Analysis of Temporal Difference Learning: Linear System Approach

Donghwan Lee and Do Wan Kim

Abstract—The goal of this technical note is to introduce a new finite-time convergence analysis of temporal difference (TD) learning based on stochastic linear system models. TD-learning is a fundamental reinforcement learning (RL) to evaluate a given policy by estimating the corresponding value function for a Markov decision process. While there has been a series of successful works in theoretical analysis of TD-learning, it was not until recently that researchers found some guarantees on its statistical efficiency by developing finite-time error bounds. In this paper, we propose a simple control theoretic finite-time analysis of TD-learning, which exploits linear system models and standard notions in linear system communities. The proposed work provides new simple templates for RL analysis, and additional insights on TD-learning and RL based on ideas in control theory.

Index Terms—Reinforcement learning, TD-learning, linear time-invariant (LTI) system, control theory, finite-time analysis

I. INTRODUCTION

Proposed by [1], temporal difference learning (TD-learning) is a fundamental reinforcement learning (RL) [2] to estimate the value function of a given policy for a Markov decision process (MDP) [3]. The idea has been applied to many more advanced algorithms such as classical Q-learning [4], SARSA [5], actor-critic [6], and more contemporary RLs such as deep Q-learning [7], double Q-learning [8], gradient TD-learning [9], [10], and deterministic actor-critic [11] to name just a few. There has been a series of successful works in theoretical analysis of TD-learning such as [12]–[14]. However, these classical analysis approaches usually focus on the asymptotic behavior of TD-learning. It was not until recently that researchers found some guarantees on its statistical efficiency [15]–[18]. In particular, these recent advances investigate how fast the TD iterates converge to the desired solution, which are expressed as convergent error bounds depending on time-steps. Such analysis is called a finite-time analysis.

A. Contribution

This technical note investigates a simple and unique control theoretic finite-time analysis TD-learning, which exploits linear system models and standard notions in linear system literatures [19]. We provide both mean-squared error bounds of the final and averaged iterates together with sample complexities and probabilistic error bounds based on concentration inequalities. The approaches for the final iterate and averaged iterate are different with duality relations. In particular, the core idea for the analysis of the final iterate is to propagate the correlation of the linear system states. On the other hand, the averaged iterate is based on the Lyapunov theory [20]. The new analysis is applied to the tabular settings, but can be extended to the linear function approximation scenario with additional efforts. On the other hand, it can cover off-policy scenarios, which means that the target policy to evaluate can be different from the behavior policy to collect experiences. Moreover, the finite-time error bounds have different features compared to the previous works [15]–[18] with some benefits. Our new analysis reveals clear connections between TD-learning and notions in linear systems, and provides additional insights on TD-learning and RL with simple concepts and analysis tools in control theory. Finally, we note that this paper only covers an i.i.d. observation model with a constant step-size for simplicity of the overall analysis. Extensions to more complicated scenarios are not the main purpose of this technical note.

B. Related Works

1) Finite-Time Analysis of TD-Learning: Recently, some progress have been made in finite-time analysis of TD-learning algorithms [15]–[18]. The recent work [18] studies a finite-time analysis of TD-learning with linear function approximation and an i.i.d. observation model. They focus on analysis with the problem independent diminishing step-sizes of the form $1/k^\sigma$ for a fixed $\sigma \in (0, 1)$, where $k$ is the iteration step, and establish that mean-squared error convergence at a rate of $O(1/k^\sigma)$. The work [16] investigates general linear stochastic approximation algorithms, including TD-learning, with i.i.d. observation model, constant step-sizes, and iterate averaging. As for the TD-learning, their work provides $O(1/k)$ bounds on the mean-squared error. The paper [15] develops a simple and explicit finite-time analysis of TD-learning with linear function approximation, and provides analysis with both i.i.d. observation model and Markovian observation model. The analysis is based on properties mirroring those of stochastic gradient descent methods. The analysis in [17] studies TD-learning with Markovian observation model and constant step-sizes by considering the drift of an appropriately chosen Lyapunov function. The Lyapunov function is a standard Lyapunov function used to study the stability of continuous-time linear ordinary differential equations.

Compared to these previous results, this technical note provides a finite-time analysis of TD-learning from an entirely different and simple viewpoint. We view the TD-learning update as a discrete-time linear system dynamics with stochastic noises, and analyze it using tools in linear system theories. The proposed mean-squared error bounds have different features compared to the previous works, and cover different cases detailed throughout this technical note. All the proofs in this note are simple with few lines. The proposed frameworks provide additional insights on TD-learning and RL, and complements existing methods.

2) Control System Analysis of RL: It is worth mentioning recent works on analysis of RLs based on control system frameworks [21]–[23]. Dynamical system perspectives of reinforcement learning and general stochastic iterative algorithms have a long tradition, which dates back to O.D.E analysis [13], [24]–[26]. More recently, [22] investigated asymptotic convergence TD-learning based on a Markovian jump linear systems (MJLSs). They tailor the MJLS theory developed in the control community to characterize the exact behaviors of the first and second order moments of a large family of TD-learning algorithms. The analysis in [22] includes both asymptotic and finite-time natures, while some parameters in the bounds are not explicit. The paper [21] studies asymptotic convergence of Q-learning [4] through a continuous-time switched linear system model [27], [28].
A finite-time analysis of Q-learning is also investigated in [23] using discrete-time switched linear system model. Lastly, a conference version of this technical note has been submitted to CDC2022, which only includes the averaged-iterate cases. Significant progresses have been made after the conference submission such as the correlation analysis and the final iterate convergence analysis.

