

---

# Characterizing the Exact Behaviors of Temporal Difference Learning Algorithms Using Markov Jump Linear System Theory

---

**Bin Hu, Usman Ahmed Syed**  
Department of Electrical and Computer Engineering  
Coordinated Science Laboratory  
University of Illinois at Urbana-Champaign

## Abstract

In this paper, we provide a unified analysis of temporal difference learning algorithms with linear function approximators by exploiting their connections to Markov jump linear systems (MJLS). We tailor the MJLS theory developed in the control community to characterize the exact behaviors of the first and second order moments of a large family of temporal difference learning algorithms. For both the IID and Markov noise cases, we show that the evolution of some augmented versions of the mean and covariance matrix of the TD estimation error exactly follows the trajectory of a deterministic linear time-invariant (LTI) dynamical system. Applying the well-known LTI system theory, we obtain closed-form expressions for the mean and covariance matrix of the TD estimation error at any time step. We provide a tight matrix spectral radius condition to guarantee the convergence of the covariance matrix of the TD estimation error, and perform a perturbation analysis to characterize the dependence of the TD behaviors on learning rate. For the IID case, we provide an exact formula characterizing how the mean and covariance matrix of the TD estimation error converge to the steady state values at a linear rate. For the Markov case, we use our formulas to explain how the behaviors of TD learning algorithms are affected by learning rate and the underlying Markov chain. For both cases, upper and lower bounds for the mean square TD error are derived. An exact formula for the steady state mean square TD error is also provided.

## 1 Introduction

Reinforcement learning (RL) has shown great promise in solving sequential decision making tasks [5, 48]. One important topic for RL is policy evaluation whose objective is to evaluate the value function of a given policy. A large family of temporal difference (TD) learning methods including standard TD, GTD, TDC, GTD2, DTD, and ATD [47, 50, 49, 38] have been developed to solve the policy evaluation problem. These TD learning algorithms have become important building blocks for RL algorithms. See [17] for a comprehensive survey. Despite the popularity of TD learning, the behaviors of these algorithms have not been fully understood from a theoretical viewpoint. The standard ODE technique [51, 9, 7, 36, 8] can only be used to prove asymptotic convergence. Finite sample bounds are challenging to obtain and typically developed in a case-by-case manner. Recently, there have been intensive research activities focusing on establishing finite sample bounds for TD learning methods with linear function approximations under various assumptions. The IID noise case is covered in [16, 37, 41]. In [6], the analysis is extended for a Markov noise model but an extra projection step in the algorithm is required. Very recently, finite sample bounds for the TD method (without the projection step) under the Markov assumption have been obtained in [45]. The bounds in [45] actually work for any TD learning algorithm that can be modeled by a linear

stochastic approximation scheme. It remains unclear how tight these bounds are (especially for the large learning rate region). To complement the existing analysis results and techniques, we propose a general unified analysis framework for TD learning algorithms by borrowing the Markov jump linear system (MJLS) theory [14] from the controls literature. Our approach is inspired by a recent research trend in applying control theory for analysis of optimization algorithms [39, 30, 31, 29, 21, 52, 15, 46, 28, 32, 22, 3, 40, 26, 4, 18, 43], and extends the jump system perspective for finite sum optimization methods in [31] to TD learning.

Our key insight is that TD learning algorithms with linear function approximations are essentially just Markov jump linear systems. Notice that a MJLS is described by a linear state space model whose state/input matrices are functions of a jump parameter sampled from a finite state Markov chain. Since the behaviors of MJLS have been well established in the controls field [14, 23, 1, 12, 13, 33, 34, 19, 20, 44], we can borrow the analysis tools there to analyze TD learning algorithms in a more unified manner. Our main contributions are summarized as follows.

1. We present a unified Markov jump linear system perspective on a large family of TD learning algorithms including TD, TDC, GTD, GTD2, ATD, and DTD. Specifically, we make the key observation that these methods are just MJLS subject to some prescribed input.
2. By tailoring the existing MJLS theory, we show that the evolution of some augmented versions of the mean and covariance matrix of the estimation error in all above TD learning methods exactly follows the trajectory of a deterministic linear time-invariant (LTI) dynamical system for both the IID and Markov noise cases. As a result, we obtain unified closed-form formulas for the mean and covariance matrix of the TD estimation error at any time step.
3. We provide a tight matrix spectral radius condition to guarantee the convergence of the covariance matrix of the TD estimation error under the general Markov assumption. By using the matrix perturbation theory [42, 35, 2, 24], we perform a perturbation analysis to show the dependence of the behaviors of TD learning on learning rate in a more transparent manner. For the IID case, we provide an exact formula characterizing how the mean and covariance matrix of the TD estimation error converge to the steady state values at a linear rate. For the Markov case, we use our formulas to explain how the behaviors of TD learning algorithms are affected by learning rate and the underlying Markov chain. For both cases, we have shown that the mean square error of TD learning converges linearly to a limit whose exact formula is also provided. In addition, upper and lower bounds for the mean square error of TD learning are simultaneously obtained.

We view our proposed analysis as a complement rather than a replacement for existing analysis techniques. The existing analysis focuses on upper bounds for the TD estimation error. Our closed-form formulas provide both upper and lower bounds for the mean square error of TD learning. Our analysis also characterizes the exact limit of the steady state TD error and related convergence rates.

## 2 Background

### 2.1 Notation

The set of  $m$ -dimensional real vectors is denoted as  $\mathbb{R}^m$ . The Kronecker product of two matrices  $A$  and  $B$  is denoted by  $A \otimes B$ . Notice  $(A \otimes B)^T = A^T \otimes B^T$  and  $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$  when the matrices have compatible dimensions. Let  $\text{vec}$  denote the standard vectorization operation that stacks the columns of a matrix into a vector. We have  $\text{vec}(AXB) = (B^T \otimes A) \text{vec}(X)$ . Let  $\text{sym}$  denote the symmetrization operation, i.e.  $\text{sym}(A) = \frac{A^T + A}{2}$ . Let  $\text{diag}(H_i)$  denote a matrix whose  $(i, i)$ -th block is  $H_i$  and all other blocks are zero. Specifically, given  $H_i$  for  $i = 1, \dots, n$ , we have

$$\text{diag}(H_i) = \begin{bmatrix} H_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & H_n \end{bmatrix}.$$

A square matrix is Schur stable if all its eigenvalues have magnitude strictly less than 1. A square matrix is Hurwitz if all its eigenvalues have strictly negative real parts. The spectral radius of a matrix  $H$  is denoted as  $\sigma(H)$ . The eigenvalue with the largest magnitude of  $H$  is denoted as  $\lambda_{\max}(H)$  and the eigenvalue with the largest real part of  $H$  is denoted as  $\lambda_{\max \text{ real}}(H)$ .

## 2.2 Linear time-invariant systems

A linear time-invariant (LTI) system is typically governed by the following state-space model

$$x^{k+1} = \mathcal{H}x^k + \mathcal{G}u^k, \quad (1)$$

where  $x^k \in \mathbb{R}^{n_x}$ ,  $u^k \in \mathbb{R}^{n_u}$ ,  $\mathcal{H} \in \mathbb{R}^{n_x \times n_x}$ , and  $\mathcal{G} \in \mathbb{R}^{n_x \times n_u}$ . The LTI system theory has been well documented in standard control textbooks [27, 10]. Here we briefly review several useful results.

- **Closed-form formulas for  $x^k$ :** Given an initial condition  $x^0$  and an input sequence  $\{u^k\}$ , the sequence  $\{x^k\}$  can be determined using the following closed-form expression

$$x^k = (\mathcal{H})^k x^0 + \sum_{t=0}^{k-1} (\mathcal{H})^{k-1-t} \mathcal{G}u^t, \quad (2)$$

where  $(\mathcal{H})^k$  stands for the  $k$ -th power of the matrix  $\mathcal{H}$ .

