On Some Geometric Behavior of Value Iteration on the Orthant: Switching System Perspective

Donghwan Lee

Abstract— In this paper, the primary goal is to offer additional insights into the value iteration through the lens of switching system models in the control community. These models establish a connection between value iteration and switching system theory and reveal additional geometric behaviors of value iteration in solving discounted Markov decision problems. Specifically, the main contributions of this paper are twofold: 1) We provide a switching system model of value iteration and, based on it, offer a different proof for the contraction property of the value iteration. 2) Furthermore, from the additional insights, new geometric behaviors of value iteration are proven when the initial iterate lies in a special region. We anticipate that the proposed perspectives might have the potential to be a useful tool, applicable in various settings. Therefore, further development of these methods could be a valuable avenue for future research.

I. INTRODUCTION

Dynamic programming [1], [2] is a general and effective methodology for finding an optimal solution for Markov decision problems [3]. Value iteration is a popular class of dynamic programming algorithms, and its exponential convergence is a fundamental and well-established [1], [2] result through the contraction mapping property of the Bellman operator.

In this paper, we offer some additional insights on the value iteration based on switching system models [4]–[11] in the control community. These models provide new geometric behaviors of value iteration to solve Markov decision problems [3]. In particular, the main contributions of this paper are twofold: 1) We present a switching system model of value iteration and, based on it, provide a different proof for the contraction property of value iteration. 2) Furthermore, from the new perspectives and frameworks, we study additional geometric behaviors of value iteration when the initial iterate lies in a special region. More specifically, the detailed contributions are summarized as follows:

1) A switching system model of value iteration is introduced along with its special properties. Based on it, we offer a different proof for the following contraction property of the value iteration:

$$\|Q_{k+1} - Q^*\|_\infty \le \gamma \|Q_k - Q^*\|_\infty, \qquad (1)$$

D. Lee is with the Department of Electrical and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, South Korea donghwan@kaist.ac.kr.

where Q_k is the kth iterate of the value iteration, Q^* is the optimal Q-function, and $\gamma \in (0, 1)$ is the so-called discount factor in the underlying Markov decision problem. Although this result is fundamental and classical, the proof technique utilized in this paper is distinct from the classical approaches, offering additional insights.

2) Furthermore, from the additional insights provided in the first part, we prove new geometric behaviors of value iteration when the initial iterate lies in a special region. Specifically, when the initial iterate Q_0 is within the set $Q_0 \le Q^*$ (shifted orthant), the iterates of the value iteration remain in the set, displaying different behaviors than the cases with a general initial iterate. When $Q_0 \le Q^*$, one can establish the new contraction property

$$\|Q_{k+1} - Q^*\|_M \le (\gamma + \varepsilon) \|Q_k - Q^*\|_M \qquad (2)$$

where $\|(\cdot)\|_M = \sqrt{(\cdot)^T M (\cdot)}$ is the weighted Euclidean norm, $\varepsilon > 0$ is any real number such that $\gamma + \varepsilon \in (0, 1)$, and M is a positive definite matrix dependent on $\varepsilon > 0$. This result corresponds to a contraction mapping property of the Bellman operator in terms of the weighted Euclidean norm, which cannot be derived from the classical approaches. Geometrically, the sublevel sets corresponding to the infinity norm $\|\cdot\|_\infty$ are squares, while the sublevel sets corresponding to the weighted Euclidean norm $\|\cdot\|_M$ are ellipsoids.
A more notable result is that when $Q_0 \le Q^*$, the iterates satisfy the linear inequality

$$(\gamma + \varepsilon)v^T(Q_k - Q^*) \le v^T(Q_{k+1} - Q^*) \le 0, \quad \forall k \ge 0, \tag{3}$$

where $\varepsilon > 0$ is any real number such that $\gamma + \varepsilon \in (0, 1)$, and v is a positive vector dependent on ε. Geometrically, this implies that the component along the direction of v in $Q_k - Q^*$ diminishes exponentially.

The analysis presented in this paper employs insights and tools from switching systems [4]–[11], positive switching systems [12]–[15], nonlinear systems [16], and linear systems [17]. We emphasize that our goal in this paper is to offer additional insights and analysis templates for value iteration via its connections to switching systems, rather than improving existing convergence rates. We anticipate that the proposed switching system perspectives may have the potential to be a useful tool, applicable in various settings, and stimulate the synergy between dynamic system theory and dynamic programming. Thus, further development of these methods could be a valuable avenue for future research. Finally, some ideas in this paper have been inspired

by [11], [18], which develop switching system models of Q-learning [19]. However, most technical results and derivation processes are nontrivial and entirely different from those in [11], [18]. Another related work is [20], which provides a linear programming-based analysis of value iteration and also employs positive switching system models for their analysis. However, the previous work merely proposed a numerical linear programming-based analysis to verify convergence properties and did not obtain the explicit relations given in (1), (2), (3). Therefore, our analysis offers significantly different results than those in [20].

