Convex Programs and Lyapunov Functions for Reinforcement Learning: A Unified Perspective on the Analysis of Value-Based Methods

Xingang Guo and Bin Hu

***Abstract*— Value-based methods play a fundamental role in Markov decision processes (MDPs) and reinforcement learning (RL). In this paper, we present a unified control-theoretic framework for analyzing valued-based methods such as value computation (VC), value iteration (VI), and temporal difference (TD) learning (with linear function approximation). Built upon an intrinsic connection between value-based methods and dynamic systems, we can directly use existing convex testing conditions in control theory to derive various convergence results for the aforementioned value-based methods. These testing conditions are convex programs in form of either linear programming (LP) or semidefinite programming (SDP), and can be solved to construct Lyapunov functions in a straightforward manner. Our analysis reveals some intriguing connections between feedback control systems and RL algorithms. It is our hope that such connections can inspire more work at the intersection of system/control theory and RL.**

I. Introduction

Over the past 10 years, many research ideas have emerged from the fields of control, optimization, and machine learning. A big research focus is on fundamental connections between control systems and iterative algorithms. The research on this topic has led to exciting new results on algorithm analysis and design. For example, iterative optimization methods have been analyzed as feedback control systems [1]–[15], and control-theoretic tools have been leveraged to design new optimization algorithms in various settings [16]–[22]. Recently, there has been an attempt to extend such control perspectives to reinforcement learning (RL). In [23], a fundamental connection between temporal difference learning and Markovian jump linear systems (MJLS) has been established. In [24], the switching system theory has been combined with the ODE method [25], [26] to analyze the asymptotic convergence of Q-learning. More recently, value iteration has also been connected to PID control [27]. Our paper is inspired by these prior results, and establishes a new connection between RL and control theory. Specifically, we tailor various convex testing conditions in control theory for unifying the analysis of value-based algorithms.

RL refers to a collection of techniques for solving Markov decision processes (MDPs), and has shown great promise in many sequential decision making tasks [28]–[30]. Value-based methods including value computation (VC), value iteration (VI), and temporal difference (TD) learning [29] have played a fundamental role in modern RL. The convergence proofs for these methods are typically derived in a case-by-case manner [28]–[37]. Such case-by-case analysis may not be easily generalized. For example, the convergence proof for VI is based on applying the contraction mapping theorem, and requires identifying the right distance metric via deep expert insights. The same distance metric may not be directly used in analyzing other algorithms such as TD learning.

In this paper, we present a unified control-theoretic framework for the convergence analysis of value-based methods. A key observation is that value-based methods can be viewed as dynamical control systems whose behaviors can be directly analyzed using convex programs. In this paper, VC is modeled as a linear time invariant (LTI) positive system, and VI is viewed as a switched positive affine system. In addition, we also borrow the Markovian jump linear system (MJLS) perspective on TD learning from [23]. Notice that there exist many convex testing conditions for analyzing LTI positive systems [38], [39], switched positive systems [40]–[45], and MSLS [46]–[51]. We show that valued-based methods can be analyzed by directly applying the existing linear programming (LP) or semidefinite programming (SDP) conditions from the positive system or MJLS theory. Importantly, we can solve these convex conditions analytically to build our Lyapunov-based proofs in a more transparent manner. It is our hope that the proposed framework can inspire more work at the intersection of system/control theory and RL.

Our analysis makes direct use of existing convex programs in control theory, and complements the work in [23], [24], [32] which rely on other types of stability analysis tools. There are many other convex conditions in control theory, and our work opens the possibility of re-examining these conditions in the context of RL. Compared with [23], our SDP approach has led to new stepsize bounds for TD learning. This result will be given in Section III-C.

II. Preliminaries and Problem Formulation

A. Notation

The set of n-dimensional real vectors is denoted as $\mathbb{R}^n$. The set of $m \times n$ real matrices is denoted as $\mathbb{R}^{m\times n}$. We use $\mathbb{R}^n_+$ to denote the set of the n-dimensional real vectors whose entries are all non-negative. For $x \in \mathbb{R}^n$, we denote its i-th element as $x(i)$. The inequality $x > 0$ ($x \geq 0$) means that $x(i) > 0$ ($x(i) \geq 0$) for all i. For $A \in \mathbb{R}^{n\times n}$, the inequality $A > 0$ ($A \geq 0$) means all the entries of A are positive (non-negative). We use A^T and $\rho(A)$ to denote the transpose and the spectral radius of A, respectively. A matrix A is said to be Schur stable if $\rho(A) < 1$. The inequality $G \succ 0$ ($G \succeq 0$) means that the matrix X is positive (semi-)definite.

Xingang Guo and Bin Hu are with the Coordinated Science Laboratory (CSL) and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. Email: {xingang2, binhu7}@illinois.edu

B. Markov Decision Process and Reinforcement Learning

First, we present some background materials on MDPs and RL. Many decision making tasks can be formulated as MDPs. Consider a MDP defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, P is the transition kernel, R is the reward function, and $\gamma \in (0,1)$ is the discount factor. In this paper, both $\mathcal{S}$ and $\mathcal{A}$ are assumed to be finite. Without loss of generality, we assume $\mathcal{S} = \{1, 2, \ldots, n\}$ and $\mathcal{A} = \{1, 2, \ldots, l\}$. The transition kernel is specified by $P((s,a), s') = \mathbf{P}(s_{k+1} = s' | s_k = s, a_k = a)$.

