Control Theoretic Analysis of Temporal Difference Learning

1st Donghwan Lee*
Department of Electrical Engineering
KAIST
Daejeon, South Korea
donghwan@kaist.ac.kr

2nd Do Wan Kim
Department of Electrical Engineering
Hanbat National University
Daejeon, South Korea
dowankim@hanbat.ac.kr

*Corresponding author

Abstract—The goal of this manuscript is to conduct a control-theoretic analysis of Temporal Difference (TD) learning algorithms. TD-learning serves as a cornerstone in the field of reinforcement learning, offering a methodology for approximating the value function associated with a given policy in a Markov Decision Process. Despite several existing works that have contributed to the theoretical understanding of TD-learning, it is only in recent years that researchers have been able to establish concrete guarantees on its statistical efficiency. In this paper, we introduce a finite-time, control-theoretic framework for analyzing TD-learning, leveraging established concepts from the field of linear systems control. Consequently, this paper provides additional insights into the mechanics of TD learning and the broader landscape of reinforcement learning, all while employing straightforward analytical tools derived from control theory.

Index Terms—Reinforcement learning, temporal difference learning (TD-learning), Markov decision problem, finite-time analysis, convergence

I. INTRODUCTION

Originally proposed in [1], temporal difference learning (TD-learning) serves as a foundational algorithm in the field of reinforcement learning (RL) [2], [3]. It is designed to approximate the value function of a specified policy for a Markov Decision Process (MDP) [4]. This seminal algorithm has been extended and incorporated into a wide array of more advanced algorithms, including but not limited to classical Q-learning [5], SARSA [6], actor-critic methods [7], as well as contemporary RL approaches such as deep Q-learning [8], double Q-learning [9], gradient TD-learning [10], [11], and deterministic actor-critic algorithms [12].

While there exists a rich body of work dedicated to the theoretical analysis of TD-learning [1], [13]–[18], much of the classical literature has predominantly focused on the asymptotic behavior of the algorithm. It is only in recent years that researchers have begun to provide guarantees on the statistical efficiency of TD-learning [19]–[22]. Specifically, these recent contributions explore the rate at which TD iterates converge to the optimal solution. This rate is often quantified through error bounds that are a function of the number of time-steps, a form of analysis commonly referred to as finite-time analysis.

In this paper, we introduce a novel finite-time error bound for TD-learning, offering additional perspectives on TD-learning. By additional perspectives, we refer to an analysis interpretable through straightforward concepts and standard analytical tools in control theory such as Lyapunov theory [23]. Specifically, we present a unique control-theoretic, finite-time analysis of TD-learning that leverages linear system models and established notions from linear systems literature [24], and derive mean-squared error bounds for both final and averaged iterates. The analysis provides clear connections between TD-learning and principles in linear systems, thereby enriching our understanding of TD-learning and RL through the lens of control theory.

It is worth noting that the scope of this paper is confined to an i.i.d. observation model with a constant step-size, to maintain the simplicity and clarity of the overall analysis. While extensions to Markovian observation scenarios can be addressed following the methods from recent works [19], [21], the corresponding derivations would significantly complicate the overall analysis and potentially obfuscate the core contributions of this paper. Therefore, such extensions are beyond the purview of this paper. We also note that the i.i.d. observation model is a commonly adopted framework in existing literature [10], [11], [14], [20], [22]. Lastly, our approach exclusively considers a constant step-size to align with the linear time-invariant (LTI) system model framework.

Related works: Recently, significant advancements have been achieved in the finite-time analysis of TD-learning algorithms [19], [21], [22]. Specifically, [22] pioneered the finite-time analysis of TD-learning, while the work in [21] enhanced the convergence properties of TD-learning, building upon the foundation laid by [22], through the incorporation of Markovian sampling and higher-order moments. Furthermore, [19] conducted a finite-time analysis of TD-learning by employing standard techniques commonly found in the literature on stochastic gradient descent.

In addition to these works, it is worth noting recent contributions related to the control system analysis of reinforcement learning algorithms [16], [25], [26]. The dynamical system...
perspective on reinforcement learning and general stochastic iterative algorithms has a long history, tracing its roots back to ordinary differential equation (O.D.E.) analysis [14], [27]–[29]. More recently, [25] examined the asymptotic convergence of Q-learning [5] through the lens of a continuous-time switched linear system model [30]. Concurrently, [16] explored the asymptotic convergence of TD-learning based on Markovian Jump Linear Systems (MJLSs). A discrete-time switched linear system model was employed for finite-time analysis of Q-learning in [26].