C. Notation

The adopted notation is as follows: \( \mathbb{R} \): set of real numbers; \( \mathbb{R}^n \): n-dimensional Euclidean space; \( \mathbb{R}^{n \times m} \): set of all \( n \times m \) real matrices; \( A^T \): transpose of matrix \( A \); \( A \succ 0 \) (\( A \preceq 0 \), \( A \succeq 0 \), \( A \preceq 0 \), respectively); symmetric positive definite (negative definite, positive semi-definite, and negative semi-definite, respectively) matrix \( A \); \( I \): identity matrix with appropriate dimensions; \( \rho (\cdot) \): spectral radius; for any matrix \( A \), \( [A]_{ij} \) is the element of \( A \) in \( i \)-th row and \( j \)-th column; \( \lambda_{\text{min}} (A) \) and \( \lambda_{\text{max}} (A) \) for any symmetric matrix \( A \); the minimum and maximum eigenvalues of \( A \); \( |S| \): cardinality of a finite set \( S \); \( \triangledown (A) \): trace of any matrix \( A \).

II. PRELIMINARIES

A. Markov decision problem

We consider the infinite-horizon discounted Markov decision problem (MDP) [3], [29], where the agent sequentially takes actions to maximize cumulative discounted rewards. In a Markov decision process with the state-space \( S := \{1, 2, \ldots, |S|\} \) and action-space \( A := \{1, 2, \ldots, |A|\} \), the decision maker selects an action \( a \in A \) with the current state \( s \), then the state transits to a state \( s' \) with probability \( P(s'|s, a) \), and the transition incurs a reward \( r(s, a, s') \). For convenience, we consider a deterministic reward function and simply write \( r(s, a, s') := r \), \( k \in \{0, 1, \ldots \} \). A (stochastic) policy is a map \( \pi: S \times A \rightarrow [0, 1] \) representing the probability, \( \pi(a|s) \), of selecting action \( a \) at the current state \( s \). The objective of the Markov decision problem (MDP) is to find a deterministic optimal policy, \( \pi^* \), such that the cumulative discounted rewards over infinite time horizons is maximized, i.e., \( \pi^* := \arg \max_{\pi \in \Theta} \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k r_k \right] \), where \( \gamma \in [0, 1) \) is the discount factor, \( \Theta \) is the set of all admissible deterministic policies, \((s_0, a_0, s_1, a_1, \ldots)\) is a state-action trajectory generated by the Markov chain under policy \( \pi \), and \( \mathbb{E} \{ \cdot | s \} \) is an expectation conditioned on the policy \( \pi \). The value function under policy \( \pi \) is defined as

\[
V^\pi (s) = \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k r_k | s_0 = s, \pi \right], \quad s \in S,
\]

Based on these notions, the policy evaluation problem is defined as follows.

**Definition 1** (Policy evaluation problem). Given a policy \( \pi \), find the corresponding value function \( V^\pi \).

The policy evaluation problem is an important component of policy optimization problems for the Markov decision problem. The (model-free) policy evaluation problem is defined as follows: given a policy \( \pi \), find the corresponding value function \( V^\pi \) without the model knowledge, i.e., \( P \), only using experiences or transitions \((s, a, r, s')\). In the policy evaluation problem, the policy we want to evaluate is called a target policy. On the other hand, the behavior policy, denoted by \( b \), is the policy that is used to generate experiences. For a learning algorithm, if \( b = \pi \), it is called on-policy learning. Otherwise, it is called off-policy learning. The stationary state distribution, if exists, is defined as

\[
\lim_{k \to \infty} P(s_k = s|b) := d(s), \quad s \in S,
\]

where \( b \) is any behavior policy. Throughout, we assume that the stationary state distribution exists, which is a standard assumption.

B. TD-Learning

We consider a version of TD-learning given in Algorithm 1. Compared to the original TD-learning, the step-size \( \alpha \) is constant.

**Algorithm 1** Off-policy TD-learning

1. Set the step-size sequence \( \{\alpha_k\}_{k=0}^{\infty} \).
2. Initialize \( V_0 \in \mathbb{R}^{|S|} \) arbitrarily such that \( \|V_0\|_\infty \leq 1 \).
3. for iteration \( k = 0, 1, \ldots \) do
   4. Observe \( s_k \sim d, a_k \sim b(s_k) \), \( s'_k \sim P(s_k, a_k, \cdot) \) and \( r_k = \pi(s_k, a_k, s'_k) \)
   5. Update \( V_{k+1}(s_k) = V_k(s_k) + \alpha_k \{\pi(s_k, a_k) r_k + \gamma \pi(s_k, a_k) V_k(s'_k) - V_k(s_k)\} \) where \( \pi(s, a) := \frac{\pi(a|s)}{d(s)} \).
6. end for

In this paper, we assume that \( m = 1 \). Note that an importance sampling ratio, \( \pi(s_k, a_k) := \frac{\pi(a_k|s_k)}{b(a_k|s_k)} \), is introduced for off-policy learning [30]. In other words, Algorithm 1 is an off-policy TD-learning. Note also that if \( \pi = b \), then \( \sigma \equiv 1 \). We make the following standard assumptions throughout the paper.