A policy is a feedback law mapping from states to actions. A policy can be stochastic and maps each state to a probability distribution over $\mathcal{A}$. The goal is to find an optimal policy that maximizes the total accumulated rewards:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R(s_k, a_k) \big| a_k \sim \pi(\cdot|s_k), s_0\right].$$

To obtain an optimal policy, one can solve the optimal value function J^* from the optimal Bellman equation:

$$J^*(s) = \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P((s,a), s') J^*(s') \right). \quad (1)$$

Once J^* is found, one can construct the optimal policy as

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P((s,a), s') J^*(s') \right).$$

The optimal Bellman equation depends on the transition kernel. If the transition model is unknown, RL methods (e.g. TD learning, Q-learning, policy gradient, etc) can be applied.

C. Value Computation

The performance of a given policy π can be evaluated from the associated value function J_π, which is defined as

$$J_\pi(i) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R(s_k, a_k) \big| a_k \sim \pi(\cdot|s_k), s_0 = i\right].$$

For given π, denote the probability transition matrix of $\{s_k\}$ as P_π. Then J_π can be solved from the Bellman equation:

$$J_\pi(i) = R_\pi(i) + \gamma \sum_j P_\pi(i,j) J_\pi(j), \quad (2)$$

where $P_\pi(i,j)$ is the (i,j)-th entry of P_π, and $R_\pi(i)$ is the immediate reward obtained from state i under the policy π. The above Bellman equation can be compactly rewritten as

$$J_\pi = R_\pi + \gamma P_\pi J_\pi. \quad (3)$$

Obviously, J_π can be calculated as $J_\pi = (I - \gamma P_\pi)^{-1} R_\pi$ for any $0 < \gamma < 1$. To avoid matrix inversion, a popular approach for solving J_π is to apply the following iterative value computation (VC) scheme:

$$J_{k+1} = \gamma P_\pi J_k + R_\pi. \quad (4)$$

It is known that the above method is guaranteed to converge to J_π at a linear rate γ. This is actually obvious from the linear system theory. For the right stochastic matrix P_π, we have $\rho(P_\pi) = 1$. Hence the convergence of (4) can be guaranteed by the fact that we have $\rho(\gamma P_\pi) = \gamma \in (0,1)$.

D. Value Iteration

One can solve the optimal value function J^* by recursively applying the Bellman operator $T(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$. This leads to the famous value iteration (VI) method which iterates as

$$J_{k+1}(s) = \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s'} P((s,a), s') J_k(s') \right).$$

A pseudo-code for VI is provided as follows.

Algorithm 1: Value iteration algorithm

Initialization: $J(s) \leftarrow J_0(s), \forall\ s \in \mathcal{S}$;
Repeat
For all $s \in \mathcal{S}$;
$J(s) \leftarrow \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s'} P((s,a), s') J(s') \right)$;
Until J converge

The iteration of VI can be compactly rewritten as

$$J_{k+1} = T(J_k), \quad (5)$$

where $T(\cdot)$ is the Bellman optimality operator. It is known that VI converges to J^* at the rate γ, and a standard way to prove this is to apply the contraction mapping theorem [28].

E. TD Learning with Linear Function Approximation

It is very common that the transition kernel of MDP is unknown. In this case, the VC scheme (4) is not applicable. Instead, TD learning can be used to estimate the value function from sampled trajectories of the underlying Markov chain $\{s_k\}$. Most applications have enormous state spaces, making policy evaluation difficult. Then one needs to incorporate function approximation techniques. Suppose the value function is estimated as $J_\pi(s) \approx \phi(s)^{\mathsf{T}} \theta_\pi$ where ϕ is the feature vector and θ_π is the weight to be estimated. One model-free way to estimate θ_π is to apply the following TD(0) recursion:

$$\theta_{k+1} = \theta_k - \alpha \phi(s_k) \left((\phi(s_k) - \gamma \phi(s_{k+1}))^{\mathsf{T}} \theta_k - R_\pi(s_k) \right),$$

where R_π is the reward, γ is the discount factor, and α is the learning rate. The Markov nature of $\{s_k\}$ has caused trouble for the finite time analysis of the above method. Very recently, various specialized tricks [23], [31], [32] have been developed to address this technical difficulty, leading to several useful finite time results for TD(0) with sufficiently small α.

F. Main Objective: Unified Analysis of VC, VI, and TD

The objective of this work is to develop a simple routine unifying the analysis of VC, VI, and TD(0) with linear function approximation. Built upon the connections between value-based methods and dynamic systems, we can directly use existing convex programs (LP/SDP) in control theory to analyze the above value-based methods. In addition, these convex programs lead to different types of Lyapunov functions, making the Lyapunov-based convergence analysis transparent. Table I summarizes our main results in this work. Our analysis sheds new light on how to combine convex programs and Lyapunov analysis in the context of RL.

TABLE I

CONVEX PROGRAMS FOR VALUE-BASED METHODS

Value-based Algorithms	Type of Dynamic Systems	Convex Programs	Lyapunov functions
Value Computation	LTI Positive system Eq.(6)	LP & SDP (Theorem 1)	Eq. (10)
Value Iteration	Switched positive system Eq.(11)	LP (Condition (13))	Eq. (16)
TD(0) with Linear Function Approximation	MJLS Eq.(22)	SDP (Proposition 3)	Eq. (24)

III. UNIFIED ANALYSIS OF VALUE-BASED METHODS

A. LPs and SDPs for Analyzing VC

To analyze VC, we apply (3) and (4) to rewrite VC as

$$\zeta_{k+1} = A_\pi \zeta_k \tag{6}$$

where $\zeta_k = J_k - J_\pi$, and $A_\pi = \gamma P_\pi \in \mathbb{R}^{n\times n}$. Notice that (6) is actually a positive system since $A_\pi = \gamma P_\pi \geq 0$. To verify the Schur stability of A_π, the following convex conditions for positive linear systems can be directly applied.

