm. After taking expectation, we get

$$
\begin{aligned}
&\mathbb{E}\left[\|V_{k+1} - V^\pi\|_2^2\right] \\
&\leq ||V_0 - V^\pi||_2^2 \|\mathcal{S}|||A^{k+1}||_2^2 \\
&\quad + \alpha \frac{64|\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^2} \left(\sqrt{\frac{8\log(2|\mathcal{S}|)}{L}} + |\mathcal{S}|^{\frac{3}{2}} |\mathcal{A}| \sqrt{\frac{\tau^{\text{mix}}}{N}} \right) \sum_{i=0}^{k} ||A||_\infty^{2k-2i+1} \\
&\quad + \alpha^2 \sqrt{|\mathcal{S}} \frac{4|\mathcal{S}|(R_{\max}+1)^2}{(1-\gamma)^2} \left(\frac{120(\log(2|\mathcal{S}|)^2}{|M_k^\pi|} + 4|\mathcal{S}|^4 |\mathcal{A}|^2 \frac{\tau^{\text{mix}}}{|B_k^\pi|} \right) \sum_{i=0}^{k} ||A||_\infty^{2k-2i} \\
&\leq ||V_0 - V^\pi||_2^2 \|\mathcal{S}|(1 - \alpha(1-\gamma)\mu_{\min}^\pi)^{2k+2} \\
&\quad + \frac{64|\mathcal{S}|^2 R_{\max}^2}{(1-\gamma)^3 \mu_{\min}^\pi} \left(\sqrt{\frac{8\log(2|\mathcal{S}|)}{L}} + |\mathcal{S}|^{\frac{3}{2}} |\mathcal{A}| \sqrt{\frac{\tau^{\text{mix}}}{N}} \right) \\
&\quad + \alpha \frac{4|\mathcal{S}|(R_{\max}+1)^2}{(1-\gamma)^3 \mu_{\min}^\pi} \left(\frac{120(\log(2|\mathcal{S}|)^2}{|M_k^\pi|} + 4|\mathcal{S}|^4 |\mathcal{A}|^2 \frac{\tau^{\text{mix}}}{|B_k^\pi|} \right),
\end{aligned}
$$

where the first inequality follows from Lemma 3.3, 3.4, and the fact that $d_{\text{TV}}(\mu_{S_\infty}^\pi, \mu_{S_{k-N}}^\pi)$ is zero for all $k \geq 0$ because the initial state distribution is identical to the stationary distribution. Moreover, the second inequality comes from the bound on $||A||_\infty$ in Lemma 3.2. This proves the first statement. Finally, applying Jensen's inequality to the first item which is the bound on $\mathbb{E}\left[\|V_k - V^\pi\|_2^2\right]$, leads to the second statement. □

A.11 Sample complexity result of Theorem 3.6

The result of Bhandari et al. (2018) requires following number of samples to bound (16) with $\epsilon > 0$:

$$
\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2} \frac{t^{\text{mix}}}{(1-\gamma)^3 (\mu_{\min}^\pi)^2} \right), \tag{25}
$$

where we used $t^{\text{mix}}(\alpha) = \tilde{\mathcal{O}}\left(t^{\text{mix}}\right)$ from Lemma A.5 in the Appendix, and $t^{\text{mix}}(\alpha)$ is defined in (1).

To bound the second line of the the second item in Theorem 3.6 with ϵ, we require

$$
L = \tilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^4}{(1-\gamma)^3 (\mu_{\min}^\pi)^2} \frac{1}{\epsilon^4} \right), \quad N = \tilde{\mathcal{O}}\left(\frac{\tau_{\text{mix}} |\mathcal{S}|^7 |\mathcal{A}|^2}{(1-\gamma)^6 (\mu_{\min}^\pi)^4} \frac{1}{\epsilon^4} \right).
$$

Next, to bound the third line of the the second item in Theorem 3.6 with ϵ, we require

$$
L = \tilde{\mathcal{O}}\left(\frac{|\mathcal{S}|}{(1-\gamma)^4 \mu_{\min}^\pi} \frac{1}{\alpha} \frac{1}{\epsilon^2} \right), \quad N = \tilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^5 |\mathcal{A}|^2 \tau^{\text{mix}}}{(1-\gamma)^3 \mu_{\min}^\pi} \frac{1}{\alpha} \frac{1}{\epsilon^2} \right).
$$

Lastly, to bound the first line of the the second item in Theorem 3.6 with ϵ, the number of steps k needs to satisfy the following:

$$
k = \mathcal{O}\left(\frac{1}{\alpha(1-\gamma)\mu_{\min}^\pi} \ln\left(\frac{1}{\epsilon}\right) \right).
$$

The overall sample complexity results to

$$k + N = \tilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^7|\mathcal{A}|^2\tau^{\text{mix}}}{(1-\gamma)^3\mu^{\pi}_{\min}}\frac{1}{\alpha}\frac{1}{\epsilon^4}\right).$$

Compared to the sample complexity in (25) by Bhandari et al. (2018), the above sample complexity result leaves room for improvement in terms of $\frac{1}{1-\gamma}$ and $\frac{1}{\mu^{\pi}_{\min}}$. However, if we consider, N, as given pre-collected dataset, the sample complexity result is only $\mathcal{O}\left(\frac{1}{\alpha(1-\gamma)\mu^{\pi}_{\min}}\ln\left(\frac{1}{\epsilon}\right)\right)$, where $\alpha \in (0,1)$.