Lastly, we present several features that distinguish the proposed method from existing approaches in the literature. In particular, [21], [22] employ a continuous-time O.D.E. model, and their analysis hinges on the stability of continuous-time LTI models, i.e., Hurwitz stability. In contrast, our work directly focuses on discrete-time LTI models and leverages properties associated with Schur stability. Compared to [19], the proposed analysis presents different convergence rates and conditions on the step-size. Additionally, the proposed approach offers a unique discrete-time LTI system perspective on TD-learning, thereby providing further insights into the subject. The proposed method also exhibits distinct features when compared to the discrete-time Markovian jump linear systems (MJLS) model presented in [16]. Specifically, while [16] consider an equivalent discrete-time MJLS model for TD-learning, our approach employs an equivalent discrete-time LTI system augmented with random noises, encapsulating all randomness into a single noise vector. In addition, while [16] focuses on asymptotic convergence, the proposed work provides finite-time analysis. Although recent contributions [19], [21], [22] have yielded valuable results in finite-time analysis under various conditions, our approach offers additional finite-time bounds under distinct conditions, as well as novel insights. These insights serve to deepen our understanding of linear stochastic approximation and TD-learning through the lens of linear system theory. In this regard, we view our proposed methods as complementary to, rather than replacements for, existing methodologies.

**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 $A$; $A \succ 0$ $(A \prec 0$, $A \succeq 0$, and $A \preceq 0$, respectively): symmetric positive definite (negative definite, positive semi-definite, and negative semi-definite, respectively) matrix $A$; $I$: identity matrix with appropriate dimensions; for any matrix $A$, $[A]_{ij}$ is the element of $A$ in $i$-th row and $j$-th column; $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ for any symmetric matrix $A$: the minimum and maximum eigenvalues of $A$; $|S|$: cardinality of a finite set $S$.

II. PRELIMINARIES

A. Markov decision problem

We consider the infinite-horizon discounted Markov decision problem (MDP) [3], [4], where the agent sequentially takes actions to maximize cumulative discounted rewards. In a Markov decision process with the state-space $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$ and action-space $\mathcal{A} := \{1, 2, \ldots, |\mathcal{A}|\}$, the decision maker selects an action $a \in \mathcal{A}$ 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 $r(s_k, a_k, s_{k+1}) := r_k, k \in \{0, 1, \ldots\}$. A stochastic policy is a map $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ representing the probability, $\pi(a|s)$, of selecting action $a$ at the current state $s$, while a deterministic policy is a map $\pi : S \to A$. In this paper, we usually focus on the stochastic policy because the deterministic counterpart can be seen as a special case of the stochastic case. 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., $\pi^* := \arg\max_{\pi} \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_k | \pi]$, 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 $\mathbb{E}[|\pi|]$ 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 \mathcal{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 addressed in this paper 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 $\pi$ we want to evaluate is called the target policy. On the other hand, the behavior policy, denoted by $\beta$, is the policy that is used to generate experiences. For a learning algorithm, if $\beta = \pi$, it is called on-policy learning. Otherwise, it is called off-policy learning. The stationary state distribution, if exists, is defined as

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

where $\beta$ is any behavior policy. Throughout, we assume that the induced Markov chain of the underlying MDP with $\pi$ is aperiodic and irreducible so that it has a unique stationary state distribution, and the Markov decision problem is well posed, which is a standard assumption.

B. TD-Learning

We consider a version of TD-learning given in Algorithm 1, which is an on-policy learning, i.e., $\beta = \pi$, because the transition $(s_k, a_k, r_k, s'_k)$ is generated using the target policy $\pi$.

Compared to the original TD-learning, the step-size $\alpha$ is constant in this paper. We make the following 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) (Unit bound on rewards) The reward is bounded as follows:
$$\max_{(s,a,s') \in S \times A \times S} |r(s,a,s')| \leq 1.$$ 
4) (Unit bound on initial parameter) The initial iterate $V_0$ satisfies $\|V_0\|_{\infty} \leq 1$.

The first statement in Assumption 1 is crucial for the proposed finite-time analysis, and this assumption is considered standard in the literature [31–33] or even more relaxed compared to existing conditions on the step-sizes [19], [21]. The second statement ensures that every state can be visited infinitely often, facilitating sufficient exploration, which is a standard assumption in the literature [3]. The third and fourth statements regarding the unit bounds imposed the reward function and $V_0$ are introduced for the sake of simplicity in analysis, without sacrificing generality. To conclude this subsection, we also introduce the concept of boundedness for TD-learning iterates [31], which plays a crucial role in our analysis.