- **Necessary and sufficient stability condition:** When  $\mathcal{H}$  is Schur stable, we know  $(\mathcal{H})^k x^0 \rightarrow 0$  for any arbitrary  $x^0$ . When  $\sigma(\mathcal{H}) \geq 1$ , there always exists  $x^0$  such that  $(\mathcal{H})^k x^0$  does not converge to 0. When  $\sigma(\mathcal{H}) > 1$ , there even exists  $x^0$  such that  $(\mathcal{H})^k x^0 \rightarrow \infty$ . See Section 7.2 in [27] for a detailed discussion. A well-known result in the controls literature is that the LTI system (1) is stable if and only if  $\mathcal{H}$  is Schur stable.
- **Exact limit for  $x^k$ :** If  $\mathcal{H}$  is Schur stable and  $u^k$  converges to a limit  $u^\infty$ , then  $x^k$  will converge to an exact limit. This is formalized as follows.

**Proposition 1.** *Consider the LTI system (1). If  $\sigma(\mathcal{H}) < 1$  and  $\lim_{k \rightarrow \infty} u^k = u^\infty$ , then  $\lim_{k \rightarrow \infty} x^k$  exists and we have  $x^\infty = \lim_{k \rightarrow \infty} x^k = (I - \mathcal{H})^{-1} \mathcal{G}u^\infty$ .*

- **Response for constant input:** If  $u^k = u \forall k$  and  $\sigma(\mathcal{H}) < 1$ , then the closed-form expression for  $x^k$  can be further simplified to give the following tight convergence rate result.

**Proposition 2.** *Suppose  $\sigma(\mathcal{H}) < 1$ , and  $x^k$  is determined by (1). If  $u^k = u \forall k$ , then  $x^k$  converges to a limit point  $x^\infty = \lim_{k \rightarrow \infty} x^k = (I - \mathcal{H})^{-1} \mathcal{G}u$ . And we can compute  $x^k$  as*

$$x^k = x^\infty + (\mathcal{H})^k (x^0 - x^\infty). \quad (3)$$

*In addition,  $\|x^k - x^\infty\| \leq C_0(\sigma(\mathcal{H}) + \varepsilon)^k$  for some  $C_0$  and any arbitrarily small  $\varepsilon > 0$ .*

From the above proposition, we can clearly see that now  $x^k$  is a sum of a constant steady state term  $x^\infty$  and a matrix power term that decays at a linear rate specified by  $\sigma(\mathcal{H})$  (see Section 2.2 in [39] for more explanations). The convergence rate characterized by  $(\sigma(\mathcal{H}) + \varepsilon)$  is tight. More discussions on the tightness of this convergence rate are provided in the supplementary material.

- **Response for exponentially shrinking input:** When  $u^k$  itself converges at a linear rate  $\tilde{\rho}$  and  $\mathcal{H}$  is Schur stable,  $x^k$  will converge to its limit point at a linear rate specified by  $\max\{\sigma(\mathcal{H}) + \varepsilon, \tilde{\rho}\}$ . A formal statement is provided as follows.

**Proposition 3.** *Suppose  $\sigma(\mathcal{H}) < 1$ , and  $x^k$  is determined by (1). If  $u^k$  converges to  $u^\infty$  as  $\|u^k - u^\infty\| \leq C\tilde{\rho}^k$ , then we have  $x^\infty = \lim_{k \rightarrow \infty} x^k = (I - \mathcal{H})^{-1} \mathcal{G}u^\infty$  and  $\|x^k - x^\infty\| \leq C_0(\max\{\sigma(\mathcal{H}) + \varepsilon, \tilde{\rho}\})^k$  for some  $C_0$  and any arbitrarily small  $\varepsilon > 0$ .*

The results in Propositions 1, 2, and 3 are well known in the control community. For completeness, we will include their proofs in the supplementary material.

## 2.3 Markov jump linear systems

Another important class of dynamic systems that have been extensively studied in the controls literature is the so-called Markov jump linear system (MJLS) [14]. Let  $\{z^k\}$  be a Markov chain sampled from a finite state space  $\mathcal{S}$ . A MJLS is governed by the following state-space model:

$$\xi^{k+1} = H(z^k)\xi^k + G(z^k)y^k, \quad (4)$$

where  $H(z^k)$  and  $G(z^k)$  are matrix functions of  $z^k$ . Here,  $\xi^k$  is the state, and  $y^k$  is the input. There is a one-to-one mapping from  $\mathcal{S}$  to the set  $\mathcal{N} := \{1, 2, \dots, n\}$  where  $n = |\mathcal{S}|$ . We can assume  $H(z^k)$  is sampled from a set of matrices  $\{H_1, H_2, \dots, H_n\}$  and  $G(z^k)$  is sampled from  $\{G_1, G_2, \dots, G_n\}$ . We have  $H(z^k) = H_i$  and  $G(z^k) = G_i$  when  $z^k = i$ . The MJLS theory has been well developed in the controls community [14]. We will apply the MJLS theory to analyze TD learning algorithms.

### 3 A general Markov jump system perspective for TD learning

In this section, we provide a general jump system perspective for TD learning with linear function approximations. Notice that many TD learning algorithms including TD, TDC, GTD, GTD2, A-TD, and D-TD can be modeled by the following linear stochastic recursion:

$$\xi^{k+1} = \xi^k + \alpha (A(z^k)\xi^k + b(z^k)), \quad (5)$$

where  $\{z^k\}$  forms a finite state Markov chain and  $b(z^k)$  satisfies  $\lim_{k \rightarrow \infty} \mathbb{E}b(z^k) = 0$ .<sup>1</sup> We have  $A(z^k) = A_i$  and  $b(z^k) = b_i$  when  $z^k = i$ . For simplicity, we mainly focus on analyzing (5). Other models including two time-scale schemes [25, 54] will be discussed in the supplementary material.

Our key observation is that (5) can be rewritten as the following MJLS

$$\xi^{k+1} = (I + \alpha A(z^k))\xi^k + \alpha b(z^k). \quad (6)$$

The above model is a special case of (4) if we set  $H(z^k) = I + \alpha A(z^k)$ ,  $G(z^k) = \alpha b(z^k)$ , and  $y^k = 1 \forall k$ . Consequently, many TD learning algorithms can be analyzed using the MJLS theory.

We will borrow the analysis idea from the standard MJLS theory. Our analysis is built upon the fact that some augmented versions of the mean and the covariance matrix of  $\{\xi^k\}$  for the MJLS model (4) actually follow the dynamics of a deterministic LTI model in the form of (1) [14, Chapter 3]. To see this, we denote the transition probabilities for the Markov chain  $\{z^k\}$  as  $p_{ij} := \mathbb{P}(z^{k+1} = j | z^k = i)$  and specify the transition matrix  $P$  by setting its  $(i, j)$ -th entry to be  $p_{ij}$ . Obviously, we have  $p_{ij} \geq 0$  and  $\sum_{j=1}^n p_{ij} = 1$  for all  $i$ . Next, the indicator function  $\mathbf{1}_{\{z^k=i\}}$  is defined as  $\mathbf{1}_{\{z^k=i\}} = 1$  if  $z^k = i$  and  $\mathbf{1}_{\{z^k=i\}} = 0$  otherwise. Now we define the following key quantities:

$$q_i^k = \mathbb{E}(\xi^k \mathbf{1}_{\{z^k=i\}}), \quad Q_i^k = \mathbb{E}(\xi^k (\xi^k)^\top \mathbf{1}_{\{z^k=i\}}).$$