II. PRELIMINARIES

A. Notations

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; $|\mathcal{S}|$: cardinality of a finite set $\mathcal{S}$; $A \otimes B$: Kronecker's product of matrices A and B; $\mathbf{1}$: the vector with ones in all elements; $\lambda_{\max}(\cdot)$: maximum eigenvalue; $\lambda_{\min}(\cdot)$: minimum eigenvalue.

B. Markov decision problem

We consider the infinite-horizon discounted Markov decision problem (MDP) [3], 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. The state then transitions to a state s' with probability $P(s'|s, a)$, and the transition incurs a reward $r(s, a, s')$, where r is a reward function. For convenience, we consider a deterministic reward function and simply write $r(s_k, a_k, s_{k+1}) =: r_k, k \geq 0$. A deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$, maps a state $s \in \mathcal{S}$ to an action $\pi(s) \in \mathcal{A}$. The objective of the Markov decision problem (MDP) is to find an optimal (deterministic) policy, π^*, such that the cumulative discounted rewards over infinite time horizons are maximized, i.e.,

$$\pi^* := \arg\max_{\pi \in \Theta} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_k \,\middle|\, \pi\right],$$

where $\gamma \in [0, 1)$ is the discount factor, Θ 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 π, and $\mathbb{E}[\cdot|\pi]$ represents an expectation conditioned on the policy π. The Q-function under policy π is defined as

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_k \,\middle|\, s_0 = s, a_0 = a, \pi\right]$$

for all $s \in \mathcal{S}, a \in \mathcal{A}$, and the optimal Q-function is defined as $Q^*(s, a) = Q^{\pi^*}(s, a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. Once Q^* is

known, then an optimal policy can be retrieved by $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a)$. Throughout, we make the following standard assumption.

Assumption 1: The reward function is unit bounded as follows:

$$\max_{(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} |r(s, a, s')| \leq 1.$$

The unit bound imposed on r is just for simplicity of analysis. Under Assumption 1, Q-function is bounded as follows.

Lemma 1: We have $\|Q^*\|_\infty \leq \frac{1}{1-\gamma}$.

Proof: The proof is straightforward from the definition of Q-function as follows: $|Q(s, a)| \leq \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$. ∎

C. Switching system

Let us consider the *switched linear system (SLS)* [4]–[11],

$$x_{k+1} = A_{\sigma_k} x_k, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\}, \quad (4)$$

Where $x_k \in \mathbb{R}^n$ is the state, $\sigma \in \mathcal{M} := 1, 2, \ldots, M$ is called the mode, $\sigma_k \in \mathcal{M}$ is called the switching signal, and $A_\sigma, \sigma \in \mathcal{M}$ are called the subsystem matrices. The switching signal can be either arbitrary or controlled by the user under a certain switching policy. Specifically, a state-feedback switching policy is denoted by $\sigma_k = \sigma(x_k)$. The analysis and control synthesis of SLSs have been actively studied during the last decades [4]–[11]. A more general class of systems is the *switched affine system (SAS)*

$$x_{k+1} = A_{\sigma_k} x_k + b_{\sigma_k}, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\},$$

where $b_{\sigma_k} \in \mathbb{R}^n$ is the additional input vector, which also switches according to σ_k. Due to the additional input b_{σ_k}, its stabilization becomes much more challenging. Lastly, when all elements of the subsystem matrices, $\{A_\sigma, \sigma \in \mathcal{M}\}$, are nonnegative, then SLS is called positive SLS [12]–[15].

D. Definitions

Throughout the paper, we will use the following compact notations:

$$P := \begin{bmatrix} P_1 \\ \vdots \\ P_{|\mathcal{A}|} \end{bmatrix}, \quad R := \begin{bmatrix} R(\cdot, 1) \\ \vdots \\ R(\cdot, |A|) \end{bmatrix}, \quad Q := \begin{bmatrix} Q(\cdot, 1) \\ \vdots \\ Q(\cdot, |\mathcal{A}|) \end{bmatrix},$$

where $P_a = P(\cdot|\cdot, a) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$, $Q(\cdot, a) \in \mathbb{R}^{|\mathcal{S}|}, a \in \mathcal{A}$, and $R \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is an enumerate of $R(s, a) := \mathbb{E}[r_k|s_k = s, a_k = a]$ with an appropriate order compatible with other definitions. Note that $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$, and $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. In this notation, Q-function is encoded as a single vector $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, which enumerates $Q(s, a)$ for all $s \in S$ and $a \in A$ with an appropriate order. In particular, the single value $Q(s, a)$ can be written as $Q(s, a) = (e_a \otimes e_s)^T Q$, where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$ are s-th basis vector (all components are 0 except for the s-th component which is 1) and a-th basis vector, respectively. Therefore, in the above definitions, all entries are ordered compatible with this vector Q.