**Assumption 1.**

1. The step-size \( \alpha \) satisfies \( \alpha \in (0, 1) \).
2. (Positive stationary state distribution) \( d(s) > 0 \) holds for all \( s \in S \).
3. (Bounded rewards) The reward is bounded as follows:

   \[
   \mathbb{E} \{r(s, a, s')|s\} := R_{\max} \leq 1.
   \]

4. (Bounded initial parameter) The initial iterate \( V_0 \) satisfies \( \|V_0\|_\infty \leq 1 \).
5. (Bounded importance sampling ratio)

   \[
   \sigma_{\max} := \mathbb{E} \{\pi(s, a)|s, a\} / b(a|s) \in [1, \infty).
   \]

The first statement in Assumption 1 is required to guarantee the convergence of the proposed results. The second statement is standard, and guarantees that every state may be visited infinitely often for sufficient exploration. The third statement is standard, and is required to ensure the boundedness of the iterates of TD-learning. The bound imposed on \( V_0 \) in the fourth item is just for simplicity of analysis. It is introduced without loss of generality. Before closing this subsection, the boundedness of TD-learning iterates [31] is introduced, which plays an important role in our analysis.

**Lemma 1** (Boundedness of TD-learning iterates [31]). If the step-size is less than one in Algorithm 1, then for all \( k \geq 0 \),

\[
\|V_k\|_\infty \leq V_{\max} := \max \{R_{\max}, \max_{s \in S} V_0(s)\} / (1 - \gamma).
\]

From Assumption 1, \( R_{\max} \leq 1 \) and \( \|V_0\|_\infty \leq 1 \), one can easily see that \( V_{\max} \leq 1 / (1 - \gamma) \).

III. LINEAR SYSTEM MODEL

In this subsection, we study a discrete-time linear system model of [30], and establish its finite-time convergence based on the stability analysis of linear systems. The following notations will be frequently used throughout this paper; hence, we define them in advance for convenience.

**Definition 2.**

1. Maximum state-action visit probability:

   \[
   d_{\max} := \max_{s \in S} d(s) \in (0, 1).
   \]
2) Minimum state-action visit probability:
\[ d_{\text{min}} := \min_{s \in S} d(s) \in (0, 1). \]

3) Exponential convergence rate:
\[ \rho := 1 - \alpha d_{\text{min}} (1 - \gamma) \in (0, 1). \]

4) The diagonal matrix \( D \) is defined as
\[ D := \begin{bmatrix} d(1) & \cdots & d(|S|) \end{bmatrix} \in \mathbb{R}^{|S| \times |S|}, \]

5) \( P^\pi \in \mathbb{R}^{|S| \times |S|} \) is a matrix such that \( [P^\pi]_{ij} = \mathbb{P}[s' = j | s = i, \pi] \)
6) \( R^\pi \in \mathbb{R}^{|S|} \) is a vector such that \( [R^\pi]_i = \mathbb{E}[r(s_k, a_k, s_{k+1}) | a_k \sim \pi(s_k), s_k = i] \)

Note that \( P^\pi \) and \( R^\pi \) can be expressed in terms of the importance sampling ratio \( \rho(s, a) \)
\[ P^\pi(s'|s) := \sum_{a \in A} P(s'|s, a) \rho(s, a), \]
\[ R^\pi(s) := \sum_{a \in A} \pi(a) P(s'|s, a) r(s, a, s') = \sum_{a \in A} \sum_{s' \in S} b(a|s) P(s'|s, a) \rho(a|s) \]
where \( P^\pi(s'|s) \) implies the probability that the current state \( s \) transits to \( s' \) under the policy \( \pi \), and \( P(s'|s, a) \) means the probability that the current state \( s \) transits to \( s' \) under the action \( a \) similarly.
\[ R^\pi(s) := \sum_{a \in A} \pi(a) P(s'|s, a) r(s, a, s') = \sum_{a \in A} \sum_{s' \in S} b(a|s) P(s'|s, a) \rho(a|s). \]

Using the notation introduced, the update in Algorithm 1 can be equivalently rewritten as
\[ V_{k+1} = V_k + \alpha \{ DR^\pi + \gamma D P^\pi V_k - DV_k + w_k \}, \]
where \( w_k := e_{s_k} \delta_k - (DR^\pi + \gamma D P^\pi V_k - DV_k) \),
\[ \delta_k := e_{s_k} \gamma \sigma(s_k, a_k) r_k + \gamma \sigma(s_k, a_k) e_{s_{k+1}}^T V_k - e_{s_k}^T V_k \],
and \( e_s \in \mathbb{R}^{|S|} \) is the \( s \)-th basis vector (all components are 0 except for the \( s \)-th component which is 1). Here, \( (s_k, a_k, r_k, s'_{k}) \) is the sample transition in the \( k \)-th time-step. The expressions can be equivalently reformulated as the linear system
\[ V_{k+1} = (I + \alpha(\gamma D P^\pi - D) V^\pi + DR^\pi) V_k + \alpha DR^\pi + \alpha w_k. \]

Invoking the optimal Bellman equation \( (\gamma D P^\pi - D) V^\pi + DR^\pi = 0 \) leads to the equivalent equation
\[ V_{k+1} = (I + \alpha(\gamma D P^\pi - D)) (V_k - V^\pi) + \alpha w_k \]
(6)
Defining \( x_k := V_k - V^\pi \) and \( A := I + \alpha(\gamma D P^\pi - D) \), the TD-learning iteration in Algorithm 1 can be concisely represented as the stochastic linear system
\[ x_{k+1} = Ax_k + \alpha w_k, \quad x_0 \in \mathbb{R}^n, \quad \forall k \geq 0. \]
(7)
where \( \pi := |S| \), and \( w_k \in \mathbb{R}^n \) is a stochastic noise. In the remaining parts of this section, we focus on some properties of the system. The first important property is that the noise \( w_k \) has the zero mean, and is bounded. It is formally proved in the following lemma.