Suppose  $y^k = 1 \forall k$ . Based on [14, Proposition 3.35],  $q^k$  and  $Q^k$  can be iteratively calculated as

$$q_j^{k+1} = \sum_{i=1}^n p_{ij} (H_i q_i^k + G_i p_i^k), \quad (7)$$

$$Q_j^{k+1} = \sum_{i=1}^n p_{ij} (H_i Q_i^k H_i^\top + 2 \text{sym}(H_i q_i^k G_i^\top) + p_i^k G_i G_i^\top), \quad (8)$$

where  $p_i^k := \mathbb{P}(z^k = i)$ . If we further augment  $q_i^k$  and  $Q_i^k$  as

$$q^k = \begin{bmatrix} q_1^k \\ \vdots \\ q_n^k \end{bmatrix}, \quad Q^k = \begin{bmatrix} Q_1^k & Q_2^k & \dots & Q_n^k \end{bmatrix},$$

then it is straightforward to rewrite (7) (8) as the following LTI system

$$\begin{bmatrix} q^{k+1} \\ \text{vec}(Q^{k+1}) \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \begin{bmatrix} q^k \\ \text{vec}(Q^k) \end{bmatrix} + \begin{bmatrix} u_q^k \\ u_Q^k \end{bmatrix}, \quad (9)$$

where  $\mathcal{H}_{11}$ ,  $\mathcal{H}_{21}$ ,  $\mathcal{H}_{22}$ ,  $u_q^k$ , and  $u_Q^k$  are given by

$$\begin{aligned} \mathcal{H}_{11} &= \begin{bmatrix} p_{11}H_1 & \dots & p_{n1}H_n \\ \vdots & \ddots & \vdots \\ p_{1n}H_1 & \dots & p_{nn}H_n \end{bmatrix}, \quad \mathcal{H}_{22} = \begin{bmatrix} p_{11}H_1 \otimes H_1 & \dots & p_{n1}H_n \otimes H_n \\ \vdots & \ddots & \vdots \\ p_{1n}H_1 \otimes H_1 & \dots & p_{nn}H_n \otimes H_n \end{bmatrix}, \\ \mathcal{H}_{21} &= \begin{bmatrix} p_{11}(H_1 \otimes G_1 + G_1 \otimes H_1) & \dots & p_{n1}(H_n \otimes G_n + G_n \otimes H_n) \\ \vdots & \ddots & \vdots \\ p_{1n}(H_1 \otimes G_1 + G_1 \otimes H_1) & \dots & p_{nn}(H_n \otimes G_n + G_n \otimes H_n) \end{bmatrix}, \\ u_q^k &= \begin{bmatrix} p_{11}G_1 & \dots & p_{n1}G_n \\ \vdots & \ddots & \vdots \\ p_{1n}G_1 & \dots & p_{nn}G_n \end{bmatrix} \begin{bmatrix} p_1^k I \\ \vdots \\ p_n^k I \end{bmatrix}, \quad u_Q^k = \begin{bmatrix} p_{11}G_1 \otimes G_1 & \dots & p_{n1}G_n \otimes G_n \\ \vdots & \ddots & \vdots \\ p_{1n}G_1 \otimes G_1 & \dots & p_{nn}G_n \otimes G_n \end{bmatrix} \begin{bmatrix} p_1^k I \\ \vdots \\ p_n^k I \end{bmatrix}. \end{aligned} \quad (10)$$

<sup>1</sup>This standard assumption is typically related to the projected Bellman equation and can always be enforced by a shifting argument. More explanations are provided in Remark 1.

A detailed derivation for the above result is presented in the supplementary material. A key implication here is that  $q^k$  and  $\text{vec}(Q^k)$  follow the LTI dynamics (9) and can be analyzed using the standard LTI theory reviewed in Section 2.2. Obviously, we have  $\mathbb{E}\xi^k = \sum_{i=1}^n q_i^k$ ,  $\mathbb{E}(\xi^k(\xi^k)^\top) = \sum_{i=1}^n Q_i^k$ , and  $\mathbb{E}\|\xi^k\|^2 = \text{trace}(\sum_{i=1}^n Q_i^k) = (\mathbf{1}_n^\top \otimes \text{vec}(I_{n_\xi})^\top) \text{vec}(Q^k)$ . Hence the mean, covariance, and mean square norm of  $\xi^k$  can all be calculated using closed-form expressions. We will present a detailed analysis for (6) and provide related implications for TD learning in the next two sections.

For illustrative purposes, we explain the jump system perspective for the standard TD method.

**Example 1: TD method.** The standard TD method (or TD(0)) uses the following update rule:

$$\theta^{k+1} = \theta^k - \alpha \phi(s^k) ((\phi(s^k) - \gamma \phi(s^{k+1}))^\top \theta^k - r(s^k)), \quad (11)$$

where  $\{s^k\}$  is the underlying Markov chain,  $\phi$  is the feature vector,  $r$  is the reward,  $\gamma$  is the discounting factor, and  $\theta^k$  is the weight vector to be estimated. Suppose  $\theta^*$  is the vector that solves the projected Bellman equation. We can set  $z^k = [(\phi(s^{k+1}) - \gamma \phi(s^k))^\top]^\top$  and then rewrite the TD update as

$$\theta^{k+1} - \theta^* = (I + \alpha A(z^k)) (\theta^k - \theta^*) + \alpha b(z^k), \quad (12)$$

where  $A(z^k) = \phi(s^k)(\gamma \phi(s^{k+1}) - \phi(s^k))^\top$  and  $b(z^k) = \phi(s^k)(r(s^k) + (\phi(s^k) - \gamma \phi(s^{k+1}))^\top \theta^*)$ . Suppose  $\lim_{k \rightarrow \infty} p_i^k = p_i^\infty$ . Since the projected Bellman equation and the equation  $\sum_{i=1}^n p_i^\infty b_i = 0$  are actually equivalent, we have naturally enforced  $\lim_{k \rightarrow \infty} \mathbb{E}b(z^k) = 0$ . Therefore, the TD update can be modeled as (6) with  $b(z^k)$  satisfying  $\lim_{k \rightarrow \infty} \mathbb{E}b(z^k) = 0$ . See Section 3.1 in [45] for a similar formulation. Now we can apply the MJLS theory and the LTI model (9) to analyze the covariance  $\mathbb{E}((\theta^k - \theta^*)(\theta^k - \theta^*)^\top)$  and the mean square error  $\mathbb{E}\|\theta^k - \theta^*\|^2$ . In this case, we have

$$q^k = \begin{bmatrix} \mathbb{E}((\theta^k - \theta^*)\mathbf{1}_{\{z^k=1\}}) \\ \vdots \\ \mathbb{E}((\theta^k - \theta^*)\mathbf{1}_{\{z^k=n\}}) \end{bmatrix}, \quad \text{vec}(Q^k) = \begin{bmatrix} \text{vec}(\mathbb{E}((\theta^k - \theta^*)(\theta^k - \theta^*)^\top \mathbf{1}_{\{z^k=1\}})) \\ \vdots \\ \text{vec}(\mathbb{E}((\theta^k - \theta^*)(\theta^k - \theta^*)^\top \mathbf{1}_{\{z^k=n\}})) \end{bmatrix}.$$

Then we can easily analyze  $q^k$  and  $Q^k$  by applying the LTI model (9). In general, the covariance matrix  $\mathbb{E}((\theta^k - \theta^*)(\theta^k - \theta^*)^\top)$  and the mean value  $\mathbb{E}(\theta^k - \theta^*)$  do not directly follow an LTI system. However, when working with the augmented covariance matrix  $Q^k$  and the augmented mean value vector  $q^k$ , we do obtain an LTI model in the form of (1). Once the closed-form expression for  $Q^k$  is obtained, the mean square estimation error for the TD update can be immediately calculated as  $\mathbb{E}\|\theta^k - \theta^*\|^2 = \text{trace}(\sum_{i=1}^n Q_i^k) = (\mathbf{1}_n^\top \otimes \text{vec}(I_{n_\theta})^\top) \text{vec}(Q^k)$ .