Lemma 1 (Boundedness of TD-learning iterates [31]). Under Assumption 1, for all $k \geq 0$, we have
$$\|V_k\|_{\infty} \leq V_{\text{max}} := \frac{1}{1 - \gamma}.$$ 

The boundedness has been established for Q-learning in [31] but not for TD-learning. For this reason, we provide its proof in Appendix A for completeness of the presentation.

C. Linear Systems

In this subsection, we briefly review basic notions in linear system theory [23], [24]. Let us consider the discrete-time linear system
$$x_{k+1} = Ax_k, \quad x_0 \in \mathbb{R}^n, \quad k \in \{1,2,\ldots\}, \quad (1)$$
Here, $x_k \in \mathbb{R}^n$ represents the state, $k \in \{1,2,\ldots\}$ denotes the discrete time, and $A \in \mathbb{R}^{n \times n}$ is referred to as the system matrix. The system is considered asymptotically stable [23] if, for any initial state $x_0 \in \mathbb{R}^n$, $x_k$ converges to zero as $k$ approaches infinity. Moreover, it is called exponentially stable if the state converges exponentially. For linear systems, both notions are known to be equivalent. The matrix $A$ is classified as a Schur matrix if the absolute values of its eigenvalues are strictly less than one, or equivalently, the spectral radius of $A$ is strictly less than one. In this case, the system in (1) is referred to as a Schur stable system. It is known that Schur stability is equivalent to exponential stability as well as asymptotic stability.

On the other hand, let us consider the continuous-time linear system
$$\dot{x}_t = Ax_t, \quad x_0 \in \mathbb{R}^n, \quad t \in \mathbb{R}_+^n, \quad (2)$$
Here, $x_t \in \mathbb{R}^n$ represents the state, $t \in \mathbb{R}_+^n$ denotes the discrete time, and $\mathbb{R}_+^n$ represents the set of all vectors of dimension $n$ with nonnegative real numbers as their entries. The notions of asymptotic stability and exponential stability are defined similarly for this continuous-time system. The matrix $A$ is referred to as a Hurwitz matrix if all eigenvalues reside in the open left half-plane of the complex plane. In such cases, the system (2) is known as a Hurwitz stable system. Similar to the discrete-time scenario, Hurwitz stability is equivalent to both exponential and asymptotic stability.

III. LINEAR SYSTEM MODEL OF TD-LEARNING

In this section, we investigate a discrete-time linear system model that corresponds to Algorithm 1 and conduct a finite-time analysis based on stability analysis of linear systems. To enhance clarity and facilitate easy reference, we summarize the definitions of the following notations, which will be extensively utilized throughout this paper.

Definition 2.

1) Maximum state visit probability:
$$d_{\text{max}} := \max_{s \in S} d(s) \in (0,1).$$
2) Minimum state 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
$$D := \begin{bmatrix} d(1) & \vdots & d(|S|) \end{bmatrix} \in \mathbb{R}^{|S| \times |S|},$$
5) $P^\pi \in \mathbb{R}^{|S| \times |S|}$ is a matrix such that $[P^\pi]_{i,j} = P[s' = j|s = i|\pi]$
6) $R^\pi \in \mathbb{R}^{|S|}$ is a vector such that $[R^\pi]_i = \mathbb{E}[r(s_k, a_k, s_{k+1})|a_k \sim \pi(s_k), s_k = i]$

Using the notation introduced, the update in Algorithm 1 can be equivalently rewritten as
$$V_{k+1} = V_k + \alpha\{DR^\pi + \gamma DP^\pi V_k - DV_k + w_k\}, \quad (3)$$
where
$$w_k := e_{a_k} \delta_k - (DR^\pi + \gamma DP^\pi V_k - DV_k), \quad (4)$$

Algorithm 1 On-policy TD-learning

1: Initialize $V_0 \in \mathbb{R}^{|S|}$ arbitrarily such that $\|V_0\|_{\infty} \leq 1$.
2: for iteration $k = 0, 1, \ldots$ do
3: \quad Observe $s_k \sim d$, $a_k \sim \pi(\cdot|s_k)$, $s'_k \sim P(\cdot|s_k, a_k)$ and $r_k = r(s_k, a_k, s'_k)$
4: \quad Update $V_{k+1}(s_k) = V_k(s_k) + \alpha(r_k + \gamma V_k(s'_k) - V_k(s_k))$
5: end for
\[ \delta_k := r_k + \gamma c_{s_k+1}^T V_k - e_{s_k}^T V_k, \]

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, \((s_k, a_k, r_k, s_{k+1})\) is the sample transition in the \( k \)-th time-step. The expressions can be equivalently reformulated as the linear system

\[ V_{k+1} = (I + \lambda D P \pi - D) V_k + \alpha R^\pi + \alpha w_k. \]