**Lemma 2.** We have

1) \( \mathbb{E}[w_k] = 0 \);
2) \( \mathbb{E}[\|w_k\|_\infty] \leq \sqrt{W_{\max}} \);
3) \( \mathbb{E}[\|w_k\|_2] \leq \sqrt{W_{\max}} \);
4) \( \mathbb{E}[w_k^T w_k] \leq \frac{\alpha^2 \max_{i, j} \{D_i\}}{(1-\gamma)^2} := W_{\max} \)
for all \( k \geq 0 \).

**Proof.** For the first statement, we take the conditional expectation on (4) to have \( \mathbb{E}[w_k | x_k] = 0 \). Taking the total expectation again with the law of total expectation leads to the first conclusion. Moreover, the conditional expectation, \( E[w_k^T w_k | V_k] \), is bounded as
\[ \mathbb{E}[w_k^T w_k | V_k] = \mathbb{E}[\|w_k\|_2^2 | V_k] \leq \mathbb{E}[\|DR^\pi + \gamma D P^\pi V_k - DV_k\|_2^2 | V_k] \]
\[ \leq \mathbb{E}[\|\delta_k\|_2^2 | V_k] \]
\[ = \mathbb{E}[\|\sigma(s_k, a_k) e_{s_{k+1}}^T V_k + 2r_k \gamma \sigma(s_k, a_k) e_{s_k}^T V_k - e_{s_k}^T V_k\|_2^2 | V_k] \]
\[ \leq \mathbb{E}[(\gamma^2 \sigma(s_k, a_k) e_{s_{k+1}}^T V_k e_{s_{k+1}}^T V_k) | V_k] \]
\[ + \mathbb{E}[(2 \sigma(s_k, a_k) e_{s_k}^T V_k e_{s_k}^T V_k) | V_k] \]
\[ \leq 2 \mathbb{E}[(\gamma^2 \sigma(s_k, a_k) e_{s_{k+1}}^T V_k e_{s_{k+1}}^T V_k) | V_k] \]
\[ + \mathbb{E}[(2 \sigma(s_k, a_k) e_{s_k}^T V_k e_{s_k}^T V_k) | V_k] \]
\[ \leq \frac{\alpha \sigma_{\max}^2}{(1-\gamma)^2} := W_{\max}, \]
where \( \delta_k \) is defined in (5), and the last inequality comes from Assumption 1 and Lemma 1. Taking the total expectation, we have the fourth result. Next, taking the square root on both sides of \( \mathbb{E}[\|w_k\|_2^2] \leq W_{\max} \), one gets
\[ \mathbb{E}[\|w_k\|_\infty] \leq \mathbb{E}[\|w_k\|_2] \leq \sqrt{\mathbb{E}[\|w_k\|_2^2]} \leq \sqrt{W_{\max}}, \]
where the first inequality comes from \( \| \cdot \|_\infty \leq \| \cdot \|_2 \). This completes the proof.

To proceed further, let us define the covariance of the noise
\[ E[w_k^T w_k] := W_k = W_{k}^T \geq 0. \]
An important quantity we use in the main result is the maximum eigenvalue, \( \lambda_{\max}(W) \), whose bound can be easily established as follows.

**Lemma 3.** The maximum eigenvalue of \( W \) is bounded as
\[ \lambda_{\max}(W_k) \leq W_{\max}, \quad \forall k \geq 0, \]
where \( W_{\max} > 0 \) is given in Lemma 2.

**Proof.** The proof is completed by noting \( \lambda_{\max}(W_k) \leq \text{tr}(W_k) = \text{tr}(E[w_k^T w_k]) = E[\text{tr}(w_k^T w_k^T)] = E[w_k^T w_k] \leq W_{\max} \), where the inequality comes from Lemma 2.

Lastly, we investigate the property of the system matrix \( A \). We establish the fact that the \( \infty \)-norm of \( A \) is strictly less than one, in particular, is bounded by \( \rho \in (0, 1) \), where \( \rho \) is defined in Definition 2.
Lemma 4. \( \| A \|_\infty \leq \rho \) holds, where the matrix norm \( \| A \|_\infty := \max_{1 \leq i \leq n} \sum_{j=1}^{n} |A_{ij}| \) and \( A_{ij} \) is the element of \( A \) in \( i \)-th row and \( j \)-th column.

Proof. Noting that \( A = I + \alpha (\gamma DP^n - D) \), we have

\[
\sum_{j} |A_{ij}| = \sum_{j} |[I - \alpha D + \alpha \gamma DP^n]_{ij}|
\]

\[
= |[I - \alpha D]_{ij} + \sum_{j} |\alpha \gamma DP^n|_{ij}|
\]

\[
= 1 - \alpha |D|_{ij} + \alpha \gamma |D|_{ij} \sum_{j} |P^n|_{ij}
\]

\[
= 1 - \alpha |D|_{ij} + \alpha \gamma |D|_{ij}
\]

\[
= 1 + \alpha |D|_{ij} (\gamma - 1),
\]

where the first line is due to the fact that \( A \) is a positive matrix. Taking the maximum over \( i \), we have

\[
\| A \|_\infty = \max_{i \in \{1, 2, \ldots, n\}} \{ 1 + \alpha |D|_{ij} (\gamma - 1) \}
\]

\[
= 1 - \alpha \min_{s \in S} d(s) (1 - \gamma),
\]

which completes the proof.