Algorithm 1 Q-VI

1: Initialize $Q_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ randomly.
2: **for** iteration $k = 0, 1, \ldots$ **do**
3:    Update

$$Q_{k+1}(s,a) = R(s,a) + \underbrace{\gamma \sum_{s' \in S} P(s'|s,a) \max_{a' \in A} Q_k(s',a')}_{=:FQ_k}$$

4: **end for**

For any stochastic policy, $\pi : \mathcal{S} \to \Delta_{|\mathcal{S}|}$, where $\Delta_{|\mathcal{A}|}$ is the set of all probability distributions over $\mathcal{A}$, we define the corresponding action transition matrix as

$$\Pi^\pi := \begin{bmatrix} \pi(1)^T \otimes e_1^T \\ \pi(2)^T \otimes e_2^T \\ \vdots \\ \pi(|S|)^T \otimes e_{|\mathcal{S}|}^T \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|}, \quad (5)$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$. Then, it is well-known that $P\Pi^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ is the transition probability matrix of the state-action pair under policy π. If we consider a deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$, the stochastic policy can be replaced with the corresponding one-hot encoding vector $\vec{\pi}(s) := e_{\pi(s)} \in \Delta_{|\mathcal{A}|}$, where $e_a \in \mathbb{R}^{|\mathcal{A}|}$, and the corresponding action transition matrix is identical to (5) with π replaced with $\vec{\pi}$. For any given $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, denote the greedy policy with respect to Q as $\pi_Q(s) := \arg\max_{a \in \mathcal{A}} Q(s,a) \in \mathcal{A}$. We will use the following shorthand frequently: $\Pi_Q := \Pi^{\pi_Q}$.

E. Q-value iteration (Q-VI)

In this paper, we consider the so-called Q-value iteration (Q-VI) [1] given in Algorithm 1, where F is called Bellman operator. It is well-known that the iterates of Q-VI converge exponentially to Q^* in terms of the infinity norm $\|\cdot\|_\infty$ [1, Lemma 2.5].

Lemma 2: We have the bounds for Q-VI iterates $\|Q_{k+1} - Q^*\|_\infty \leq \gamma \|Q_k - Q^*\|_\infty$.
The proof is given in [1, Lemma 2.5], which is based on the contraction property of Bellman operator. Note that [1, Lemma 2.5] deals with the value iteration for value function instead of Q-function addressed in our work. However, it is equivalent to Q-VI. Next, a direct consequence of Lemma 2 is the convergence of Q-VI

$$\|Q_k - Q^*\|_\infty \leq \gamma^k \|Q_0 - Q^*\|_\infty \quad (6)$$

In what follows, an equivalent switching system model that captures the behavior of Q-VI is introduced, and based on it, we provide a different proof of Lemma 2.

III. SWITCHING SYSTEM MODEL

In this section, we study a discrete-time switching system model of Q-VI and establish its finite-time convergence based on the stability analysis of the switching system.

Using the notation introduced in Section II-D, the update in Algorithm 1 can be rewritten as

$$Q_{k+1} = R + \gamma P\Pi_{Q_k} Q_k, \quad (7)$$

Recall the definitions $\pi_Q(s)$ and Π_Q. Invoking the optimal Bellman equation $(\gamma P\Pi_{Q^*} - I)Q^* + R = 0$, (7) can be further rewritten by

$$(Q_{k+1} - Q^*) = \gamma P\Pi_{Q_k}(Q_k - Q^*) + \gamma P(\Pi_{Q_k} - \Pi_{Q^*})Q^*, \quad (8)$$

which is a SAS (switched affine system). In particular, for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, define

$$A_Q := \gamma P\Pi_Q, \quad b_Q := \gamma P(\Pi_Q - \Pi_{Q^*})Q^*$$

Hence, Q-VI can be concisely represented as the SAS

$$Q_{k+1} - Q^* = A_{Q_k}(Q_k - Q^*) + b_{Q_k}, \quad (9)$$

where A_{Q_k} and b_{Q_k} switch among matrices from $\{\gamma P\Pi^\pi : \pi \in \Theta\}$ and vectors from $\{\gamma P(\Pi^\pi - \Pi^{\pi^*})Q^* : \pi \in \Theta\}$ according to the changes of Q_k. Therefore, the convergence of Q-VI is now reduced to analyzing the stability of the above SAS. Th main obstacle in proving the stability arises from the presence of the affine term. Without it, we can easily establish the exponential stability of the corresponding deterministic switching system, under arbitrary switching policy. Specifically, we have the following result.