Proposition 1: Suppose $A_\pi \geq 0$. Then each of the following conditions provides a necessary and sufficient condition for the stability of (6):

1) $\exists\ \xi \in \mathbb{R}^n$ s.t. $\xi > 0,$ and $A_\pi \xi - \xi < 0,$
2) $\exists\ \nu \in \mathbb{R}^n$ s.t. $\nu > 0,$ and $\nu^\mathsf{T} A_\pi - \nu^\mathsf{T} < 0,$
3) $\exists\ G \in \mathbb{R}^{n\times n}$ s.t. $G \succ 0$ $A_\pi^\mathsf{T} G A_\pi \prec G.$

Proof: This result is just the discrete-time counterpart of Proposition 1 in [39] and can be proved similarly. ■

The above convex programs can be solved to obtain three types of Lyapunov functions for (6). Let 1_n denote the n-dimensional vector whose entries are all equal to 1. Then our results can be stated as follows.

Theorem 1: Consider the recursion (6). Set (ξ, ν, G) as

$$\xi = 1_n,\ \ \nu = \omega,\ \ G = \mathrm{diag}\left(\frac{\nu(1)}{\xi(1)}, \cdots, \frac{\nu(n)}{\xi(n)}\right),$$

where ω is the stationary distribution of P_π. Then we have

$$\xi > 0 \text{ and } A_\pi \xi \leq \gamma \xi, \tag{7}$$

$$\nu > 0 \text{ and } \nu^\mathsf{T} A_\pi \leq \gamma \nu^\mathsf{T}, \tag{8}$$

$$G \succ 0 \text{ and } A_\pi^\mathsf{T} G A_\pi \preceq \gamma^2 G. \tag{9}$$

This leads to the following three types of Lyapunov functions

$$V_1(\zeta) = \max_i |\zeta(i)|, \quad V_2(\zeta) = |\nu^\mathsf{T}\zeta|, \quad V_3(\zeta) = \zeta^\mathsf{T} G \zeta, \tag{10}$$

which satisfy $V_1(\zeta_k) \leq C_1\gamma^k$, $V_2(\zeta_k) \leq C_2\gamma^k$, and $V_3(\zeta_k) \leq C_3\gamma^{2k}$ for some fixed positive constants (C_1, C_2, C_3).

Proof: Since P_π is always a right stochastic matrix, we have $P_\pi 1_n = 1_n$ and $\omega^\mathsf{T} P_\pi = \omega^\mathsf{T}$. Therefore,

$$A_\pi 1_n - \gamma 1_n = \gamma(P_\pi 1_n - 1_n) = 0$$
$$\omega^\mathsf{T} A_\pi - \gamma \omega^\mathsf{T} = \gamma\left(\omega^\mathsf{T} P_\pi - \omega^\mathsf{T}\right) = 0.$$

Hence (7) and (8) hold. The third condition can be proved as discussed by Proposition 2 in [39]. The rest of the results follow from standard arguments in positive system theory. ■

Notice that the conditions (7) (8) are LPs, and (9) can be solved as SDPs. Theorem 1 provides three different types of Lyapunov functions for the positive system (6):

1) ℓ_∞-type: $V_1(\zeta) = \max_i |\zeta(i)|$;
2) Linear-type copositive: $V_2(\zeta) = |\nu^\mathsf{T}\zeta|$;
3) Quadratic: $V_3(\zeta) = \zeta^\mathsf{T} G \zeta$.

It is trivial to show $V_1(\zeta_{k+1}) \leq \gamma V_1(\zeta_k)$ and $V_3(\zeta_{k+1}) \leq \gamma^2 V_3(\zeta_k)$. For $i = 2$, the Lyapunov function is copositive and works slightly differently. Here we briefly explain how it works. For any $\zeta_0 \in \mathbb{R}^n$, $\exists \zeta_0^+, \zeta_0^- \in \mathbb{R}^n_+$ s.t. $\zeta_0 = \zeta_0^+ - \zeta_0^-$. Let $\{\zeta_k^+\}$ and $\{\zeta_k^-\}$ be the state trajectories of (6) initialized from ζ_0^+ and ζ_0^-, respectively. By linearity, we have $\zeta_k = \zeta_k^+ - \zeta_k^-$. Based on the condition (8), we can show $V_2(\zeta_k) \leq V_2(\zeta_k^+) + V_2(\zeta_k^-) \leq \gamma^k(V_2(\zeta_0^+) + V_2(\zeta_0^-))$. This ensures the convergence of VC.

A key message from the above analysis is that the Lyapunov function construction for positive linear systems can be simpler than general LTI systems due to the use of LPs. Since the construction of the max-type (or ℓ_∞-type) Lyapunov function $V_1(\cdot)$ is independent of the underlying policy, we may construct an ℓ_∞-type common Lyapunov function for cases where the policy is changing over time.