Invoking the Bellman equation \((\gamma D P \pi - D) V^\pi + DR^\pi = 0\) leads to the equivalent equation

\[ V_{k+1} - V^\pi = (I + \alpha (\gamma D P \pi - D)) (V_k - V^\pi) + \alpha w_k. \]

Note that the term \( b \) in (6) has been cancelled out in (7) by adding the Bellman equation \((-\alpha (\gamma D P \pi - D) V^\pi + D R^\pi) = 0\). Next, 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 discrete-time stochastic linear system

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

where \( n := |S| \), 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 above system. The first important property is that the noise \( w_k \) has the zero mean, and is bounded. It is formally stated in the following lemma with proofs given in Appendix B.

**Lemma 2.** We have

1) \( \mathbb{E}[w_k] = 0; \)

2) \( \mathbb{E}[\|w_k\|_\infty] \leq \sqrt{W_{\max}}; \)

3) \( \mathbb{E}[\|w_k\|_2] \leq W_{\max}; \)

4) \( \mathbb{E}[w_k^T w_k] \leq \frac{1}{(1 - \gamma)^2} : = W_{\max}. \)

for all \( k \geq 0 \).

To proceed further, let us define the covariance of the noise

\[ \mathbb{E}[w_k w_k^T] = : W_k = W^T_k \geq 0. \]

The covariance matrix will play a central role in the proposed analysis. In particular, 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 defined in Lemma 2.

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

Lastly, we investigate the property of the system matrix \( A \) in (8). 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 \( |A|_\infty := \max_{1 \leq i \leq m} \sum_{j=1}^{|S|} |[A]_{ij}| \) and \([A]_{ij}\) is the element of \( A \) in \( i \)-th row and \( j \)-th column.

**Proof.** Noting that \( A = I + \alpha (\gamma D P \pi - D) \), we have

\[ \sum_j |[A]_{ij}| = \sum_j |[I - \alpha D + \alpha \gamma D P \pi]_{ij}| = |I - \alpha D|_{ii} + \sum_j [\alpha \gamma D P \pi]_{ij}, \]

\[ = 1 - \alpha |D|_{ii} + \alpha \gamma |D|_{ii} \sum_j [P \pi]_{ij}, \]

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

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

\[ |A|_\infty = \max_{i \in \{1, 2, ..., |S|\}} \{ 1 + \alpha |D|_{ii} (\gamma - 1) \} \]

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

which completes the proof.

**Proof.** Noting that \( A = I + \alpha (\gamma D P \pi - D) \), we have

\[ \sum_j |[A]_{ij}| = \sum_j |[I - \alpha D + \alpha \gamma D P \pi]_{ij}| = |I - \alpha D|_{ii} + \sum_j [\alpha \gamma D P \pi]_{ij}, \]

\[ = 1 - \alpha |D|_{ii} + \alpha \gamma |D|_{ii} \sum_j [P \pi]_{ij}, \]

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

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

\[ |A|_\infty = \max_{i \in \{1, 2, ..., |S|\}} \{ 1 + \alpha |D|_{ii} (\gamma - 1) \} \]

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

which completes the proof.

**Proof.** Noting that \( A = I + \alpha (\gamma D P \pi - D) \), we have

\[ \sum_j |[A]_{ij}| = \sum_j |[I - \alpha D + \alpha \gamma D P \pi]_{ij}| = |I - \alpha D|_{ii} + \sum_j [\alpha \gamma D P \pi]_{ij}, \]

\[ = 1 - \alpha |D|_{ii} + \alpha \gamma |D|_{ii} \sum_j [P \pi]_{ij}, \]

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

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

\[ |A|_\infty = \max_{i \in \{1, 2, ..., |S|\}} \{ 1 + \alpha |D|_{ii} (\gamma - 1) \} \]

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

which completes the proof.

**Proof.** Noting that \( A = I + \alpha (\gamma D P \pi - D) \), we have

\[ \sum_j |[A]_{ij}| = \sum_j |[I - \alpha D + \alpha \gamma D P \pi]_{ij}| = |I - \alpha D|_{ii} + \sum_j [\alpha \gamma D P \pi]_{ij}, \]

\[ = 1 - \alpha |D|_{ii} + \alpha \gamma |D|_{ii} \sum_j [P \pi]_{ij}, \]

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

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

\[ |A|_\infty = \max_{i \in \{1, 2, ..., |S|\}} \{ 1 + \alpha |D|_{ii} (\gamma - 1) \} \]

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

which completes the proof.