Proposition 1: For arbitrary $H_k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, k \geq 0$, the linear switching system $Q_{k+1} - Q^* = A_{H_k}(Q_k - Q^*), Q_0 - Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, is exponentially stable such that $\|Q_{k+1} - Q^*\|_\infty \leq \gamma \|Q_k - Q^*\|_\infty, k \geq 0$ and $\|Q_k - Q^*\|_\infty \leq \gamma^k \|Q_0 - Q^*\|_\infty, k \geq 0$.

The above result follows immediately from the key fact that $\|A_Q\|_\infty \leq \gamma$, which we formally state in the lemma below.

Lemma 3: For any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \|A_Q\|_\infty \leq \gamma$, where the matrix norm $\|A\|_\infty := \max_{1 \leq i \leq m} \sum_{j=1}^n |A_{ij}|$ and A_{ij} is the element of A in i-th row and j-th column.

Proof: Note $\sum_j |[A_Q]_{ij}| = \sum_j |[\gamma P\Pi_Q]_{ij}| = \gamma$, which completes the proof. ∎

However, because of the additional affine term in the original switching system (9), it is not obvious how to directly derive its finite-time convergence. To circumvent the difficulty with the affine term, we will resort to two simpler upper and lower bounds, which are given below.

Proposition 2 (Upper and lower bounds): For all $k \geq 0$, we have

$$A_{Q^*}(Q_k - Q^*) \leq Q_{k+1} - Q^* \leq A_{Q_k}(Q_k - Q^*).$$
Proof: For the lower bound, we have

$$\begin{aligned}&(Q_{k+1} - Q^*)\\=&A_{Q^*}(Q_k - Q^*) + (A_{Q_k} - A_{Q^*})(Q_k - Q^*) + b_{Q_k}\\=&A_{Q^*}(Q_k - Q^*) + \gamma P(\Pi_{Q_k} - \Pi_{Q^*})Q_k\\\geq&A_{Q^*}(Q_k - Q^*)\end{aligned}$$

where the third line is due to $P(\Pi_{Q_k} - \Pi_{Q^*})Q_k \geq P(\Pi_{Q^*} - \Pi_{Q^*})Q_k = 0$. For the upper bound, one gets

$$(Q_{k+1} - Q^*) = A_{Q_k}(Q_k - Q^*) + b_{Q_k}$$

$$\leq A_{Q_k}(Q_k - Q^*),$$

where we used the fact that $b_{Q_k} = (\gamma P \Pi_{Q_k} Q^* - \gamma P \Pi_{Q^*} Q^*) \leq (\gamma P \Pi_{Q^*} Q^* - \gamma P \Pi_{Q^*} Q^*) = 0$ in the first inequality. This completes the proof. ∎

The lower bound is in the form of an LTI system. Specifically, the lower bound corresponds to a special LTI system form known as a positive linear system [21], where all the elements of the system matrix A_{Q^*} are nonnegative. In our analysis, this property will be utilized to derive new behaviors of Q-VI. Similarly, the system matrix A_{Q_k} in the upper bound switches according to the changes in Q_k. Therefore, it can be proven that the upper bound is in the form of positive SLSs [12]–[15], where all the elements of the system matrix for each mode are nonnegative.

A. Finite-time error bound of Q-VI

In this subsection, we provide a proof for the contraction property outlined in Lemma 2. Although this result represents one of the most basic and fundamental outcomes in classical dynamic programming, we offer an alternative proof in this paper. This proof, grounded in the switching system model discussed in earlier sections, is provided below.

Since $A_{Q^*}(Q_k - Q^*) \leq Q_{k+1} - Q^* \leq A_{Q_k}(Q_k - Q^*)$ from Proposition 2, it follows that $(e_a \otimes e_s)^T A_{Q^*}(Q_k - Q^*) \leq (e_a \otimes e_s)^T (Q_k - Q^*) \leq (e_a \otimes e_s)^T A_{Q_k}(Q_k - Q^*)$. If $(e_a \otimes e_s)^T(Q_k - Q^*) \leq 0$, then $|(e_a \otimes e_s)^T(Q_k - Q^*)| \leq |(e_a \otimes e_s)^T A_{Q^*}(Q_k - Q^*)|$, where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$ are the s-th and a-th standard basis vectors, respectively. If $(e_a \otimes e_s)^T(Q_k - Q^*) > 0$, then $|(e_a \otimes e_s)^T(Q_k - Q^*)| \leq |(e_a \otimes e_s)^T A_{Q_k}(Q_k - Q^*)|$. Therefore, one gets

$$\begin{aligned}
&\|Q_{k+1} - Q^*\|_\infty \\
&\leq \max\{\|A_{Q_k}(Q_k - Q^*)\|_\infty, \|A_{Q^*}(Q_k - Q^*)\|_\infty\} \\
&\leq \max\{\gamma \|Q_k - Q^*\|_\infty, \gamma \|Q_k - Q^*\|_\infty\} \\
&= \gamma \|Q_k - Q^*\|_\infty,
\end{aligned}$$

which completes the proof.