Lemma 4 plays an important role in proving the convergence of the linear system (7) or equivalently, the convergence of TD-learning in Algorithm 1. In the next section, we provide the main result, a finite-time convergence analysis of Algorithm 1.

IV. CONVERGENCE ANALYSIS: FINAL ITERATION

In this section, we prove the convergence of the TD-learning in Algorithm 1 based on the linear system model in (7), which is done by analyzing the propagation of both mean and correlation of the state \( x_k \). First of all, taking the mean on (6) leads to

\[
\mathbb{E}[x_{k+1}] = \mathbb{E}[x_k] + \mathbb{E} [\alpha D x_k], \quad x_0 \in \mathbb{R}^n, \quad \forall k \geq 0. \tag{8}
\]

Therefore, the mean state \( \mathbb{E}[x_k] \) follows the linear system dynamics. Using Lemma 4, we can establish the exponential convergence of the mean dynamics driven by (8).

Lemma 5. For all \( k \geq 0 \), \( \| \mathbb{E}[x_k] \|_\infty \) is bounded as

\[
\| \mathbb{E}[x_k] \|_\infty \leq \rho^k \| x_0 \|_\infty, \quad \forall x_0 \in \mathbb{R}^n.
\]

Proof. Taking the norm on (8) leads to \( \| \mathbb{E}[x_{k+1}] \|_\infty = \| \mathbb{E}[x_k] \|_\infty \leq \| A \|_\infty \| \mathbb{E}[x_k] \|_\infty \leq \rho \| \mathbb{E}[x_k] \|_\infty \), where the last inequality is due to Lemma 4. Recursively applying the inequality yields the desired conclusion.

As a next step, we investigate how the correlation, \( \mathbb{E}[x_k x_k^T] \), propagates over the time. In particular, the correlation is updated through the recursion

\[
\mathbb{E}[x_{k+1} x_{k+1}^T] = \mathbb{E}[x_k x_k^T] A^T + \alpha^2 W_k,
\]

where \( \mathbb{E}[w_k w_k^T] = W_k \). Defining \( X_k := \mathbb{E}[x_k x_k^T] \), \( k \geq 0 \), it is equivalently written as

\[
X_{k+1} = AX_k A^T + \alpha^2 W_k, \quad \forall k \geq 0,
\]

with \( X_0 := x_0 x_0^T \). A natural question is whether or not the iterate, \( X_k \), converges as \( k \to \infty \). We can at least prove that \( X_k \) bounded.

Lemma 6 (Boundedness). The iterate, \( X_k \), is bounded as

\[
\| X_k \|_2 \leq \frac{\alpha^2 W_{\max}}{1 - \rho^2} + n \| x_0 \|_2
\]

The proof is given in Appendix A. Similarly, the following lemma proves that the trace of \( X_k \) is bounded, which will be used for the main development.

Lemma 7. We have the following bound:

\[
\text{tr}(X_k) \leq \frac{36 \sigma^2 \max n^2 \alpha}{d_{\min}^2 (1 - \gamma)^2} + n^2 \rho^{2k}
\]

Proof. We first bound \( \lambda_{\max}(X_k) \) as follows:

\[
\lambda_{\max}(X_k) \leq \alpha^2 \sum_{i=0}^{k-1} \lambda_{\max}(A^i W_{k-i-1} (A^T)^i) + \lambda_{\max}(A^k X_0 (A^T)^k)
\]

\[
\leq \alpha^2 \sum_{j=0}^{k-1} \lambda_{\max}(W_j) \sum_{i=0}^{k-1} \lambda_{\max}(A^i (A^T)^i)
\]

\[
+ \lambda_{\max}(X_0) \lambda_{\max}(A^k (A^T)^k)
\]

\[
= \alpha^2 \sum_{j=0}^{k-1} \lambda_{\max}(W_j) \sum_{i=0}^{k-1} \| A^i \|_2^2 + \lambda_{\max}(X_0) \| A^k \|_2^2
\]

\[
\leq \alpha^2 \| W_{\max} \|_n \sum_{i=0}^{k-1} \| A^i \|_\infty^2 + n \lambda_{\max}(X_0) \| A^k \|_2^2
\]

\[
\leq \alpha^2 W_{\max} n \sum_{i=0}^{k-1} \rho^{2i} + n \lambda_{\max}(X_0) \rho^{2k}
\]

\[
\leq \alpha^2 W_{\max} n \frac{1 - \rho^2}{1 - \rho} + n \lambda_{\max}(X_0) \rho^{2k}
\]

where the first inequality is due to \( A^i W_{k-i-1} (A^T)^i \geq 0 \) and \( A^k X_0 (A^T)^k \geq 0 \), and the sixth inequality is due to \( \rho \in (0, 1) \). On the other hand, since \( X_k \geq 0 \), the diagonal elements are nonnegative. Therefore, we have \( \text{tr}(X_k) \leq n \lambda_{\max}(X_k) \). Combining the last two inequalities lead to

\[
\text{tr}(X_k) \leq n \lambda_{\max}(X_k) \leq \frac{\lambda_{\max}(W)n^2 \alpha^2}{1 - \rho^2} + n \lambda_{\max}(X_0) n^2 \rho^{2k}
\]

Moreover, \( \lambda_{\max}(X_0) \leq \text{tr}(X_0) = \text{tr}(x_0 x_0^T) = \| x_0 \|_2^2 \). Plugging \( \rho = 1 - \alpha d_{\min} (1 - \gamma) \), using the bound on \( \lambda_{\max}(W) \) in Lemma 3, and the last bound on \( \lambda_{\max}(X_0) \), one gets

\[
\text{tr}(X_k) \leq \frac{36 \sigma^2 \max n^2 \alpha}{d_{\min}^2 (1 - \gamma)^2} + \| x_0 \|_2^2 n^2 \rho^{2k}.
\]

This completes the proof.