Here we omit the detailed formulations for other TD learning methods. The key message is that  $\{z^k\}$  can be viewed as a jump parameter and TD learning methods are essentially just MJLS. Notice that all the TD learning algorithms that can be analyzed using the ODE method are in the form of (6). Jump system perspectives for other TD learning algorithms are discussed in the supplementary material.

**Remark 1** (Assumptions). Denote  $\bar{A} = \lim_{k \rightarrow \infty} \mathbb{E}A(z^k) = \sum_{i=1}^n p_i^\infty A_i$ . In this paper, we will assume  $\bar{A}$  is Hurwitz. This assumption is standard and even required by the ODE approach. For the standard TD method,  $\bar{A}$  is Hurwitz when the discount factor is smaller than 1,  $p_i^\infty$  is positive for all  $i$ , and the feature matrix is full column rank [51]. It is worth emphasizing that the assumption  $\lim_{k \rightarrow \infty} \mathbb{E}b(z^k) = 0$  is also general. Suppose  $\sum_{i=1}^n p_i^\infty b_i \neq 0$ . This case can still be handled using a shifting argument since  $\bar{A}$  is Hurwitz. Notice the iteration  $\xi^{k+1} = (I + \alpha A(z^k))\xi^k + \alpha b(z^k)$  can be rewritten as  $\xi^{k+1} - \tilde{\xi} = \xi^k - \tilde{\xi} + \alpha (A(z^k)(\xi^k - \tilde{\xi}) + A(z^k)\tilde{\xi} + b(z^k))$  for any  $\tilde{\xi}$ . Now we denote  $\tilde{b}_i = A_i \tilde{\xi} + b_i$  and the above iteration just becomes  $\xi^{k+1} - \tilde{\xi} = (I + \alpha A(z^k))(\xi^k - \tilde{\xi}) + \alpha \tilde{b}(z^k)$ . When  $\bar{A}$  is Hurwitz (and hence invertible), we can choose  $\tilde{\xi} = -(\sum_{i=1}^n p_i^\infty A_i)^{-1}(\sum_{i=1}^n p_i^\infty b_i)$  such that  $\sum_{i=1}^n p_i^\infty \tilde{b}_i = \sum_{i=1}^n p_i^\infty (A_i \tilde{\xi} + b_i) = 0$ .

**Remark 2** (Generality of (4)). Notice that (4) provides a general jump system model for linear stochastic schemes that may have more complicated forms than (5). However, (4) can not be directly used to cover nonlinear stochastic approximation schemes. See [53, 11] for recent finite sample analysis results on nonlinear stochastic approximation over non-IID data.

## 4 Analysis under the IID assumption

For illustrative purposes, we first present the analysis for (6) under the IID assumption ( $\mathbb{P}(z^k = i) = p_i \forall i$ ). In this case, the analysis is significantly simpler, since  $\{\mathbb{E}\xi^k\}$  and  $\{\mathbb{E}(\xi^k(\xi^k)^\top)\}$  directly form LTI systems with much smaller dimensions. We denote  $\mu^k := \mathbb{E}\xi^k$  and  $\mathbb{Q}^k := \mathbb{E}(\xi^k(\xi^k)^\top)$ . Then the following equations hold for the general jump system model (4)

$$\begin{aligned}\mu^{k+1} &= \sum_{i=1}^n p_i (H_i \mu^k + G_i) = \bar{H} \mu^k + \bar{G}, \\ \text{vec}(\mathbb{Q}^{k+1}) &= \left( \sum_{i=1}^n p_i H_i \otimes H_i \right) \text{vec}(\mathbb{Q}^k) + \left( \sum_{i=1}^n p_i (H_i \otimes G_i + G_i \otimes H_i) \right) \mu^k + \sum_{i=1}^n p_i G_i \otimes G_i.\end{aligned}\tag{13}$$

There are many ways to derive the above formulas. One way is to first show  $q_i^k = p_i \mu^k$  and  $Q_i^k = p_i \mathbb{Q}^k$  in this case and then apply (7) and (8). Another way is to directly modify the proof of Theorem 1 (which is presented in the supplementary material). Now consider the jump system model (6) under the assumption  $\mathbb{E}b(z^k) = \sum_{i=1}^n p_i b_i = 0$ . In this case, we have  $H_i = I + \alpha A_i$ ,  $G_i = \alpha b_i$ , and  $y^k = 1$ . Denote  $\bar{A} := \sum_{i=1}^n p_i A_i$ . We can directly obtain the following result.

**Theorem 1.** *Consider the jump system model (6) with  $H_i = I + \alpha A_i$ ,  $G_i = \alpha b_i$ , and  $y^k = 1$ . Suppose  $\{z^k\}$  is sampled from  $\mathcal{N}$  using an IID distribution  $\mathbb{P}(z^k = i) = p_i$ . In addition, assume  $\sum_{i=1}^n p_i b_i = 0$ . Then  $\mu^k$  and  $\text{vec}(\mathbb{Q}^k)$  are governed by the following LTI system:*

$$\begin{bmatrix} \mu^{k+1} \\ \text{vec}(\mathbb{Q}^{k+1}) \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \begin{bmatrix} \mu^k \\ \text{vec}(\mathbb{Q}^k) \end{bmatrix} + \begin{bmatrix} 0 \\ \alpha^2 \sum_{i=1}^n p_i (b_i \otimes b_i) \end{bmatrix},\tag{14}$$

where  $\mathcal{H}_{11}$ ,  $\mathcal{H}_{21}$  and  $\mathcal{H}_{22}$  are determined as

$$\begin{aligned}\mathcal{H}_{11} &= I + \alpha \bar{A}, \\ \mathcal{H}_{21} &= \alpha^2 \sum_{i=1}^n p_i (A_i \otimes b_i + b_i \otimes A_i), \\ \mathcal{H}_{22} &= I_{n_\xi^2} + \alpha (I \otimes \bar{A} + \bar{A} \otimes I) + \alpha^2 \sum_{i=1}^n p_i (A_i \otimes A_i).\end{aligned}\tag{15}$$

In addition, if  $\sigma(\mathcal{H}_{22}) < 1$ , we have

$$\begin{bmatrix} \mu^k \\ \text{vec}(\mathbb{Q}^k) \end{bmatrix} = \left( \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \right)^k \left( \begin{bmatrix} \mu^0 \\ \text{vec}(\mathbb{Q}^0) \end{bmatrix} - \begin{bmatrix} \mu^\infty \\ \text{vec}(\mathbb{Q}^\infty) \end{bmatrix} \right) + \begin{bmatrix} \mu^\infty \\ \text{vec}(\mathbb{Q}^\infty) \end{bmatrix}\tag{16}$$

where  $\mu^\infty = \lim_{k \rightarrow \infty} \mu^k = 0$ , and  $\text{vec}(\mathbb{Q}^\infty)$  is given as

$$\text{vec}(\mathbb{Q}^\infty) = \lim_{k \rightarrow \infty} \text{vec}(\mathbb{Q}^k) = -\alpha \left( I \otimes \bar{A} + \bar{A} \otimes I + \alpha \sum_{i=1}^n p_i (A_i \otimes A_i) \right)^{-1} \left( \sum_{i=1}^n p_i (b_i \otimes b_i) \right)\tag{17}$$

*Proof.* For completeness, a detailed proof is presented in the supplementary material.  $\square$