**IV. Finite-Time Analysis: Final Iteration**

In this section, we investigate the finite-time error bound of the TD-learning in Algorithm 1 for the final iterate based on the linear system model in (8), which is done by analyzing
the propagations of both mean and correlation of the state \( x_k \).

First of all, taking the mean on (7) leads to

\[
E[x_{k+1}] = AE[x_k], \quad x_0 \in \mathbb{R}^n, \quad \forall k \geq 0,
\]

where \( E[w_k] = 0 \) in (7) due to the i.i.d. assumption on the samples of transitions.

Therefore, the mean state \( E[x_k] \) follows the behavior of the discrete-time linear system in (1). By utilizing Lemma 4, we can establish a finite-time bound for the mean dynamics driven by (9), which incorporates an exponentially converging term.

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

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

**Proof.** Taking the norm on (9) leads to \( \|E[x_{k+1}]\|_\infty = \|AE[x_k]\|_\infty \leq \|A\| \|E[x_k]\|_\infty \leq \rho \|E[x_k]\|_\infty \), 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 covariance matrix, \( E[x_kw_k^T] \), propagates over the time. In particular, the covariance matrix is updated through the recursion

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

where \( E[w_kw_k^T] = W_k \). Defining \( X_k := E[x_kx_k^T], k \geq 0 \), it is equivalently written as

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

with \( X_0 := x_0x_0^T \). It is worth noting that the above recursion corresponds to the Lyapunov matrix associated with the system matrix \( A^T \). In other words, the covariance matrix can be seen as a Lyapunov matrix corresponding to the discrete-time system \( x_{k+1} = A^T x_k \).

A natural question arises regarding the convergence of the iterate, \( X_k \), as \( k \to \infty \). We can at least prove that \( X_k \) bounded.

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

\[
\|X_k\|_2 \leq \frac{\alpha^2 W_{\max}}{1 - \rho^2} + n \|X_0\|_2
\]

where \( \rho \) is defined in Definition 2.

The proof is given in Appendix C. 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:

\[
\text{tr}(X_k) \leq \frac{36n^2\alpha}{d_{\min}(1 - \gamma)^3} + \|x_0\|_2^2 n^2 \rho^{2k}
\]

where \( \rho \) is defined in Definition 2.

The proof is given in Appendix D. Now, we provide a finite-time bound on the mean-squared error motivated by some standard notions in discrete-time linear system theory.

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

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

where \( \rho \) is defined in Definition 2.

**Proof.** Noting the relations

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

and using the bound in Lemma 7, one gets

\[
\mathbb{E}[(\|V_k - V^\pi\|_2^2) \leq \frac{36n^2\alpha}{d_{\min}(1 - \gamma)^3} + \|x_0\|_2^2 n^2 \rho^{2k}
\]

Taking the square root on both side 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.

The first term on the right-hand side of (10) can be reduced arbitrarily by decreasing the step size \( \alpha \in (0, 1) \). The second bound diminishes exponentially as \( k \to \infty \), at a rate of \( \rho = 1 - \alpha d_{\min}(1 - \gamma) \in (0, 1) \).

The bound presented in Theorem 1 is derived based on the discrete-time linear stochastic system model described in (3). In standard control theory, the stochastic noise term \( w_k \) is typically assumed to follow a Gaussian distribution. However, the system in (3) incorporates a stochastic noise term \( w_k \) with special structures. Specifically, \( w_k \) in Theorem 1 represents discrete random variables that capture essential structures of TD-learning. Furthermore, the noise term is bounded, as demonstrated in Lemma 2, which plays a crucial role in establishing the bound in Lemma 6. The noise term is also special in the sense that it depends on the state \( x_k \), and hence, to establish its boundedness in Lemma 6, additional analysis on the boundedness of the state vector \( x_k \) is required as shown in Lemma 1.

Finally, note that while our analysis is primarily conducted in tabular settings, it can be extended to incorporate on-policy linear function approximation with additional effort.

V. Finite-Time Analysis: Averaged Iteration

In this section, we additionally provide a finite-time bound in terms of the averaged iteration \( \frac{1}{k} \sum_{i=0}^{k-1} V_i \). We can first obtain the following result.

**Theorem 2.** For any \( k \geq 0 \), it holds that

\[
\mathbb{E} \left[ \left( \frac{1}{k} \sum_{i=0}^{k-1} V_i - V^\pi \right)^2 \right] \leq \frac{1}{k} \alpha d_{\min}(1 - \gamma)^3 \left( \|V_0 - V^\pi\|_2 + \sqrt{\frac{36|S|^2}{d_{\min}(1 - \gamma)^3}} \right)^2.
\]

The analysis of the final iterate in the previous section is based on the Lyapunov matrix associated with the system matrix \( A^T \). On the other hand, the analysis of the averaged iterate in this section relies on the Lyapunov matrix associated with the dual system matrix \( A \). In the analysis of the final iterate,
the covariance matrix of the state is propagated. On the other hand, the analysis of the averaged iterate in Theorem 2 draws upon the well-established principles of Lyapunov theory [23], [24], which are summarized below.