Now, we are ready to present the main results. In the first result, we provide a finite-time bound on the mean-squared error.

Theorem 1. For any \( k \geq 0 \), we have

\[
\mathbb{E}[\| V_k - V^* \|_2^2] \leq \frac{6 \sigma_{\max} |S \sqrt{\alpha}|}{d_{\min}^2 (1 - \gamma)^{1.5}} + \| V_0 - V^* \|_2^2 |S|^k.
\]

Proof. Noting the relations

\[
\mathbb{E}[\| V_k - V^* \|_2^2]
\]

\[
= \mathbb{E}[\| (V_k - V^*)^T (V_k - V^*) \|_2^2]
\]

\[
= \mathbb{E}[\text{tr}((V_k - V^*)^T (V_k - V^*)^T)]
\]

\[
= \mathbb{E}[\text{tr}(V_k - V^*) (V_k - V^*)^T]
\]

\[
= \mathbb{E}[\text{tr}(X_k)],
\]
and using the bound in Lemma 7, one gets
\[ E[\|V_k - V^*\|_2^2] \leq \frac{36\sigma_{\max}^2 n^2 \alpha}{\delta d_{\min}(1-\gamma)^3} + \|x_0\|_2^2 n^2 \rho^{2k} \]
(10)

Taking the square root on both sides of the last inequality, using the subadditivity of the square root function, the Jensen inequality, and the concavity of the square root function, we have the desired conclusion.

Theorem 2. For any \( k \geq 0 \) and \( \varepsilon > 0 \), we have
\[
P \left[ \|V_k - V^*\|_2 < \varepsilon + \rho^k \sqrt{\|S^\prime\|} \|V_0 - V^*\|_2 \right] \
\geq 1 - \frac{36\sigma_{\max}^2 \|S^\prime\|^2 \alpha}{\varepsilon^2 d_{\min}(1-\gamma)^3} \left[ \|V_0 - V^*\|_2^2 \rho^{2k} \right].
\]

Proof. For notational simplicity, let \( n = |S| \) and \( x_k = V_k - V^* \). Moreover, define the random variables \( Y_i := e_i^T x_k, i \in \{1, 2, \ldots, n\} \). By the multivariate Chebyshev’s inequality [32], we have
\[
P \left[ \sum_{i=1}^n \|Y_i - E[Y_i]\|_2^2 \geq \varepsilon^2 \right] \leq \sum_{i=1}^n \frac{E[\|Y_i - E[Y_i]\|_2^2]}{\varepsilon^2},
\]
and equivalently,
\[
P \left[ \|x_k - E[x_k]\|_2 \geq \varepsilon \right] \leq \sum_{i=1}^n \frac{E[\|e_i^T x_k - e_i^T E[x_k]\|_2^2]}{\varepsilon^2} \leq \sum_{i=1}^n \frac{E[\|e_i^T x_k\|_2^2]}{\varepsilon^2},
\]
which further leads to
\[
P \left[ \|x_k - E[x_k]\|_2 \geq \varepsilon \right] \leq \sum_{i=1}^n \frac{E[\|e_i^T x_k\|_2^2]}{\varepsilon^2} \leq \sum_{i=1}^n \frac{E[\|e_i^T x_k - e_i^T E[x_k]\|_2^2]}{\varepsilon^2} \leq \frac{1}{\varepsilon^2} \frac{36\sigma_{\max}^2 n^2 \alpha}{\delta d_{\min}(1-\gamma)^3} \|x_0\|_2^2 n^2 \rho^{2k},
\]
where the last inequality of the corresponding complements event set is bounded by
\[
P \left[ \|x_k - E[x_k]\|_2 \leq \varepsilon \right] \geq 1 - \frac{36\sigma_{\max}^2 n^2 \alpha}{\varepsilon^2 d_{\min}(1-\gamma)^3} \left[ \|x_0\|_2^2 n^2 \rho^{2k} \right].
\]

The reverse triangle inequality leads to
\[
P \left[ \|x_k - E[x_k]\|_2 < \varepsilon \right] \leq \P \left[ \|x_k\|_2 - \|E[x_k]\|_2 < \varepsilon \right].
\]
Therefore, combining the last two inequalities yields
\[
P \left[ \|x_k\|_2 < \varepsilon + \|E[x_k]\|_2 \right] \geq 1 - \frac{36\sigma_{\max}^2 n^2 \alpha}{\varepsilon^2 d_{\min}(1-\gamma)^3} \left[ \|x_0\|_2^2 n^2 \rho^{2k} \right].
\]
Then, using \( \|E[x_k]\|_2 \leq \sqrt{n} \|E[x_k]\|_\infty \leq \rho^k \sqrt{n} \|x_0\|_\infty \), we have
\[
P \left[ \|x_k\|_2 < \varepsilon + \rho^k \sqrt{n} \|x_0\|_\infty \right] \geq 1 - \frac{36\sigma_{\max}^2 n^2 \alpha}{\varepsilon^2 d_{\min}(1-\gamma)^3} \left[ \|x_0\|_2^2 n^2 \rho^{2k} \right].
\]