Now we discuss the implications of the above theorem for TD learning. For simplicity, we denote

$$\mathcal{H} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix}.$$

**Stability condition for TD learning.** From the LTI theory, the system (14) is stable if and only if  $\mathcal{H}$  is Schur stable. We can apply Proposition 3.6 in [14] to show that  $\mathcal{H}$  is Schur stable if and only if  $\mathcal{H}_{22}$  is Schur stable. Hence, a necessary and sufficient stability condition for the LTI system (14) is that  $\mathcal{H}_{22}$  is Schur stable. Under this condition, the first term on the right side of (16) converges to 0 at a linear rate specified by  $\sigma(\mathcal{H})$ , and the second term on the right side of (16) is a constant

matrix quantifying the steady state covariance. An important question for TD learning is how to choose  $\alpha$  such that  $\sigma(\mathcal{H}_{22}) < 1$  for some given  $\{A_i\}$ ,  $\{b_i\}$ , and  $\{p_i\}$ . We provide some clue to this question by applying an eigenvalue perturbation analysis to the matrix  $\mathcal{H}_{22}$ . We assume  $\alpha$  is small. Then under mild technical condition<sup>2</sup>, we can ignore the quadratic term  $\alpha^2 \sum_{i=1}^n p_i (A_i \otimes A_i)$  in the expression of  $\mathcal{H}_{22}$  and use  $\lambda_{\max}(I_{n_\xi} + \alpha(I \otimes \bar{A} + \bar{A} \otimes I))$  to estimate  $\lambda_{\max}(\mathcal{H}_{22})$ . We have

$$\lambda_{\max}(\mathcal{H}_{22}) = 1 + 2\lambda_{\max \text{ real}}(\bar{A})\alpha + O(\alpha^2). \quad (18)$$

Then we immediately obtain  $\sigma(\mathcal{H}_{22}) \approx 1 + 2\lambda_{\max \text{ real}}(\bar{A})\alpha + O(\alpha^2)$ . Therefore, as long as  $\bar{A}$  is Hurwitz, there exists sufficiently small  $\alpha$  such that  $\sigma(\mathcal{H}_{22}) < 1$ . More details of the perturbation analysis are provided in the supplementary material.

**Exact limit for the mean square error of TD learning.** Obviously,  $\mu^k$  converges to 0 at the rate specified by  $\sigma(I + \alpha\bar{A})$  due to the relation  $\mu^k = (I + \alpha\bar{A})^k \mu^0$ . Applying Proposition 3 and making use of the block structure in  $\mathcal{H}$ , one can show  $\text{vec}(\mathbb{Q}^\infty) = \alpha^2(I_{n_\xi} - \mathcal{H}_{22})^{-1}(\sum_{i=1}^n p_i(b_i \otimes b_i))$ , which leads to the result in (17). A key message here is that the covariance matrix converges linearly to an exact limit under the stability condition  $\sigma(\mathcal{H}_{22}) < 1$ . We can clearly see  $\lim_{k \rightarrow 0} \text{vec}(\mathbb{Q}^k) = O(\alpha)$  and can be controlled by decreasing  $\alpha$ . When  $\alpha$  is large, we need to keep the quadratic term  $\alpha \sum_{i=1}^n p_i (A_i \otimes A_i)$ . Therefore, our theory captures the steady-state behavior of TD learning for both small and large  $\alpha$ , and complement the existing finite sample bounds in literatures. To further compare our results with existing finite sample bounds, we obtain the following result for the mean square error of TD learning.

**Corollary 1.** *Consider the TD update (12) with  $\bar{A}$  being Hurwitz. Suppose  $\sigma(\mathcal{H}_{22}) < 1$  and  $\mathbb{P}(z^k = i) = p_i \forall i$ . Then  $\lim_{k \rightarrow \infty} \mathbb{E}\|\theta^k - \theta^*\|^2$  exists and is determined as  $\delta^\infty := \lim_{k \rightarrow \infty} \mathbb{E}\|\theta^k - \theta^*\|^2 = \text{trace}(\mathbb{Q}^\infty)$  where  $\mathbb{Q}^\infty$  is given by (17). In addition, the following mean square TD error bounds hold for some constant  $C_0$  and any arbitrary small positive  $\varepsilon$ :*

$$\delta^\infty - C_0(\sigma(\mathcal{H}) + \varepsilon)^k \leq \mathbb{E}\|\theta^k - \theta^*\|^2 \leq \delta^\infty + C_0(\sigma(\mathcal{H}) + \varepsilon)^k. \quad (19)$$

Finally, for sufficiently small  $\alpha$ , one has  $\lim_{k \rightarrow \infty} \mathbb{E}\|\theta^k - \theta^*\|^2 = O(\alpha)$ . If  $\lambda_{\max}(I_{n_\xi} + \alpha(I \otimes \bar{A} + \bar{A} \otimes I))$  is a semisimple eigenvalue, then  $\sigma(\mathcal{H}) = \sigma(\mathcal{H}_{11}) = 1 + \text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha$  for small  $\alpha$ .

*Proof.* Recall that we have  $\mathbb{E}\|\theta^k - \theta^*\|^2 = \text{trace}(\mathbb{Q}^k)$ . Taking limits on both sides leads to the expression for  $\delta^\infty$ . Then we can apply Proposition 2 to obtain a linear convergence bound for  $\mathbb{Q}^k$  which eventually leads to (19). Notice  $\bar{A}$  is assumed to be Hurwitz. Therefore, we can apply standard matrix perturbation theory to show  $\delta^\infty = O(\alpha)$  and  $\sigma(\mathcal{H}) = \sigma(\mathcal{H}_{11}) = 1 + \text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha$  for sufficiently small  $\alpha$ .  $\square$

The above corollary gives both upper and lower bounds for the mean square error of TD learning. From the above result, the final TD estimation error is actually exactly on the order of  $O(\alpha)$ . This justifies the tightness of the existing upper bounds for the final TD error up to a constant factor. From the above corollary, we can also see that one can obtain a faster convergence rate at the price of getting a bigger steady state error. This is consistent with the finite sample bound in the literature [6, 45]. Since  $\mathcal{H}_{21} = O(\alpha^2)$ , it is possible to tighten the rate as  $\sigma(\mathcal{H}_{22}) \approx 1 + 2\text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha$  by allowing some extra error on the order of  $O(\alpha)$ . We omit the details for such modifications.

## 5 Analysis under the Markov assumption

Now we analyze the behaviors of TD learning under the general assumption that  $\{z^k\}$  is a Markov chain. Recall that the augmented mean vector  $q^k$  and the augmented covariance matrix  $Q^k$  have been defined in Section 3. We can directly modify (9) to obtain the following result.

**Theorem 2.** *Consider the jump system model (6) with  $H_i = I + \alpha A_i$ ,  $G_i = \alpha b_i$ , and  $y^k = 1$ . Suppose  $\{z^k\}$  is a Markov chain sampled from  $\mathcal{N}$  using the transition matrix  $P$ . In addition, define  $p_i^k = \mathbb{P}(z^k = i)$  and set the augmented vector  $p^k := [p_1^k \ p_2^k \ \dots \ p_n^k]^\top$ . Clearly  $p^k = (P^\top)^k p^0$ . Further denote the augmented vectors as  $b := [b_1^\top \ b_2^\top \ \dots \ b_n^\top]^\top$ ,*

<sup>2</sup>One such condition is that  $\lambda_{\max}(I_{n_\xi} + \alpha(I \otimes \bar{A} + \bar{A} \otimes I))$  is a semisimple eigenvalue.