B. LPs and Common Lyapunov Functions for VI

Next we establish the connection between VI and switched positive affine systems. This will lead to LP conditions for analyzing VI. The VI scheme $J_{k+1} = T(J_k)$ can be recast as

$$J_{k+1} = \gamma P_{\sigma_k} J_k + R_{\sigma_k} \tag{11}$$

where $\sigma_k \in \{1, 2, \ldots, l^n\}$. Recall l and n denote the size of action space and state space, respectively. When $\sigma_k = m$, we set $P_{\sigma_k} = P_m$ and $R_{\sigma_k} = R_m$. For all $m \in \{1, 2, \cdots, l^n\}$, P_m is an $n \times n$ matrix whose i-th row (for all i) is a row vector in the form of $\begin{bmatrix} P((i,a),1) & P((i,a),2) & \ldots & P((i,a),n)\end{bmatrix}$ with some $a \in \mathcal{A}$. Similarly, for all m, the vector R_m is a vector whose i-th element (for all i) is equal to $R(i,a)$ for some $a \in \mathcal{A}$. The total number of the (P_m, R_m) pairs is l^n, and we denote the set of all such pairs as $\mathbf{\Lambda}$. Therefore, we can just view (11) as a switched positive affine system, and it is not that surprising that we can analyze VI via switched system theory.

For ease of exposition, we first address the case where $R_m = 0$ for all m. In this case, we have $J^* = 0$, and (11) can be rewritten as a switched positive linear system:

$$J_{k+1} = A_m J_k, \quad m \in \{1, 2, \cdots, l^n\}, \tag{12}$$

where $A_m = \gamma P_m \in \mathbb{R}^{n\times n}$. A well-known fact is that the system state of (12) may diverge for some switching sequence even when A_m is Schur stable for all m [52]. The stability guarantees for (12) are typically obtained by extending the Lyapunov approach presented in Proposition 1. One way is to use the common Lyapunov function (CLF).

Proposition 2: Suppose $A_m \geq 0$ for all m. Then each of the following conditions provides a sufficient condition for the stability of the switched positive system (12):

1) $\exists \xi \in \mathbb{R}^n$ s.t. $\xi > 0$ and $A_m \xi - \xi < 0$ for all A_m.
2) $\exists \nu \in \mathbb{R}^n$ s.t. $\nu > 0$ and $\nu^\mathsf{T} A_m - \nu^\mathsf{T} < 0$ for all A_m.
3) $\exists$ a matrix $G \succ 0$ s.t. $A_m^\mathsf{T} G A_m - G \prec 0$ for all A_m.

Proof: The proof is standard. The third condition actually does not require A_m to be positive. If $A_m^\mathsf{T} G A_m - G \prec 0$, then $\exists\ \epsilon > 0$ such that $A_m^\mathsf{T} G A_m - (1-\epsilon) G \preceq 0$. Hence we can define a Lyapunov function $V(J_k) = J_k^\mathsf{T} G J_k$ satisfying $V(J_{k+1}) \leq (1-\epsilon) V(J_k)$. This ensures the stability of (12). The first and second conditions do require A_m to be positive, and can be proved similarly. See [44], [45] for details. ∎

Similar to Theorem 1, the testing conditions in Proposition 2 can be modified to analyze convergence rates of (12). One will obtain similar rate bounds as presented in Theorem 1 if any of the following is feasible:

$$\exists\ \xi \in \mathbb{R}^n\ s.t.\ \xi > 0 \text{ and } A_m \xi \leq \gamma \xi\ \forall\ m. \tag{13}$$

$$\exists\ \nu \in \mathbb{R}^n\ s.t.\ \nu > 0 \text{ and } \nu^\mathsf{T} A_m \leq \gamma \nu^\mathsf{T}\ \forall\ m. \tag{14}$$

$$\exists\ G \in \mathbb{R}^{n \times n}\ s.t.\ G \succ 0 \text{ and } A_m^\mathsf{T} G A_m \preceq \gamma^2 G\ \forall\ m. \tag{15}$$

It is interesting to see that in general, the system (12) does not have a common linear copositive Lyapunov function since the stationary distributions for different P_m are typically not the same. It also seems difficult to construct a common solution G for the SDP (15). However, since all P_m share the same right eigenvector 1_n, we have $\gamma P_m 1_n = \gamma 1_n$ for all m. Hence we can solve (13) to obtain an ℓ_∞-type CLF:

$$V(J_k) = \|J_k - J^*\|_\infty. \tag{16}$$

Condition (13) can be used to guarantee $V(J_k) \leq \gamma^k V(J_0)$.

Now, we can extend the above analysis to the general case where $R_m \neq 0$. In this case, we will show that the iterations of VI can be upper and lower bounded by the trajectories of two stable positive linear systems. Hence positive system theory can still be applied. We need the following lemma.

Lemma 1: Consider the switched positive affine system (11) with a switching sequence $\{\sigma_k\}$ completely determined by the Bellman optimality operator[1]. Then the following inequality holds for all k

$$\gamma P^*(J_k - J^*) \leq J_{k+1} - J^* \leq \gamma P_{\sigma_k}(J_k - J^*). \tag{17}$$

Proof: Since $J^* = T(J^*)$, there exists a pair $(P^*, R^*) \in \mathbf{\Lambda}$ such that $J^* = \gamma P^* J^* + R^*$. By the definition of the Bellman optimality operator, one can show that the following two inequalities holds for all m:

$$\gamma P_m J^* + R_m \leq \gamma P^* J^* + R^*, \tag{18}$$

$$\gamma P_m J_k + R_m \leq \gamma P_{\sigma_k} J_k + R_{\sigma_k}. \tag{19}$$

Using (18), (19), and the fact that $J^* = \gamma P^* J^* + R^*$, one can verify $\gamma P^*(J_k - J^*) \leq J_{k+1} - J^* \leq \gamma P_{\sigma_k}(J_k - J^*)$. This leads to the desired conclusion. ∎

Based on Lemma 1, we obtain the following main result.