Lemma 8. There exists a positive definite \( M > 0 \) such that
\[
A^T MA = M - I,
\]
and
\[
\lambda_{\min}(M) \geq 1, \quad \lambda_{\max}(M) \leq \frac{n}{1 - \rho}.
\]
Proof. Consider matrix \( M \) such that
\[
M = \sum_{k=0}^\infty (A^k)^T A^k.
\]
Noting that \( A^T MA + I = A^T \left( \sum_{k=0}^\infty (A^k)^T A^k \right) + I = M \), we have \( A^T MA + I = M \), resulting in the desired conclusion. Next, it remains to prove the existence of \( P \) by proving its boundedness. In particular, taking the norm on \( M \) leads to
\[
\|M\|_2 = \|I + A T A + (A^2)^T A^2 + \cdots\|_2 \leq \|I\|_2 + \|A T A\|_2 + \|A^2 T A^2\|_2 + \cdots = \|I\|_2 + \|A\|_2^2 + \|A^2\|_2^2 + \cdots = 1 + n \|A\|_2^2 + n \|A^2\|_\infty + \cdots = 1 + n \frac{n}{1 - \rho}.
\]
which implies the boundedness. Next, we prove the bounds on the maximum and minimum eigenvalues. From the definition (17), \( M \succeq I \), and hence \( \lambda_{\min}(M) \geq 1 \). On the other hand, one gets
\[
\lambda_{\max}(M) = \lambda_{\max}(I + A^T A + (A^2)^T A^2 + \cdots) \\
\leq \lambda_{\max}(I) + \lambda_{\max}(A^T A) + \lambda_{\max}(A^T A^2) + \cdots \\
= \lambda_{\max}(I) + \| A \|_2^2 + \| A^2 \|_2^2 + \cdots \\
\leq 1 + n\| A \|^2_\infty + n\| A^2 \|^2_\infty + \cdots \\
\leq \frac{n}{1 - \rho^2} \\
\leq \frac{n}{1 - \rho}
\]
The proof is completed. \( \square \)

Now, we are ready to prove Theorem 3.

Proof of Theorem 3: Consider the quadratic Lyapunov function,
\[
v(x) := x^T M x,
\]
where \( M \) is a positive definite matrix defined in (17). Using Lemma 8, we have
\[
\mathbb{E}[v(x_{k+1})] \\
= \mathbb{E}[(Ax_k + \alpha w_k)^T M (Ax_k + \alpha w_k)] \\
\leq \mathbb{E}[v(Ax_k)] + \alpha^2 \lambda_{\max}(M) W_{\max} \\
\leq \mathbb{E}[v(x_k)] - x_k^T x_k + \alpha^2 \lambda_{\max}(M) W_{\max},
\]
where the first inequality follows from Lemma 8 and Lemma 2. Rearranging the last inequality, summing it over \( i = 0 \) to \( k - 1 \), and dividing both sides by \( k \) lead to
\[
\frac{1}{k} \sum_{i=0}^{k-1} \mathbb{E}[x_i^T x_i] \leq \frac{1}{k} v(x_0) + \alpha^2 \lambda_{\max}(M) W_{\max}
\]
Using Jensen’s inequality, \( \lambda_{\min}(M) \| x \|_2^2 \leq v(x) \leq \lambda_{\max}(M) \| x \|_2^2 \), and \( \| x_0 \|_2 \leq \sqrt{n} \| x_0 \|_\infty \) with Lemma 8, we have the desired conclusion.

With a prescribed finial iteration number and the final iteration dependent constant step-size, we can obtain \( \mathcal{O}(1/\sqrt{T}) \) convergence rate with respect to the mean-squared error
\[
\mathbb{E} \left[ \left\| \frac{1}{T} \sum_{i=0}^{T-1} V_i - V^\pi \right\|^2_2 \right]
\]
and \( \mathcal{O}(1/T^{1/4}) \) convergence rate with respect to
\[
\mathbb{E} \left[ \left\| \frac{1}{T} \sum_{i=0}^{T-1} V_i - V^\pi \right\|_2 \right]
\]

Corollary 1. For any final iteration number \( T \geq 0 \) and the prescribed constant step-size \( \alpha = \sqrt{T/2} \), we have
\[
\mathbb{E} \left[ \left\| \frac{1}{T} \sum_{i=0}^{T-1} V_i - V^\pi \right\|^2_2 \right] \\
\leq \frac{1}{T^{1/4}} \sqrt{\frac{|S|}{d_{\min}(1 - \gamma)}} \| V_0 - V^\pi \|_2 + \frac{36 \sigma^2 \pi_{\max} |S|^2}{d_{\min}(1 - \gamma)^2}.
\]
The proof is straightforward, and hence, is omitted here.

VI. COMPARATIVE ANALYSIS

We compare the proposed result with some recent results in the literature on the convergence of TD-learning. Theorem 2 in [15] provides a finite-time bound under the constant step-size, i.i.d. observation model, and the linear function approximation. The bound is given by, for any \( \alpha \leq \frac{d_{\min}}{\sqrt{2}n} \) and \( k \geq 0 \),
\[
\mathbb{E}[\| V_k - V^\pi \|_2] \leq \frac{\exp(-\alpha (1 - \gamma) d_{\min} k)}{d_{\min}} \| V_0 - V^\pi \|_2 + \sqrt{\frac{2\alpha\sigma^2}{(1 - \gamma) n d_{\min}^2}}.
\]

where \( \sigma^2 = \mathbb{E} \left[ \left\| (e_{s_k} e_{s_k}^T)^T V_k + \gamma (e_{s_k} e_{s_k}^T)^T V^\pi - e_{s_k} e_{s_k}^T V^\pi \right\|^2_2 \right] \).