IV. CONVERGENCE OF Q-VI ON THE ORTHANT

In the previous section, we revisited the convergence of Q-VI from the perspective of switching system viewpoints. In this section, we study additional geometric behaviors of Q-VI, again using the switching system perspective. Let us suppose that $Q^* \geq Q_0$ holds. Please note that such an initial value can be found by setting

$$Q_0 = -\frac{1}{1-\gamma}\mathbf{1}$$

because

$$-\frac{1}{1-\gamma}\mathbf{1} \leq Q^* \leq \frac{1}{1-\gamma}\mathbf{1}$$

holds according to Lemma 1. Then, using the bound provided in Proposition 2, it can be proved that $Q^* \geq Q_k, \forall k \geq 0$. This is equivalent to saying $0 \geq Q_k - Q^*, \forall k \geq 0$. In other words, if the initial iterate Q_0 falls within the shifted orthant, $Q^* \geq Q_0$, then future iterates will also stay within the same set, i.e., $Q^* \geq Q_k, \forall k \geq 0$.

Proposition 3: Suppose that $Q_0 - Q^* \leq 0$ holds. Then, $Q_k - Q^* \leq 0, \forall k \geq 0$.

Proof: For an induction argument, suppose $Q_k - Q^* \leq 0$ for any $k \geq 0$. Then, by Proposition 2, it follows that $Q_{k+1} - Q^* \leq A_{Q_k}(Q_k - Q^*) \leq 0$ because A_{Q_k} is a nonnegative matrix. Therefore, $Q_{k+1} - Q^* \leq 0$, and the proof is completed by induction. ∎

From the results above, it can be observed that the behavior of $Q_k - Q^*$ is fully dictated by the lower bound, which is a linear function of $Q_k - Q^*$. Therefore, the fundamental theory of linear systems can be applied to analyze the behavior of Q-VI. To facilitate this, the subsequent result reviews the Lyapunov theory for discrete-time linear systems.

Proposition 4: For any $\varepsilon > 0$ such that $\gamma + \varepsilon \in (0, 1)$, there exists the corresponding positive definite $M \succ 0$ such that

$$A_{Q^*}^T M A_{Q^*} = (\gamma + \varepsilon)^2 (M - I),$$

and

$$\lambda_{\min}(M) \geq 1, \quad \lambda_{\max}(M) \leq \frac{|\mathcal{S}||\mathcal{A}|}{1 - \left(\frac{\gamma}{\gamma+\varepsilon}\right)^2}.$$

The proof of Proposition 4 is provided in the Appendix. Note that Proposition 4 can be applied to general LTI systems. However, because the lower bound takes the form of positive linear systems, it can be demonstrated that the Lyapunov matrix M also possesses the additional special property outlined below.

Proposition 5: M is a nonnegative matrix.

Proof: The proof is easily done from the construction in (12). ∎

From the perspectives provided above, we can derive a bound on $Q_k - Q^*$ in terms of the weighted Euclidean norm $\|\cdot\|_M$. This is an alternative to the infinity norm $\|\cdot\|_\infty$, which cannot be derived from classical contraction mapping arguments.

Theorem 1: For any $k \geq 0$,

$$\|Q_{k+1} - Q^*\|_M \leq (\gamma+\varepsilon) \|Q_k - Q^*\|_M - (\gamma+\varepsilon) \|Q_k - Q^*\|_2$$

holds

Proof: We have

$$\begin{aligned}
&(Q_{k+1} - Q^*)^T M (Q_{k+1} - Q^*) \\
&\leq (Q_k - Q^*)^T A_{Q^*}^T M A_{Q_k}(Q^* - Q^*) \\
&= (\gamma + \varepsilon)^2 (Q_k - Q^*)^T M (Q_k - Q^*) \\
&\quad - (\gamma + \varepsilon)^2 (Q_k - Q^*)^T (Q_k - Q^*)
\end{aligned}$$

where the first inequality follows from the fact that M is a nonnegative matrix and $A_{Q^*}(Q_k - Q^*) \leq Q_{k+1} - Q^* \leq 0$ from Proposition 2, and the equality is due to the Lyapunov theorem in Proposition 4. Taking the squared root on both sides yields the desired conclusion. ∎