[1]In other words, the trajectory of such a switched system now exactly matches the sequence generated by the VI method.

Theorem 2: Consider the switched positive affine system (11) with a switching sequence $\{\sigma_k\}$ completely determined by the Bellman optimality operator T. Suppose the sequence $\{J_k^u\}$ is generated by the system $J_{k+1}^u - J^* = \gamma P_{\sigma_k}(J_k^u - J^*)$ with the same switching sequence $\{\sigma_k\}$. Let the sequence $\{J_k^o\}$ be generated by the system $J_{k+1}^o - J^* = \gamma P^*(J_k^o - J^*)$. Suppose $J_0 = J_0^u = J_0^o$. Then we have

$$J_k^o - J^* \leq J_k - J^* \leq J_k^u - J^*,\ \forall k \tag{20}$$

Proof: This theorem can be proved using induction. When $k = 0$, it is straightforward to verify that (20) holds as a consequence of Lemma 1. Suppose (20) holds for $k = t$. For $k = t+1$, we can apply Lemma 1 to show

$$J_{t+1} - J^* \leq \gamma P_{\sigma_k}(J_t - J^*) \leq \gamma P_{\sigma_k}(J_t^u - J^*) = J_{t+1}^u - J^*$$

where the second step follows from the fact that P_{σ_k} is right stochastic. Based on Lemma 1, we can use a similar argument to show $J_{t+1} - J^* \geq J_t^o - J^*$. Hence (20) holds for $k = t+1$. This completes the proof. ∎

Therefore, we can directly apply the LP condition (13) to construct an ℓ_∞-type CLF for VI and prove the rate bound $\|J_k - J^*\|_\infty \leq \max\{\|J_k^u - J^*\|_\infty, \|J_k^o - J^*\|_\infty\} \leq \gamma^k \|J_0 - J^*\|_\infty$. This demonstrates how to apply the simple LP condition (13) to analyze VI.

C. SDPs for TD(0) with Linear Function Approximation

In this section, we provide SDP-based finite time analysis for TD(0) with linear function approximation. Since TD(0) can be viewed as a MJLS, the SDP-based stability conditions for MJLS can be directly applied. Recall that TD(0) (with linear function approximation) follows the update rule $\theta_{k+1} = \theta_k - \alpha\phi(s_k)\left((\phi(s_k) - \gamma\phi(s_{k+1}))^\mathsf{T}\theta_k - R_\pi(s_k)\right)$, where ϕ is the feature vector, and θ is the weight to be estimated. We can augment $\begin{bmatrix} s_{k+1}^\mathsf{T} & s_k^\mathsf{T} \end{bmatrix}^\mathsf{T} \in \mathcal{S} \oplus \mathcal{S}$ as a new vector z_k. Obviously, there is a one-to-one mapping from $\mathcal{S} \oplus \mathcal{S}$ to the set $\mathcal{N} = \{1, 2, \cdots, n^2\}$. Without loss of generality, $\{z_k\}$ can be set up as a Markov chain sampled from $\mathcal{N}$. Suppose θ_π is the solution to the projected Bellman equation for the fixed policy π. Due to the one-to-one correspondence between $\begin{bmatrix} s_{k+1}^\mathsf{T} & s_k^\mathsf{T} \end{bmatrix}^\mathsf{T}$ and z_k, the iteration of TD(0) can be recast as

$$\theta_{k+1} - \theta_\pi = \theta_k - \theta_\pi + \alpha\left(A_{z_k}(\theta_k - \theta_\pi) + b_{z_k}\right), \tag{21}$$

where $A_{z_k} = \phi(s_k)(\gamma\phi(s_{k+1}) - \phi(s_k))^\mathsf{T}$ and $b_{z_k} = \phi(s_k)\left(R_\pi(s_k) + (\phi(s_k) - \gamma\phi(s_{k+1}))^\mathsf{T}\theta^\pi\right)$. When $z_k = i \in \mathcal{N}$, we have $A_{z_k} = A_i$ and $b_{z_k} = b_i$. We denote $\zeta_k = \theta_k - \theta_\pi$. Then (21) can be rewritten as the following MJLS:

$$\zeta_{k+1} = H_{z_k}\zeta_k + \alpha b_{z_k} u_k. \tag{22}$$

where $H_{z_k} = I + \alpha A_{z_k}$, and $u_k = 1\ \forall k$. When $z_k = i \in \mathcal{N}$, we have $H_{z_k} = H_i$. Denote $p_{ij} = \mathbf{P}(z_{k+1} = j | z_k = i)$, and $N = n^2$. Then the following mean square stability condition [46]–[48] can be directly applied to analyze (22).

Proposition 3: The MJLS (22) is mean square stable (MSS) if and only if there exist matrices $G_i \succ 0$ for $i = 1, \cdots, N$ such that the following SDP is feasible:

$$G_i - H_i^\mathsf{T}\left(\sum_{j=1}^N p_{ij} G_j\right) H_i \succ 0, \text{for } i = 1, \cdots, N. \tag{23}$$

Proof: This stability condition is well known. For more details, see discussions in [46] or [47]. ■

There are multiple ways to prove the mean square stability of (22) from the SDP condition (23). One way is to construct the following quadratic Lyapunov function from $\{G_i\}$:

$$V(\zeta_k) = \mathbb{E}\left[\zeta_k^\mathsf{T} G_{z_k} \zeta_k\right]. \tag{24}$$

Once the MJLS (22) is shown to be MSS, Theorem 3.33 in [46] can be applied to show that (22) is also asymptotically wide sense stationary, and then the mean square TD error can be exactly calculated via Proposition 3.35 in [46]. As a matter of fact, the convergence bounds in Corollary 2 of [23] can be directly applied whenever (22) is MSS. Therefore, the finite time analysis of TD(0) boils down to checking the mean square stability of the MJLS (22). Next, we show how to construct solutions for the SDP condition (23) under the following standard assumption.