**Lemma 8.** There exists a positive definite $M > 0$ such that

$$A^TMA = M - I,$$

and

$$\lambda_{\text{min}}(M) \geq 1, \quad \lambda_{\text{max}}(M) \leq \frac{n}{1 - \rho}.$$  

The proof of Lemma 8 is given in Appendix E. Note that in the proof of Lemma 8, webound $\frac{n}{1 - \rho}$ by $\frac{1}{\rho} n$ just to simplify the final expression. Note also that $M$ in Lemma 8 can be seen as a Lyapunov matrix corresponding to the discrete-time linear system $x_{k+1} = Ax_k$, which is dual to the system associated with the covariance matrix in the previous sections. Now, we are ready to prove Theorem 2.

**Proof of Theorem 2.** Consider the quadratic Lyapunov function, $v(x) = x^T M x$, where $M$ is a positive definite matrix defined in (14). Using Lemma 8, we have

$$E[v(x_{k+1})] = E[(Ax_k + \alpha w_k)^T M (Ax_k + \alpha w_k)] \leq E[v(x_k)] + \alpha^2 \lambda_{\text{max}}(M) W_{\text{max}} \leq E[v(x_k)] - x^T_k x_k + \alpha^2 \lambda_{\text{max}}(M) W_{\text{max}}.$$

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

$$\frac{1}{k} \sum_{i=0}^{k-1} E[x_i^T x_i] \leq \frac{1}{k} v(x_0) + \alpha^2 \lambda_{\text{max}}(M) W_{\text{max}}.$$

Using Jensen’s inequality, $\lambda_{\text{min}}(M) \|x\|^2 \leq v(x) \leq \lambda_{\text{max}}(M) \|x\|^2$, and $\|x_0\|_2 \leq \sqrt{n} \|x_0\|_\infty$ with Lemma 8, we have the desired conclusion. \qed

With a prescribed final iteration number and the final iteration dependent constant step-size, we can obtain $O(1/\sqrt{T})$ convergence rate with respect to the mean-squared error

$$E \left[ \frac{1}{T} \sum_{i=0}^{T-1} V_i - \mathcal{V} \right] = O(1/T^{1/4}),$$

with respect to $E \left[ \frac{1}{T} \sum_{i=0}^{T-1} V_i - \mathcal{V} \right] = O(1/T^{1/4}).$ This result is summarized below.

**Corollary 1.** For any final iteration number $T \geq 0$ and the prescribed constant step-size $\alpha = \frac{1}{\sqrt{T}}$, we have

$$E \left[ \frac{1}{T} \sum_{i=0}^{T-1} V_i - \mathcal{V} \right] \leq \frac{1}{T^{1/4}} \left( \sqrt{\frac{|S|}{d_{\text{min}}(1 - \gamma)}} \|V_0 - \mathcal{V}\|_2 + \frac{3\alpha^2 \lambda_{\text{max}}^2 |S|^2}{d_{\text{min}}(1 - \gamma)^3} \right).$$

The proof is straightforward, and hence, is omitted here.

**Conclusion**

In this paper, we have introduced a control-theoretic, finite-time analysis of TD-learning. While recent advancements have yielded valuable insights into finite-time analysis under a variety of conditions, the proposed approach offers additional finite-time bounds under unique conditions. These new bounds, along with the insights gained, serve to enrich our understanding of TD-learning through the application of concepts and tools from linear system theory. However, it is important to note the limitations of our current work. Specifically, we have restricted our analysis to scenarios with i.i.d. observation models. Additionally, we have not considered linear function approximation or diminishing step-size in our model. These topics provide promising avenues for future research.

**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] D. P. Bertsekas and J. N. Tsitsiklis, Neuro-dynamic programming, Athena Scientific Belmont, MA, 1996.

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

[5] C. J. C. H. Watkins and P. Dayan, “Q-learning,” Machine learning, vol. 8, no. 3–4, pp. 279–292, 1992.

[6] G. A. Rummery and M. Niranjan, On-line Q-learning using connectionist systems. University of Cambridge, Department of Engineering Cambridge, England, 1994, vol. 37.

[7] V. R. Konda and J. N. Tsitsiklis, “Actor-critic algorithms,” in Advances in neural information processing systems, pp. 1008–1014.