The result in Theorem 1 provides

$$\|Q_{k+1} - Q^*\|_M \leq (\gamma+\varepsilon) \|Q_k - Q^*\|_M - (\gamma+\varepsilon) \|Q_k - Q^*\|_2$$

for an arbitrarily small $\varepsilon > 0$, which implies

$$\|Q_{k+1} - Q^*\|_M \leq (\gamma + \varepsilon) \|Q_k - Q^*\|_M$$

and

$$\|Q_k - Q^*\|_M \le (\gamma + \varepsilon)^k \|Q_0 - Q^*\|_M$$

This relationship reveals different geometric convergence behaviors: the sublevel sets associated with the infinity norm $\|\cdot\|_\infty$ are squares, whereas those corresponding to the weighted Euclidean norm $\|\cdot\|_M$ are ellipsoids.

Furthermore, our new approach enables us to derive a bound on a linear function of $Q_k - Q^*$. Specifically, for positive LTI systems, one common Lyapunov function is a linear Lyapunov function of the form $V(x) = v^T x$ for a positive vector $v \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}, v \ge 0$. We proceed by providing an explicit form of such a vector v.

Proposition 6: For any $\varepsilon > 0$ such that $\gamma + \varepsilon \in (0, 1)$ and for any positive vector $w \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, define

$$v := \left(\sum_{i=0}^{\infty} \left(\frac{1}{\gamma + \varepsilon} \right)^i A_{Q^*}^i \right)^T w \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$$

Then, it holds true that

$$v > 0, \quad \|w\|_\infty \le \|v\|_\infty \le \frac{\|w\|_1}{1 - \gamma}$$

and

$$v^T A_{Q^*} = (\gamma + \varepsilon)(v^T - w^T). \tag{10}$$

The proof is given in Appendix. Proposition 6 proves that $V(x) = v^T x$ plays the role of a linear Lyapunov function for the positive LTI system. From Proposition 6, one can prove another convergence result of Q-VI.

Theorem 2: For the positive vector v given in Proposition 6, we have

$$(\gamma + \varepsilon) v^T (Q_k - Q^*)$$
$$\le v^T (Q_{k+1} - Q^*) + (\gamma + \varepsilon) w^T (Q_k - Q^*)$$
$$\le (\gamma + \varepsilon) w^T (Q_k - Q^*),$$

for all $k \ge 0$.

Proof: Multiplying both sides of (10) in Proposition 6 by $Q_k - Q^*$ from the right leads to

$$(\gamma + \varepsilon) v^T (Q_k - Q^*) - (\gamma + \varepsilon) w^T (Q_k - Q^*)$$
$$= v^T A_{Q^*} (Q_k - Q^*)$$
$$\le v^T (Q_{k+1} - Q^*)$$
$$\le 0$$

where the first equality comes from (10), the first inequality is due to $A_{Q^*}(Q_k - Q^*) \le Q_{k+1} - Q^*$ in Proposition 2, $v \ge 0$, and the second inequality follows from $Q_{k+1} - Q^* \le 0$ in Proposition 3. This completes the proof. ∎

Proposition 6 tells us that

$$(\gamma + \varepsilon) v^T (Q_k - Q^*) \le v^T (Q_{k+1} - Q^*) \le 0, \quad \forall k \ge 0,$$

which implies

$$(\gamma + \varepsilon)^k v^T (Q_0 - Q^*) \le v^T (Q_k - Q^*) \le 0, \quad \forall k \ge 0. \tag{11}$$

Suppose that $Q_k - Q^* = [v]_k + [v]_k^\perp$, where $[v]_k$ represents the component of $Q_k - Q^*$ in the direction of v, and $[v]_k^\perp$

represents the components of $Q_k - Q^*$ orthogonal to v. Then, according to (11), the component of $Q_k - Q^*$ in the direction of v diminishes exponentially. From the construction of v in Proposition 6, it can be seen that there could be infinitely many such vectors v, depending on the positive vector w. Therefore, $Q_k - Q^*$ becomes trapped in the intersections of the nonpositive orthant and infinitely many half planes. An illustration of a single half plane is provided in Figure 1.
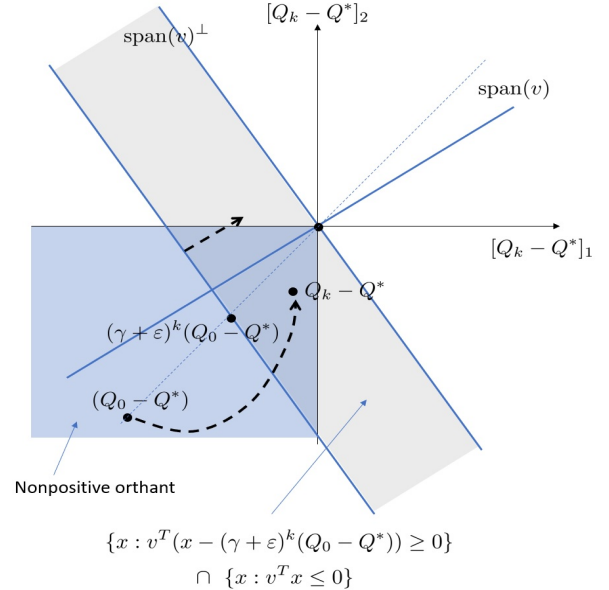