Assumption 1: Suppose $\{z_k\}$ is irreducible and aperiodic. Denote $p_i^\infty = \lim_{k\to\infty} \mathbf{P}(z_k = i)$ and $\bar{A} = \sum_{i=1}^N p_i{}^\infty A_i$. We assume $\bar{A}$ is Hurwitz, and $\sum_{i=1}^N p_i^\infty b_i = 0$.

Under Assumption 1, let $\bar{G}$ be the solution to the Lyapunov equation $\bar{A}^\mathsf{T}\bar{G} + \bar{G}\bar{A} = -I$. We also denote $X_i = A_i^\mathsf{T}\bar{G} + \bar{G}A_i + I/(p_i^\infty N)$. Now we can state the following result.

Lemma 2: Supposed Assumption 1 is given. For sufficiently small α, we can solve the SDP (23) by choosing $G_i = \bar{G} + \alpha\tilde{G}_i$, where $\tilde{G}_N = 0$ and $\tilde{G}_i$ (for $i = 1, \cdots, N-1$) is solved from the following linear equation:

$$\tilde{G}_i - \sum_{j=1}^{N-1} p_{ij}\tilde{G}_j = X_i, \text{ for } i = 1, \cdots, N-1. \tag{25}$$

Proof: First, notice that (25) does have a unique solution. To see this, let $\hat{P} \in \mathbb{R}^{(N-1)\times(N-1)}$ be a substochastic matrix whose (i,j)-th entry is equal to p_{ij}. Then $\hat{P}$ is a submatrix of the transition matrix of $\{z_k\}$, and has a spectral radius which is smaller than 1. Hence $(I_{N-1} - \hat{P})$ is invertible, and (25) admits a unique well-defined solution. Since X_i is symmetric for all i, the resultant matrices $\{\tilde{G}_i\}$ are also symmetric. Now we briefly explain our choices of G_i. If we substitute $G_i = \bar{G} + \alpha\tilde{G}_i$ into (23), we get

$$\tilde{G}_i - \sum_{j=1}^{N} p_{ij}\tilde{G}_j - (A_i^\mathsf{T}\bar{G} + \bar{G}A_i) + O(\alpha) \succ 0, \forall i \tag{26}$$

For $i = 1, \cdots, N-1$, we can substitute (25) and $\tilde{G}_N = 0$ into (26) to simplify it as $I/(p_i^\infty N) + O(\alpha) \succ 0$, which clearly holds for sufficiently small α. For $i = N$, we can use the fact $\bar{A}^\mathsf{T}\bar{G} + \bar{G}\bar{A} = -I$ to show

$$A_N^\mathsf{T}\bar{G} + \bar{G}A_N = \frac{1}{p_N^\infty}\left(-I - \sum_{i=1}^{N-1} p_i^\infty (A_i^\mathsf{T}\bar{G} + \bar{G}A_i)\right)$$

We have $A_i^\mathsf{T}\bar{G} + \bar{G}A_i = \tilde{G}_i - \sum_{j=1}^{N-1} p_{ij}\tilde{G}_j - I/(p_i^\infty N)$ for $i < N$ (see (25)). Substituting these into (26) for $i = N$ leads to $I/(p_N^\infty N) + O(\alpha) \succ 0$, which holds for small α. ■

Next, we provide an explicit upper bound on α. To make sure that $\{\bar{G} + \alpha\tilde{G}_i\}$ solves the SDP condition (23), we need

$$\bar{G} + \alpha\tilde{G}_i \succ 0, \;\; I/(p_i^\infty N) + \alpha M_i + \alpha^2 \tilde{M}_i \succ 0, \;\; \forall i \tag{27}$$

where $\tilde{M}_i = -A_i^\mathsf{T}(\sum_{j=1}^N p_{ij}\tilde{G}_j)A_i$, and $M_i = -A_i^\mathsf{T}\bar{G}A_i - A_i^\mathsf{T}(\sum_{j=1}^N p_{ij}\tilde{G}_j) - (\sum_{j=1}^N p_{ij}\tilde{G}_j)A_i$. We know $\bar{G} \succ 0$ and $I/(p_i^\infty N) \succ 0$. Notice $\{\tilde{G}_i\}$, $\{M_i\}$, and $\{\tilde{M}_i\}$ are symmetric matrices which are completely determined by $\{A_i\}$ and p_{ij}. Hence the SDP condition (23) is feasible with $G_i = \bar{G} + \alpha\tilde{G}_i$ if for all $i \in \mathcal{N}$, α satisfies $\lambda_{\min}(\bar{G}) + \alpha\lambda_{\min}(\tilde{G}_i) > 0$ and

$$1/(p_i^\infty N) + \alpha\lambda_{\min}(M_i) + \alpha^2\lambda_{\min}(\tilde{M}_i) > 0. \tag{28}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue. Let $\mathbf{1}_\mathcal{D}$ denote the indicator function for any set $\mathcal{D}$. We have $\lambda_{\min}(\bar{G}) + \alpha\lambda_{\min}(\tilde{G}_i) > 0$ for any $0 < \alpha < \frac{\lambda_{\min}(\bar{G})}{|\lambda_{\min}(\tilde{G}_i)|(1-\mathbf{1}_{\tilde{G}_i \succeq 0})}$. When $\tilde{G}_i \succeq 0$, this bound becomes $+\infty$. It is also straightforward to verify that (28) is true for any $0 < \alpha < \bar{\alpha}_i$, where $\bar{\alpha}_i$ is defined as $\bar{\alpha}_i = \frac{1}{p_i^\infty N|\lambda_{\min}(M_i)|(1-\mathbf{1}_{M_i \succeq 0})}$ if $\tilde{M}_i \succeq 0$, and $\bar{\alpha}_i = \frac{-\lambda_{\min}(M_i) - \sqrt{\lambda_{\min}^2(M_i) - 4\lambda_{\min}(\tilde{M}_i)/(p_i^\infty N)}}{2\lambda_{\min}(\tilde{M}_i)}$ otherwise. This leads to the following result.