Compared to (12), the main advantage of (9) lies in its range of the available constant step-size \((0, 1)\), which covers a wider range than that in [15]. Similar to [15], the proposed scheme also applies simple arguments. On the other hand, we applies a different approach based on the linear system model, which provides additional insights on TD-learning. Moreover, the proposed bound in (9) can cover a different case, off-policy learning, while [15] considers an on-policy learning with the linear function approximation. The recent work, Theorem 3.1. in [18], provides a finite-time bound under the i.i.d. observation model and linear function approximation. They use a diminishing step-size of the form \( \alpha_k = 1/(k + 1) - \alpha \), where \( \alpha \in (0, 1) \). Moreover, some parameters in their bound depends on the system properties which are not easily analyzable. Therefore, it is not comparable with our results.

CONCLUSION

In this technical note, we introduce a new and simple finite-time analysis of temporal difference (TD) learning based on stochastic linear system models. The analysis especially covers the tabular TD-learning case with constant step-sizes and off-policy scenarios. The analysis is based on discrete-time stochastic linear system models, which is unique in the literature. The developed finite-time bounds also include unique features summarized in the last section. The overall analysis processes are simple enough to be described in few lines compared to existing approaches. The proposed work provides additional insights on TD-learning and RLs with simple concepts and analysis tools in control theory. The frameworks and ideas in this technical note can complement existing analysis in the literature.

REFERENCES

[1] R. S. Sutton, “Learning to predict by the methods of temporal differences,” Machine learning, vol. 3, no. 1, pp. 9–44, 1988.
[2] R. S. Sutton and A. G. Barto, Reinforcement Learning: An introduction. MIT Press, 1998.
[3] M. L. Puterman, Markov decision processes: Discrete stochastic dynamic programming. John Wiley & Sons, 2014.
[4] C. J. C. H. Watkins and P. Dayan, “Q-learning,” Machine learning, vol. 8, no. 3-4, pp. 279–292, 1992.
[5] G. A. Rummery and M. Niranjan, On-line Q-learning using connectionist systems, University of Cambridge, Department of Engineering Cambridge, England, 1994, vol. 37.
[6] V. R. Konda and J. N. Tsitsiklis, “Actor-critic algorithms,” in Advances in neural information processing systems, pp. 1008–1014.
[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, p. 529, 2015.
[8] H. V. Hasselt, “Double Q-learning,” in Advances in Neural Information Processing Systems, 2010, pp. 2613–2621.
[9] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, “Fast gradient-descent methods for temporal-difference learning with linear function approximation,” in Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 993–1000.
[10] R. S. Sutton, H. R. Maei, and C. Szepesvári, “A convergent O(n) temporal-difference algorithm for off-policy learning with linear function approximation,” in Advances in neural information processing systems, 2009, pp. 1609–1616.
[11] D. Silver, G. Lever, N. Heess, T. Derezis, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” in International conference on machine learning, 2014, pp. 387–395.
[12] T. Jaakkola, M. I. Jordan, and S. P. Singh, “Convergence of stochastic iterative dynamic programming algorithms,” in Advances in neural information processing systems, 1994, pp. 703–710.
[13] V. S. Borkar and S. P. Meyn, “The ODE method for convergence of stochastic approximation and reinforcement learning,” SIAM Journal on Control and Optimization, vol. 38, no. 2, pp. 447–469, 2000.
Lemma 4

Lemma 2

C.-T. Chen, "Nonlinear systems," Operations Research, vol. 69, no. 3, pp. 950-973, 2021.

D. Lee and N. He, "A unified switching system perspective and convergence analysis for TD(0) with function approximation," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

Lemma 4 completes the proof.

We will find the least required number of samples to achieve
\[ P \left[ \|x_n\|_\infty < \eta \right] \geq 1 - \delta \]

According to Theorem 2, a sufficient condition is
\[ \varepsilon = \frac{2}{\sqrt{\eta}} \rho \sqrt{n} \|x_0\|_\infty \leq \frac{\eta}{2} \] (14)

and
\[ \frac{1}{\eta^2} \frac{36\|S\|\|S\|\|\alpha\|}{\varepsilon^2 d_{\min}(1-\gamma)^3} \leq \delta \] (15)

The inequality in (14) holds if
\[ k \geq \left( \frac{2\|S\|\|x_0\|_\infty}{\eta} \right) / \left( \ln \left( \frac{1}{\rho} \right) \right) \] (16)

Similarly, (15) holds if
\[ \frac{4\|S\|\|x_0\|_\infty^2}{\eta^2} \leq \frac{\delta}{2} \] (17)

and
\[ \frac{4\|x_0\|_\infty^2}{\eta^2} \leq \frac{\delta}{2} \] (18)

hold. The inequality in (17) is satisfied if
\[ \alpha = \frac{\delta n^2 d_{\min}(1-\gamma)^3}{288\|S\|^2} \] (19)

Moreover, the inequality in (18) is satisfied if
\[ k \geq \frac{1}{2} \ln \left( \frac{8\|x_0\|_\infty^2}{\eta^2} \right) / \left( \ln \left( \frac{1}{\rho} \right) \right) \] (20)

Both (16) and (20) hold if
\[ k \geq \frac{\ln \left( \frac{\|x_0\|_\infty^2}{\eta n^2} \right) \|S\|^2}{\ln \left( \frac{1}{\rho} \right)} \] (21)

Using the inequalities \( 1 - \frac{1}{x} \leq \ln x \leq x - 1 \) for any \( x > 0 \) and choosing the step-size (19), we have the desired conclusion.