$\hat{B} = [(b_1 \otimes b_1)^\top \dots (b_n \otimes b_n)^\top]^\top$ , and set  $S(b_i, A_i) := (b_i \otimes (I + \alpha A_i) + (I + \alpha A_i) \otimes b_i)$ . Then  $q^k$  and  $\text{vec}(Q^k)$  are governed by the following LTI model:

$$\begin{bmatrix} q^{k+1} \\ \text{vec}(Q^{k+1}) \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \begin{bmatrix} q^k \\ \text{vec}(Q^k) \end{bmatrix} + \begin{bmatrix} \alpha((P^\top \text{diag}(p_i^k)) \otimes I_{n_\xi})b \\ \alpha^2((P^\top \text{diag}(p_i^k)) \otimes I_{n_\xi^2})\hat{B} \end{bmatrix}, \quad (20)$$

where  $\mathcal{H}_{11}$ ,  $\mathcal{H}_{21}$  and  $\mathcal{H}_{22}$  are given by

$$\begin{aligned} \mathcal{H}_{11} &= (P^\top \otimes I_{n_\xi}) \text{diag}(I_{n_\xi} + \alpha A_i), \\ \mathcal{H}_{21} &= \alpha \begin{bmatrix} p_{11}S(b_1, A_1) & \dots & p_{n1}S(b_n, A_n) \\ \vdots & \ddots & \vdots \\ p_{1n}S(b_1, A_1) & \dots & p_{nn}S(b_n, A_n) \end{bmatrix}, \\ \mathcal{H}_{22} &= (P^\top \otimes I_{n_\xi^2}) \text{diag}((I_{n_\xi} + \alpha A_i) \otimes (I_{n_\xi} + \alpha A_i)). \end{aligned} \quad (21)$$

In addition, the following closed-form solution holds for any  $k$

$$\begin{aligned} q^k &= (\mathcal{H}_{11})^k q^0 + \alpha \sum_{t=0}^{k-1} (\mathcal{H}_{11})^{k-1-t} ((P^\top \text{diag}(p_i^t)) \otimes I_{n_\xi}) b, \\ \text{vec}(Q^k) &= (\mathcal{H}_{22})^k \text{vec}(Q^0) + \sum_{t=0}^{k-1} (\mathcal{H}_{22})^{k-1-t} \left( \mathcal{H}_{21} q^t + \alpha^2 ((P^\top \text{diag}(p_i^t)) \otimes I_{n_\xi^2}) \hat{B} \right), \end{aligned} \quad (22)$$

where  $\mathcal{H}_{11}$ ,  $\mathcal{H}_{21}$  and  $\mathcal{H}_{22}$  are determined by (21).

*Proof.* A detailed proof is presented in the supplementary material. We present a proof sketch here. Notice (20) is a direct consequence of (7) and (8) (which are special cases of Proposition 3.35 in [14]). Specifically, it is straightforward to verify the following equations using the Markov assumption

$$q_j^{k+1} = \sum_{i=1}^n p_{ij} ((I + \alpha A_i) q_i^k + \alpha p_i^k b_i), \quad (23)$$

$$Q_j^{k+1} = \sum_{i=1}^n p_{ij} ((I + \alpha A_i) Q_i^k (I + \alpha A_i)^\top + 2\alpha \text{sym}((I + \alpha A_i) q_i^k b_i^\top) + \alpha^2 p_i^k b_i b_i^\top). \quad (24)$$

Then we can apply the basic property of the vectorization operation to obtain (20). Applying (2) to iterate (20) directly leads to (22).  $\square$

Therefore, the evolutions of  $q^k$  and  $Q^k$  can be fully understood via the well-established LTI system theory. Now we discuss the implications of Theorem 2 for TD learning.

**Stability condition for TD learning.** Similar to the IID case, the necessary and sufficient stability condition is  $\sigma(\mathcal{H}_{22}) < 1$ . Now  $\mathcal{H}_{22}$  becomes a much larger matrix depending on the transition matrix  $P$ . An important question is how to choose  $\alpha$  such that  $\sigma(\mathcal{H}_{22}) < 1$  for some given  $\{A_i\}$ ,  $\{b_i\}$ ,  $P$ , and  $\{p^0\}$ . Again, we perform an eigenvalue perturbation analysis for the matrix  $\mathcal{H}_{22}$ . This case is quite subtle due to the fact that we are no longer perturbing an identity matrix. We are perturbing the matrix  $(P^\top \otimes I_{n_\xi^2})$  and the eigenvalues here are not simple. Under the ergodicity assumption, the largest eigenvalue for  $(P^\top \otimes I_{n_\xi^2})$  (which is 1) is semisimple. Hence we can directly apply the results in Section II of [35] or Theorem 2.1 in [42] to show

$$\lambda_{\max}(\mathcal{H}_{22}) = 1 + 2\lambda_{\max \text{ real}}(\bar{A})\alpha + o(\alpha), \quad (25)$$

where  $\bar{A} = \sum_{i=1}^n p_i^\infty A_i$  and  $p^\infty$  is the unique stationary distribution of the Markov chain under the ergodicity assumption. Then we still have  $\sigma(\mathcal{H}_{22}) \approx 1 + 2\text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha + o(\alpha)$ . Therefore, as long as  $\bar{A}$  is Hurwitz, there exists sufficiently small  $\alpha$  such that  $\sigma(\mathcal{H}_{22}) < 1$ . This is consistent with Assumption 3 in [45]. To understand the details of our perturbation argument, we refer the readers to the remark placed after Theorem 2.1 in [42]. Notice we have

$$\mathcal{H}_{22} = P^\top \otimes I_{n_\xi^2} + \alpha(P^\top \otimes I_{n_\xi^2}) \text{diag}(A_i \otimes I + I \otimes A_i) + O(\alpha^2).$$

The largest eigenvalue of  $P^\top \otimes I_{n_\xi^2}$  is semisimple due to the ergodicity assumption. Then the perturbation result directly follows as a consequence of Theorem 2.1 in [42]. More explanations are also provided in the supplementary material.



**Exact limit for the mean square TD error and related convergence rate.** Assume the Markov chain  $\{z^k\}$  is aperiodic and irreducible. Then we have  $p^t \rightarrow p^\infty$  at some linear rate where  $p^\infty$  is the stationary distribution. In this case, we can apply Proposition 3 to show that the mean square error of TD learning converges linearly to an exact limit.

**Corollary 2.** Consider the TD update (12) with  $\bar{A}$  being Hurwitz. Let  $\{z^k\}$  be a Markov chain sampled from  $\mathcal{N}$  using the transition matrix  $P$ . Suppose  $\sigma(\mathcal{H}_{22}) < 1$ . We set  $N = nn_\xi^2$ . If we assume  $p^k \rightarrow p^\infty$  where  $p^\infty$  is the stationary distribution for  $\{z^k\}$ , then we have

$$\begin{aligned} q^\infty &= \lim_{k \rightarrow \infty} q^k = \alpha(I - \mathcal{H}_{11})^{-1}((P^\top \text{diag}(p_i^\infty)) \otimes I_{n_\xi})b, \\ \text{vec}(Q^\infty) &= \lim_{k \rightarrow \infty} \text{vec}(Q^k) = \alpha^2(I_N - \mathcal{H}_{22})^{-1} \left( \alpha^{-2} \mathcal{H}_{21} q^\infty + ((P^\top \text{diag}(p_i^\infty)) \otimes I_{n_\xi}) \hat{B} \right), \quad (26) \\ \delta^\infty &= \lim_{k \rightarrow \infty} \mathbb{E} \|\theta^k - \theta^*\|^2 = (\mathbf{1}_n^\top \otimes \text{vec}(I_{n_\theta})^\top) \text{vec}(Q^\infty). \end{aligned}$$

If we further assume the geometric ergodicity, i.e.  $\|p^k - p^\infty\| \leq C\tilde{\rho}^k$ , then we have

$$\delta^\infty - C_0 \max\{\sigma(\mathcal{H}) + \varepsilon, \tilde{\rho}\}^k \leq \mathbb{E} \|\theta^k - \theta^*\|^2 \leq \delta^\infty + C_0 \max\{\sigma(\mathcal{H}) + \varepsilon, \tilde{\rho}\}^k, \quad (27)$$

where  $C_0$  is some constant and  $\varepsilon$  is an arbitrary small positive number. For sufficiently small  $\alpha$ , we have  $\delta^\infty = O(\alpha)$ . If  $\lambda_{\max}(P^\top \otimes I_{n_\xi^2} + \alpha(P^\top \otimes I_{n_\xi^2}) \text{diag}(A_i \otimes I + I \otimes A_i))$  is a semisimple eigenvalue, then we further have  $\sigma(\mathcal{H}) = 1 + \text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha + o(\alpha)$  for sufficiently small  $\alpha$ .