Theorem 3: Given Assumption 1, the TD(0) method (21) with step size $0 < \alpha < \min_{i\in\mathcal{N}}\left\{\bar{\alpha}_i, \frac{\lambda_{\min}(\bar{G})}{|\lambda_{\min}(\tilde{G}_i)|(1-\mathbf{1}_{\tilde{G}_i \succeq 0})}\right\}$ is MSS, and the mean square estimation error $\mathbb{E}\|\theta_k - \theta_\pi\|^2$ converges exponentially to its stationary value.

Proof: From the above discussion, our stepsize bound can guarantee (27), and hence (21) is MSS. Then the convergence behavior of $\mathbb{E}\|\theta_k - \theta_\pi\|^2$ can be shown using Proposition 3.35 of [46] or Corollary 2 of [23]. ■

With the MSS property, we can directly apply Corollary 2 in [23] to obtain explicit formulas for the convergence rate and the steady state error. We skip those formulas. Clearly, our result is closely related to [23] which also analyzes TD learning using MJLS theory. A key difference is that the analysis in [23] boils down to an LTI system formulation without exploiting the SDP (23). Our SDP approach brings a new benefit in providing an explicit stepsize bound guaranteeing MSS, as specified by Theorem 3. In contrast, the analysis in [23] relies on advanced eigenvalue perturbation theory and only shows that TD(0) is MSS for sufficiently small α without providing such explicit stepsize bounds.

IV. Conclusion and Future Work

In this paper, we show that existing convex programs in control theory can be directly used to analyze value-based methods such as VC, VI, and TD(0) with linear function approximation. It is possible that these convex programs can be extended to address the impacts of computation error and delay. This will be investigated in the future.

Acknowledgment

This work is generously supported by the NSF award CAREER-2048168 and the 2020 Amazon research award.

References

[1] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2016.

[2] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan, "A general analysis of the convergence of ADMM," in *International Conference on Machine Learning*, 2015, pp. 343–352.

[3] B. Hu and L. Lessard, "Dissipativity theory for Nesterov's accelerated method," in *International Conference on Machine Learning*, vol. 70, 2017, pp. 1549–1557.

[4] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, "Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2654–2689, 2018.

[5] A. Sundararajan, B. Hu, and L. Lessard, "Robust convergence analysis of distributed optimization algorithms," in *Annual Allerton Conference on Communication, Control, and Computing*, 2017, pp. 1206–1212.

[6] B. Hu and L. Lessard, "Control interpretations for first-order optimization methods," in *American Control Conference*, 2017, pp. 3114–3119.

[7] B. Hu, P. Seiler, and A. Rantzer, "A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints," in *Conference on Learning Theory*, vol. 65, 2017, pp. 1157–1189.

[8] T. Hatanaka, N. Chopra, T. Ishizaki, and N. Li, "Passivity-based distributed optimization with communication delays using PI consensus algorithm," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4421–4428, 2018.

[9] B. Hu, S. Wright, and L. Lessard, "Dissipativity theory for accelerating stochastic variance reduction: A unified analysis of SVRG and Katyusha using semidefinite programs," in *International Conference on Machine Learning*, 2018, pp. 2043–2052.

[10] S. Han, "Systematic design of decentralized algorithms for consensus optimization," *IEEE Control Systems Letters*, vol. 3, no. 4, pp. 966–971, 2019.

[11] N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar, "Robust accelerated gradient methods for smooth strongly convex functions," *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 717–751, 2020.

[12] H. Xiong, Y. Chi, B. Hu, and W. Zhang, "Analytical convergence regions of accelerated gradient descent in nonconvex optimization under regularity condition," *Automatica*, vol. 113, 2020.

[13] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Robustness of accelerated first-order algorithms for strongly convex optimization problems," *IEEE Transactions on Automatic Control*, vol. 66, no. 6, pp. 2480–2495, 2020.

[14] B. Hu, P. Seiler, and L. Lessard, "Analysis of biased stochastic gradient descent using sequential semidefinite programs," *Mathematical Programming*, vol. 187, no. 1, pp. 383–408, 2021.

[15] O. Gannot, "A frequency-domain analysis of inexact gradient methods," *Mathematical Programming*, pp. 1–42, 2021.

[16] B. Van Scoy, R. Freeman, and K. Lynch, "The fastest known globally convergent first-order method for minimizing strongly convex functions," *IEEE Control Systems Letters*, vol. 2, no. 1, pp. 49–54, 2017.

[17] S. Cyrus, B. Hu, B. Van Scoy, and L. Lessard, "A robust accelerated optimization algorithm for strongly convex functions," in *American Control Conference*, 2018, pp. 1376–1381.