$$\{x : v^T(x - (\gamma + \varepsilon)^k (Q_0 - Q^*)) \ge 0\}$$
$$\cap \ \{x : v^T x \le 0\}$$

Fig. 1. Evolution of $Q_k - Q^*$ and geometric properties from Theorem 2.

V. Conclusion

In this paper, we have presented additional insights on value iteration, approached through the lens of switching system models in the control community. This offers a connection between value iteration and switching system theory and reveals additional geometric behaviors of value iteration. Specifically, we introduced a switching system model of value iteration and, based on it, provided a novel proof for the contraction property of the value iteration. Moreover, our insights led to the proof of new geometric behaviors of value iteration when the initial iterate resides in a particular set (the shifted orthant). We believe that the perspectives proposed here could serve as useful tools applicable in various settings. As such, further development of these methods may present a valuable future direction.

References

[1] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific Belmont, MA, 1996.
[2] D. P. Bertsekas, "Dynamic programming and optimal control 4th edition, volume ii," *Athena Scientific*, 2015.
[3] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
[4] D. Liberzon, *Switching in systems and control*. Springer Science & Business Media, 2003.
[5] D. Lee and J. Hu, "Periodic stabilization of discrete-time switched linear systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3382–3394, 2017.

[6] ——, "Stabilizability of discrete-time controlled switched linear systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3516–3522, 2018.

[7] J. C. Geromel and P. Colaneri, "Stability and stabilization of discrete time switched systems," *Int. J. Control*, vol. 79, no. 7, pp. 719–728, 2006.

[8] T. Hu, L. Ma, and Z. Lin, "Stabilization of switched systems via composite quadratic functions," *IEEE Trans. Autom. Control*, vol. 53, no. 11, pp. 2571–2585, 2008.

[9] H. Lin and P. J. Antsaklis, "Stability and stabilizability of switched linear systems: a survey of recent results," *IEEE Transactions on Automatic control*, vol. 54, no. 2, pp. 308–322, 2009.

[10] W. Zhang, A. Abate, J. Hu, and M. P. Vitus, "Exponential stabilization of discrete-time switched linear systems," *Automatica*, vol. 45, no. 11, pp. 2526–2536, 2009.

[11] D. Lee and N. He, "A unified switching system perspective and convergence analysis of q-learning algorithms," in *34th Conference on Neural Information Processing Systems, NeurIPS 2020*, 2020.

[12] F. Blanchini, P. Colaneri, and M. E. Valcher, "Co-positive lyapunov functions for the stabilization of positive switched systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 12, pp. 3038–3050, 2012.

[13] E. Fornasini and M. E. Valcher, "Linear copositive lyapunov functions for continuous-time positive switched systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 8, pp. 1933–1937, 2010.

[14] J. Zhang, Z. Han, and J. Huang, "Stabilization of discrete-time positive switched systems," *Circuits, Systems, and Signal Processing*, vol. 32, pp. 1129–1145, 2013.

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

[16] H. K. Khalil, "Nonlinear systems," *Upper Saddle River*, 2002.

[17] C.-T. Chen, *Linear System Theory and Design*. Oxford University Press, Inc., 1995.

[18] D. Lee, J. Hu, and N. He, "A discrete-time switching system analysis of q-learning," *arXiv preprint arXiv:2102.08583*, 2021.

[19] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 1998.

[20] X. Guo and B. Hu, "Convex programs and lyapunov functions for reinforcement learning: A unified perspective on the analysis of value-based methods," in *2022 American Control Conference (ACC)*, 2022, pp. 3317–3322.

[21] B. Roszak and E. J. Davison, "Necessary and sufficient conditions for stabilizability of positive lti systems," *Systems & Control Letters*, vol. 58, no. 7, pp. 474–481, 2009.