*Proof.* Notice  $\mathbb{E} \|\theta^k - \theta^*\|^2 = (\mathbf{1}_n^\top \otimes \text{vec}(I_{n_\xi})^\top) \text{vec}(Q^k)$ . We can directly apply Theorem 2, Proposition 1, and Proposition 3 to prove (26) and (27). When  $\alpha$  is small, we can apply the Laurent series trick in [2, 24] to show that  $\lim_{k \rightarrow 0} \text{vec}(Q^k) = O(\alpha)$  and  $\delta^\infty = O(\alpha)$ . The difficulty here is that  $I_N - P^\top \otimes I_{n_\xi^2}$  is a singular matrix and hence  $(I_N - \mathcal{H}_{22})^{-1}$  does not have a Taylor series around  $\alpha = 0$ . Therefore, we need to apply some advanced matrix inverse perturbation result to perform a Laurent expansion of  $(I_N - \mathcal{H}_{22})^{-1}$ . By using the ergodicity assumption and the matrix inverse perturbation theory in [2, 24], we can obtain the Laurent expansion of  $(I_N - \mathcal{H}_{22})^{-1}$  and show  $\lim_{k \rightarrow 0} \text{vec}(Q^k) = O(\alpha)$ . Consequently, we have  $\delta^\infty = O(\alpha)$ . By applying Theorem 2.1 in [42], we can show  $\sigma(\mathcal{H}) = 1 + \text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha + o(\alpha)$ .  $\square$

Due to the assumption  $\sum_{i=1}^n p_i^\infty b_i = 0$ , we have  $\lim_{k \rightarrow \infty} q^k \neq 0$  in general but  $\mu^\infty = 0$ . Again, we have obtained both upper and lower bounds for the mean square TD error. Our result states that under mild technical assumptions, the final TD error is actually exactly on the order of  $O(\alpha)$ . This justifies the tightness of the existing upper bounds for the final TD error [6, 45] up to a constant factor. From the above corollary, we can also see the trade-off between the convergence rate and the steady state error. Clearly, the convergence rate in (27) also depends on the initial distribution  $p^0$  and the mixing rate of the underlying Markov jump parameter  $\{z^k\}$  (which is denoted as  $\tilde{\rho}$ ). If the initial distribution is the stationary distribution, i.e.  $p^0 = p^\infty$ , the input to the LTI dynamical system (20) is just a constant for all  $k$  and then we will be able to obtain an exact formula similar to (16). However, for a general initial distribution  $p^0$ , the mixing rate  $\tilde{\rho}$  matters more and may affect the overall convergence rate. One resultant guideline for algorithm design is that increasing  $\alpha$  may not increase the convergence rate when the mixing rate  $\tilde{\rho}$  dominates the convergence process. When  $\alpha$  becomes smaller and smaller, eventually  $\sigma(\mathcal{H})$  is going to become the dominating term and the mixing rate does not affect the convergence rate any more. Similar to the IID case, for sufficiently small  $\alpha$ , it seems possible to obtain alternative upper bounds in the form of  $\mathbb{E} \|\theta^k - \theta^*\|^2 \leq \delta^\infty + O(\alpha) + C_0(\sigma(\mathcal{H}_{22}) + \varepsilon)^k$  where  $\sigma(\mathcal{H}_{22}) \approx 1 + 2 \text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha$ . Such modifications are not pursued in this paper.

**Algorithm design.** Here we make a remark on how our proposed MJLS framework can be further extended to provide clues for designing fast TD learning. When  $\alpha$  (or even other hyperparameters including momentum term) is changing with time, we can still obtain expressions of  $\text{vec}(Q^k)$  and  $q^k$  in an iterative form. However, both  $\mathcal{H}$  and  $\mathcal{G}$  depend on  $k$  now. Then given a fixed time budget  $T$ , in theory it is possible to minimize the mean square estimation error at  $T$  subject to some optimization constraints in the form of a time-varying iteration  $\begin{bmatrix} q^{k+1} \\ \text{vec}(Q^{k+1}) \end{bmatrix} = \mathcal{H}(k) \begin{bmatrix} q^k \\ \text{vec}(Q^k) \end{bmatrix} + \mathcal{G}(k)u^k$ . One may use this control-oriented optimization formulation to gain some theoretical insights on how to choose hyperparameters adaptively for fast TD learning. Clearly, solving such an optimization problem requires knowing the underlying Markov model. However, this type of theoretical study may lead to new hyperparameter tuning heuristics that do not require the model information.

## Acknowledgments

Bin Hu acknowledges helpful discussions with Rayadurgam Srikant.

## References

- [1] H. Abou-Kandil, G. Freiling, and G. Jank. On the solution of discrete-time Markovian jump linear quadratic control problems. *Automatica*, 31(5):765–768, 1995.
- [2] K. Avrachenkov and J. Lasserre. Analytic perturbation of generalized inverses. *Linear Algebra and its Applications*, 438(4):1793–1813, 2013.
- [3] N. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar. Robust accelerated gradient method. *arXiv preprint arXiv:1805.10579*, 2018.
- [4] N. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. *arXiv preprint arXiv:1901.08022*, 2019.
- [5] D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming*, volume 5. Athena Scientific Belmont, 1996.
- [6] J. Bhandari, D. Russo, and R. Singal. A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*, 2018.
- [7] S. Bhatnagar, H. Prasad, and L. Prashanth. *Stochastic recursive algorithms for optimization: simultaneous perturbation methods*, volume 434. Springer, 2012.
- [8] V. Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.
- [9] V. Borkar and S. Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.
- [10] C. Chen. *Linear system theory and design*. Oxford University Press, Inc., 1998.
- [11] Z. Chen, S. Zhang, T. Doan, S. Maguluri, and J. Clarke. Performance of Q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*, 2019.
- [12] H. Chizeck, A. Willsky, and D. Castanon. Discrete-time Markovian-jump linear quadratic optimal control. *International Journal of Control*, 43(1):213–231, 1986.
- [13] O. Costa and M. Fragoso. Stability results for discrete-time linear systems with Markovian jumping parameters. *Journal of Mathematical Analysis and Applications*, 179(1):154–178, 1993.
- [14] O. Costa, M. Fragoso, and R. Marques. *Discrete-time Markov jump linear systems*. Springer Science & Business Media, 2006.
- [15] S. Cyrus, B. Hu, B. Van Scoy, and L. Lessard. A robust accelerated optimization algorithm for strongly convex functions. In *2018 Annual American Control Conference (ACC)*, pages 1376–1381, 2018.
- [16] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor. Finite sample analyses for td (0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [17] C. Dann, G. Neumann, and J. Peters. Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research*, 15(1):809–883, 2014.
- [18] N. Dhingra, S. Khong, and M. Jovanovic. The proximal augmented lagrangian method for nonsmooth composite optimization. *IEEE Transactions on Automatic Control*, 2018.
- [19] L. El Ghaoui and M. Rami. Robust state-feedback stabilization of jump linear systems via LMIs. *International Journal of Robust and Nonlinear Control*, 6(9-10):1015–1022, 1996.
- [20] Y. Fang and K. Loparo. Stochastic stability of jump linear systems. *IEEE Transactions on Automatic Control*, 47(7):1204–1208, 2002.
- [21] M. Fazlyab, A. Ribeiro, M. Morari, and V. Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *arXiv preprint arXiv:1705.03615*, 2017.
- [22] M. Fazlyab, A. Ribeiro, M. Morari, and V. Preciado. A dynamical systems perspective to convergence rate analysis of proximal algorithms. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 354–360, 2017.