[8] 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.

[9] H. V. Hasselt, “Double Q-learning,” in Advances in Neural Information Processing Systems, 2010, pp. 2613–2621.

[10] 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.

[11] R. S. Sutton, H. R. Maei, and C. Szepesvári, “A convergent O(\gamma) temporal-difference algorithm for off-policy learning with linear function approximation,” in Advances in neural information processing systems, 2009, pp. 1609–1616.

[12] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” in International conference on machine learning, 2014, pp. 387–395.

[13] 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.

[14] 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.

[15] J. N. Tsitsiklis and B. Van Roy, “An analysis of temporal-difference learning with function approximation,” IEEE Transactions on Automatic Control, vol. 42, no. 5, pp. 674–690, 1997.

[16] B. Hu and U. Syed, “Characterizing the exact behaviors of temporal difference learning algorithms using markov jump linear system theory,” in Advances in Neural Information Processing Systems, 2019, pp. 8477–8488.

[17] D. Lee, H.-D. Lim, J. Park, and O. Choi, “New versions of gradient temporal difference learning,” IEEE Transactions on Automatic Control, 2022.
Lemma 1 can be written as

\[ \leq (1 - \alpha)(1 + \gamma + \cdots + \gamma^k) + \alpha[1 + \gamma + \cdots + \gamma^{k+1}] \]

\[ \leq [1 + \gamma + \cdots + \gamma^k] - \alpha(1 + \gamma + \cdots + \gamma^k) + \alpha[1 + \gamma + \cdots + \gamma^{k+1}] \]

\[ \leq [1 + \gamma + \cdots + \gamma^k] + \alpha \gamma^{k+1} \]

\[ \leq 1 + \gamma + \cdots + \gamma^k + \alpha \gamma^{k+1} \]

which completes the proof.

APPENDIX B

PROOF OF LEMMA 2

For the first statement, we take the conditional expectation on (4) to have \( \mathbb{E}[w_k | \pi_k] = 0 \). Taking the total expectation again with the law of total expectation leads to the first conclusion. Moreover, the conditional expectation, \( \mathbb{E}[w_k^T w_k | V_k] \), is bounded as

\[ \mathbb{E}[w_k^T w_k | V_k] = \mathbb{E}[\| e_{s_k} \delta_k \| - (DR^x + \gamma D_P^x V_k - D V_k)^2 | V_k] \]

\[ = \mathbb{E}[\| e_{s_k} \delta_k \| - (DR^x + \gamma D_P^x V_k - D V_k)^2 | V_k] \]

\[ = \mathbb{E}[\| e_{s_k} \delta_k \| - (DR^x + \gamma D_P^x V_k - D V_k)^2 | V_k] \]

\[ = \mathbb{E}[\| e_{s_k} \delta_k \| - (DR^x + \gamma D_P^x V_k - D V_k)^2 | V_k] \]

\[ = \mathbb{E}[\| e_{s_k} \delta_k \|^2 | V_k] \]

\[ = \mathbb{E}[\| e_{s_k} \delta_k \|^2 | V_k] \]

\[ = \mathbb{E}[\| e_{s_k} \delta_k \|^2 | V_k] \]

\[ = \mathbb{E}[\| e_{s_k} \delta_k \|^2 | V_k] \]

where the second inequality is due to the boundedness of rewards in (1). By induction, we have

\[ \| V_k \|_{\infty} \leq 1 + \gamma + \cdots + \gamma^k \leq \sum_{i=0}^{\infty} \gamma^i = \frac{1}{1 - \gamma}, \]

which completes the proof.
APPENDIX C
PROOF OF Lemma 6

We first prove the boundedness of $W_k$ as follows:
\[ \|W_k\|_2 = \|E[w_kw_k^T]\|_2 \leq \|E[w_kw_k^T]\|_2 \leq W_{\text{max}}, \] (13)
where the first inequality is due to Jensen’s inequality, and the last inequality is due to the bound in Lemma 2.