[18] M. Fazlyab, M. Morari, and V. M. Preciado, "Design of first-order optimization algorithms via sum-of-squares programming," in *IEEE Conference on Decision and Control*, 2018, pp. 4445–4452.

[19] Z. Nelson and E. Mallada, "An integral quadratic constraint framework for real-time steady-state optimization of linear time-invariant systems," in *American Control Conference*, 2018, pp. 597–603.

[20] N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar, "A universally optimal multistage accelerated stochastic gradient method," in *Advances in Neural Information Processing Systems*, 2019, pp. 8525–8536.

[21] S. Michalowsky, C. Scherer, and C. Ebenbauer, "Robust and structure exploiting optimisation algorithms: an integral quadratic constraint approach," *International Journal of Control*, pp. 1–24, 2020.

[22] A. Sundararajan, B. Van Scoy, and L. Lessard, "Analysis and design of first-order distributed optimization algorithms over time-varying graphs," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 4, pp. 1597–1608, 2020.

[23] 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. 8479–8490.

[24] D. Lee and N. He, "A unified switching system perspective and convergence analysis of Q-learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 15 556–15 567.

[25] V. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.

[26] V. Borkar and S. 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.

[27] A. Farahmand and M. Ghavamzadeh, "PID accelerated value iteration algorithm," in *International Conference on Machine Learning*, 2021, pp. 3143–3153.

[28] M. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[29] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[30] D. Bertsekas and J. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific Belmont, 1996, vol. 5.

[31] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Conference on learning theory*, 2018, pp. 1691–1692.

[32] R. Srikant and L. Ying, "Finite-time error bounds for linear stochastic approximation and TD learning," in *Conference on Learning Theory*, 2019, pp. 2803–2830.

[33] T. Xu, S. Zou, and Y. Liang, "Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples," in *Advances in Neural Information Processing Systems*, 2019.

[34] J. Sun, G. Wang, G. B. Giannakis, Q. Yang, and Z. Yang, "Finite-time analysis of decentralized temporal-difference learning with linear function approximation," in *AIstats*, 2020, pp. 4485–4495.

[35] P. Xu and Q. Gu, "A finite-time analysis of Q-learning with neural network function approximation," in *International Conference on Machine Learning*, 2020, pp. 10 555–10 565.

[36] S. Zhang, Z. Zhang, and S. T. Maguluri, "Finite sample analysis of average-reward TD learning and Q-learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[37] T. T. Doan, "Finite-time analysis and restarting scheme for linear two-time-scale stochastic approximation," *SIAM Journal on Control and Optimization*, vol. 59, no. 4, pp. 2798–2819, 2021.

[38] L. Farina and S. Rinaldi, *Positive linear systems: theory and applications*. John Wiley & Sons, 2011.

[39] A. Rantzer, "Distributed control of positive systems," in *IEEE Conference on Decision and Control and European Control Conference*, 2011, pp. 6608–6611.

[40] F. Blanchini, P. Colaneri, M. E. Valcher *et al.*, "Switched positive linear systems," *Foundations and Trends® in Systems and Control*, vol. 2, no. 2, pp. 101–273, 2015.

[41] X. Liu, "Stability analysis of switched positive systems: a switched linear copositive Lyapunov function method," *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 56, no. 5, pp. 414–418, 2009.

[42] O. Mason and R. Shorten, "On linear copositive Lyapunov functions and the stability of switched positive linear systems," *IEEE Transactions on Automatic Control*, vol. 52, no. 7, pp. 1346–1349, 2007.

[43] Y. Xu, J. Dong, R. Lu, and L. Xie, "Stability of continuous-time positive switched linear systems: A weak common copositive Lyapunov functions approach," *Automatica*, vol. 97, pp. 278–285, 2018.

[44] E. Fornasini and M. Valcher, "Stability and stabilizability criteria for discrete-time positive switched systems," *IEEE Transactions on Automatic control*, vol. 57, no. 5, pp. 1208–1221, 2011.

[45] O. C. Pastravanu and M.-H. Matcovschi, "Max-type copositive Lyapunov functions for switching positive linear systems," *Automatica*, vol. 50, no. 12, pp. 3323–3327, 2014.

[46] O. Costa, M. Fragoso, and R. Marques, *Discrete-time Markov jump linear systems*. Springer Science & Business Media, 2006.

[47] O. Costa and M. Fragoso, "Stability results for discrete-time linear systems with Markovian jumping parameters," *Journal of Mathematical Analysis and Applications*, vol. 179, no. 1, pp. 154–178, 1993.

[48] L. El Ghaoui and M. Rami, "Robust state-feedback stabilization of jump linear systems via LMIs," *International Journal of Robust and Nonlinear Control*, vol. 6, no. 9-10, pp. 1015–1022, 1996.

[49] Y. Fang and K. Loparo, "Stochastic stability of jump linear systems," *IEEE Transactions on Automatic Control*, vol. 47, no. 7, pp. 1204–1208, 2002.

[50] Y. Ji, H. Chizeck, X. Feng, and K. Loparo, "Stability and control of discrete-time jump linear systems," *Control-Theory and Advanced Technology*, vol. 7, no. 2, pp. 247–270, 1991.

[51] P. Seiler and R. Sengupta, "A bounded real lemma for jump systems," *IEEE Transactions on Automatic Control*, vol. 48, no. 9, pp. 1651–1654, 2003.

[52] R. DeCarlo, M. Branicky, S. Pettersson, and B. Lennartson, "Perspectives and results on the stability and stabilizability of hybrid systems," *Proceedings of the IEEE*, vol. 88, no. 7, pp. 1069–1082, 2000.