VI. Appendix

A. Proof of Proposition 4

Proof: For simplicity, denote $A = A_{Q^*}$. Consider matrix M such that

$$M = \sum_{k=0}^{\infty} \left(\frac{1}{\gamma + \varepsilon} \right)^{2k} (A^k)^T A^k. \tag{12}$$

Noting that

$$(\gamma + \varepsilon)^{-2} A^T M A + I$$
$$= \frac{1}{(\gamma + \varepsilon)^2} A^T \left(\sum_{k=0}^{\infty} \left(\frac{1}{\gamma + \varepsilon} \right)^{2k} (A^k)^T A^k \right) A + I$$
$$= M,$$

we have $(\gamma + \varepsilon)^{-2} A^T M A + I = M$, resulting in the desired conclusion. Next, it remains to prove the existence of M by proving its boundedness. Taking the norm on M leads to

$$\|P\|_2 = \left\| I + (\gamma + \varepsilon)^{-2} A^T A + (\gamma + \varepsilon)^{-4} (A^2)^T A^2 + \cdots \right\|_2$$
$$\leq \|I\|_2 + (\gamma + \varepsilon)^{-2} \left\| A^T A \right\|_2$$
$$\quad + (\gamma + \varepsilon)^{-4} \left\| (A^2)^T A^2 \right\|_2 + \cdots$$

$$= \|I\|_2 + (\gamma + \varepsilon)^{-2} \|A\|_2^2 + (\gamma + \varepsilon)^{-4} \|A^2\|_2^2 + \cdots$$
$$= 1 + |\mathcal{S}||\mathcal{A}|(\gamma + \varepsilon)^{-2} \|A\|_\infty^2$$
$$\quad + |S||A|(\gamma + \varepsilon)^{-4} \|A^2\|_\infty^2 + \cdots$$
$$= 1 - |\mathcal{S}||\mathcal{A}| + \frac{|\mathcal{S}||\mathcal{A}|}{1 - \left(\frac{\gamma}{\gamma + \varepsilon} \right)^2}.$$

Finally, we prove the bounds on the maximum and minimum eigenvalues. From the definition (12), $M \succeq I$, and hence $\lambda_{\min}(M) \geq 1$. On the other hand, one gets

$$\lambda_{\max}(M)$$
$$= \lambda_{\max}(I + (\gamma + \varepsilon)^{-2} A^T A$$
$$\quad + (\gamma + \varepsilon)^{-4} (A^2)^T A^2 + \cdots)$$
$$\leq \lambda_{\max}(I) + (\gamma + \varepsilon)^{-2} \lambda_{\max}(A^T A)$$
$$\quad + (\gamma + \varepsilon)^{-4} \lambda_{\max}((A^2)^T A^2) + \cdots$$
$$= \lambda_{\max}(I) + (\gamma + \varepsilon)^{-2} \|A\|_2^2 + (\gamma + \varepsilon)^{-4} \|A^2\|_2^2 + \cdots$$
$$\leq 1 + |\mathcal{S}||\mathcal{A}|(\gamma + \varepsilon)^{-2} \|A\|_\infty^2$$
$$\quad + |\mathcal{S}||\mathcal{A}|(\gamma + \varepsilon)^{-4} \|A^2\|_\infty^2 + \cdots$$
$$\leq \frac{|\mathcal{S}||\mathcal{A}|}{1 - \left(\frac{\gamma}{\gamma + \varepsilon} \right)^2},$$

where $|\mathcal{S}|$ and $|\mathcal{A}|$ denote the cardinality of the sets $\mathcal{S}$ and $\mathcal{A}$, respectively. The proof is completed. ∎

B. Proof of Proposition 6

Proof: We have

$$v^T A_{Q^*} = w^T \left(\sum_{i=0}^{\infty} \left(\frac{1}{\gamma + \varepsilon} \right)^i A_{Q^*}^i \right) A_{Q^*}$$
$$= (\gamma + \varepsilon) w^T \left(\sum_{i=1}^{\infty} \left(\frac{1}{\gamma + \varepsilon} \right)^i A_{Q^*}^i \right)$$
$$= (\gamma + \varepsilon) \left\{ w^T \left(\sum_{i=0}^{\infty} \left(\frac{1}{\gamma + \varepsilon} \right)^i A_{Q^*}^i \right) - w^T \right\}$$
$$= (\gamma + \varepsilon)(v^T - w^T)$$

Moreover,

$$\|v\|_\infty = \left\| w^T \left(\sum_{i=0}^{\infty} A_{Q^*}^i \right) \right\|_\infty$$
$$\leq \left\| w^T \right\|_\infty \left\| \sum_{i=0}^{\infty} A_{Q^*}^i \right\|_\infty$$
$$\leq \|w\|_1 \sum_{i=0}^{\infty} \left\| A_{Q^*}^i \right\|_\infty$$
$$\leq \|w\|_1 \sum_{i=0}^{\infty} \gamma^i$$
$$= \frac{\|w\|_1}{1 - \gamma}$$

and

$$\|v\|_\infty \geq \|w\|_\infty$$

∎