Next, noting
\[ X_k = \alpha^2 \sum_{i=0}^{k-1} A^i W_{k-i-1} (A^T)^i + A^k X_0 (A^T)^k \]
and taking the norm on $X_k$ lead to
\[ \|X_k\|_2 \leq \alpha^2 \sum_{i=0}^{k-1} \|A^i W_{k-i-1} (A^T)^i\|_2 + \|A^k X_0 (A^T)^k\|_2 \]
\[ \leq \alpha^2 \sum_{i=0}^{k-1} \|W_{k-i-1}\|_2 \|A^i (A^T)^i\|_2 + \|X_0\|_2 \|A^k (A^T)^k\|_2 \]
\[ \leq \alpha^2 \sum_{i=0}^{k-1} \|A_i\|^2 + \|X_0\|_2 \|A_i\|_2 \]
\[ \leq \alpha^2 \sum_{i=0}^{k-1} \|A_i\|^2 + \|X_0\|_2 \|A_i\|_2 \]
\[ \leq \alpha^2 \sum_{i=0}^{k-1} \|A_i\|^2 + \|X_0\|_2 \|A_i\|_2 \]
\[ \leq \alpha^2 \sum_{i=0}^{k-1} \|A_i\|^2 + \|X_0\|_2 \|A_i\|_2 \]
where the fourth inequality is due to (13), and the sixth inequality is due to Lemma 4, and the last inequality uses $\rho \in (0, 1)$. This completes the proof.

APPENDIX D
PROOF OF Lemma 7

We first bound $\lambda_{\text{max}}(X_k)$ as follows:
\[ \lambda_{\text{max}}(X_k) \leq \alpha^2 \sum_{i=0}^{k-1} \lambda_{\text{max}}(A^i W_{k-i-1} (A^T)^i) \]
\[ + \lambda_{\text{max}}(A^k X_0 (A^T)^k) \]
\[ \leq \alpha^2 \sum_{i=0}^{k-1} \lambda_{\text{max}}(A^i) \lambda_{\text{max}}(W_{k-i-1}) \]
\[ + \lambda_{\text{max}}(X_0) \lambda_{\text{max}}(A^k) \]
\[ \leq \alpha^2 \sum_{i=0}^{k-1} \|A_i\|^2 + \lambda_{\text{max}}(X_0) \|A_i\|^2 \]
\[ \leq \alpha^2 \sum_{i=0}^{k-1} \|A_i\|^2 + \lambda_{\text{max}}(X_0) \|A_i\|^2 \]
\[ \leq \alpha^2 W_{\text{max}} \sum_{i=0}^{k-1} \|A_i\|^2 + n \lambda_{\text{max}}(X_0) \|A_i\|^2 \]
\[ \leq \alpha^2 W_{\text{max}} \sum_{i=0}^{k-1} \|A_i\|^2 + n \lambda_{\text{max}}(X_0) \|A_i\|^2 \]
\[ \leq \alpha^2 W_{\text{max}} \sum_{i=0}^{k-1} \|A_i\|^2 + n \lambda_{\text{max}}(X_0) \|A_i\|^2 \]
\[ \leq \alpha^2 W_{\text{max}} \sum_{i=0}^{k-1} \|A_i\|^2 + n \lambda_{\text{max}}(X_0) \|A_i\|^2 \]

This completes the proof.

APPENDIX E
PROOF OF Lemma 8

Consider matrix $M$ such that
\[ M = \sum_{k=0}^{\infty} (A^k)^T A^k. \] (14)

Noting that $A^T M A + I = A^T (\sum_{k=0}^{\infty} (A^k)^T A^k) A + I = M$, we conclude that $A^T M A + I = M$, resulting in the desired conclusion. Next, it remains to prove the existence of $M$ by proving its boundedness. In particular, taking the norm on $M$ leads to
\[ \|M\|_2 = \|I + A^T A + (A^2)^T A^2 + \cdots\|_2 \]
\[ \leq \|I\|_2 + \|A^T A\|_2 + \|(A^2)^T A^2\|_2 + \cdots \]
\[ = \|I\|_2 + \|A\|^2 + \|A^2\|^2 + \cdots \]
\[ = 1 + n \|A\|^2 + n \|A^2\|^2 + \cdots \]
\[ = 1 + n \|A\|^2 + n \|A^2\|^2 + \cdots \]
which implies the boundedness. Next, we prove the bounds on the maximum and minimum eigenvalues. From the definition (14), $M \geq I$, and hence $\lambda_{\text{min}}(M) \geq 1$. On the other hand, one gets
\[ \lambda_{\text{max}}(M) = \lambda_{\text{max}}(I + A^T A + (A^2)^T A^2 + \cdots) \]
\[ \leq \lambda_{\text{max}}(I) + \lambda_{\text{max}}((A^2)^T A^2) + \cdots \]
\[ = \lambda_{\text{max}}(I) + \|A\|^2 + \|A^2\|^2 + \cdots \]
\[ \leq 1 + n\|A\|_\infty^2 + n\|A^2\|_\infty^2 + \cdots \]
\[ \leq \frac{n}{1 - \rho^2} \]
\[ \leq \frac{n}{1 - \rho}. \]

The proof is completed.