- [23] X. Feng, K. Loparo, Y. Ji, and H. Chizeck. Stochastic stability properties of jump linear systems. *IEEE transactions on Automatic Control*, 37(1):38–53, 1992.
- [24] P. Gonzalez-Rodriguez, M. Moscoso, and M. Kindelan. Laurent expansion of the inverse of perturbed, singular matrices. *Journal of Computational Physics*, 299:307–319, 2015.
- [25] H. Gupta, R. Srikant, and L. Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019.
- [26] S. Han. Systematic design of decentralized algorithms for consensus optimization. *arXiv preprint arXiv:1903.01023*, 2019.
- [27] J. Hespanha. *Linear systems theory*. Princeton university press, 2009.
- [28] B. Hu and L. Lessard. Control interpretations for first-order optimization methods. In *2017 American Control Conference (ACC)*, pages 3114–3119, 2017.
- [29] B. Hu and L. Lessard. Dissipativity theory for Nesterov’s accelerated method. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.
- [30] B. Hu, P. Seiler, and L. Lessard. Analysis of approximate stochastic gradient using quadratic constraints and sequential semidefinite programs. *arXiv preprint arXiv:1711.00987*, 2017.
- [31] 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*, pages 1157–1189, 2017.
- [32] 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*, pages 2043–2052, 2018.
- [33] Y. Ji and H. Chizeck. Controllability, observability and discrete-time Markovian jump linear quadratic control. *International Journal of Control*, 48(2):481–498, 1988.
- [34] Y. Ji and H. Chizeck. Jump linear quadratic Gaussian control: Steady-state solution and testable conditions. *Control Theory Adv. Technol.*, 6:289–319, 1990.
- [35] T. Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.
- [36] H. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.
- [37] C. Lakshminarayanan and C. Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.
- [38] D. Lee and N. He. Target-based temporal difference learning. *arXiv preprint arXiv:1904.10945*, 2019.
- [39] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.
- [40] L. Lessard and P. Seiler. Direct synthesis of iterative algorithms with bounds on achievable worst-case convergence rate. *arXiv preprint arXiv:1904.09046*, 2019.
- [41] B. Liu, J. Liu, M. Ghavamzadeh, S. Mahadevan, and M. Petrik. Finite-sample analysis of proximal gradient td algorithms. In *UAI*, pages 504–513, 2015.
- [42] J. Moro, J. Burke, and M. Overton. On the Lidskii–Vishik–Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure. *SIAM Journal on Matrix Analysis and Applications*, 18(4):793–817, 1997.
- [43] Z. Nelson and E. Mallada. An integral quadratic constraint framework for real-time steady-state optimization of linear time-invariant systems. In *2018 Annual American Control Conference (ACC)*, pages 597–603, 2018.
- [44] P. Seiler and R. Sengupta. A bounded real lemma for jump systems. *IEEE Transactions on Automatic Control*, 48(9):1651–1654, 2003.
- [45] R. Srikant and L. Ying. Finite-time error bounds for linear stochastic approximation and TD learning. *arXiv preprint arXiv:1902.00923*, 2019.

- [46] A. Sundararajan, B. Hu, and L. Lessard. Robust convergence analysis of distributed optimization algorithms. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1206–1212, 2017.
- [47] R. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.
- [48] R. Sutton and A. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [49] R. Sutton, H. 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*, pages 993–1000, 2009.
- [50] R. Sutton, C. Szepesvári, and H. Maei. A convergent  $O(n)$  algorithm for off-policy temporal-difference learning with linear function approximation. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pages 1609–1616, 2008.
- [51] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.
- [52] 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*, 2(1):49–54, 2017.
- [53] G. Wang, B. Li, and G. Giannakis. A multistep Lyapunov approach for finite-time analysis of biased stochastic approximation. *arXiv preprint arXiv:1909.04299*, 2019.
- [54] 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.

# Supplementary Material

## A More facts about LTI systems

### A.1 Tightness of the spectral radius stability condition

Technically speaking, the condition  $\sigma(\mathcal{H}) < 1$  is necessary and sufficient for the asymptotic stability of the LTI system (1). Since exponential stability and asymptotic stability are equivalent notions of stability for LTI systems, the condition  $\sigma(\mathcal{H}) < 1$  is also necessary and sufficient for the exponential stability of (1). See Theorem 8.3 and Theorem 8.4 in [27] for formal statements of these facts. We will give a more intuitive explanation here. Specifically, we look at the behaviors of the matrix power term  $(\mathcal{H})^k x^0$ . This is the homogeneous state response of (1). There are three possible behaviors for this term.

1. When  $\mathcal{H}$  is Schur stable (or equivalently  $\sigma(\mathcal{H}) < 1$ ), the term  $(\mathcal{H})^k$  converges to a zero matrix and  $(\mathcal{H})^k x^0 \rightarrow 0$  for any arbitrary  $x^0$ . The convergence rate is linear and is completely specified by the spectral radius  $\sigma(\mathcal{H})$ .
2. When  $\sigma(\mathcal{H}) \leq 1$  and all the Jordan blocks corresponding to eigenvalues with magnitude equal to 1 are  $1 \times 1$ ,  $(\mathcal{H})^k$  remains bounded for any  $k$ . This is the so-called marginal stability case where  $(\mathcal{H})^k x^0$  remains bounded but may not converge to 0.
3. For all other cases,  $(\mathcal{H})^k$  is unbounded and there exists  $x^0$  such that  $(\mathcal{H})^k x^0 \rightarrow \infty$ .

See Section 7.2 in [27] for a detailed explanation of the above fact. Consequently, we can only guarantee  $(\mathcal{H})^k x^0$  to converge for all  $x^0$  when  $\sigma(\mathcal{H}) < 1$ . Therefore, the condition  $\sigma(\mathcal{H}) < 1$  is a tight condition for the stability of the LTI system (1). As mentioned above, when  $\sigma(\mathcal{H}) < 1$ ,  $(\mathcal{H})^k x^0$  converges at a linear rate completely determined by  $\sigma(\mathcal{H})$ . Technically speaking, the convergence rate is either equal to  $\sigma(\mathcal{H}) + \varepsilon$  for some arbitrary small positive  $\varepsilon$  or just equal to  $\sigma(\mathcal{H})$  itself. Now we provide a detailed discussion on this convergence rate.

### A.2 Convergence rate of the matrix power

Notice we have  $(\mathcal{H})^k x^0 = \rho^k (\rho^{-1}\mathcal{H})^k x^0$ . As long as  $(\rho^{-1}\mathcal{H})^k x^0$  stays bounded for any  $x^0$ , the term  $(\mathcal{H})^k x^0$  will converge at the linear rate  $\rho$ .

From Section 7.2 in [27], we can almost directly see how to determine the convergence rate of  $(\mathcal{H})^k x^0$ .

1. When  $\sigma(\mathcal{H}) < 1$  and all the Jordan blocks corresponding to eigenvalues with magnitude equal to  $\sigma(\mathcal{H})$  are  $1 \times 1$ , we can choose  $\rho = \sigma(\mathcal{H})$  and show  $\sigma(\rho^{-1}\mathcal{H}) \leq 1$  and all the Jordan blocks corresponding to eigenvalues (of  $\rho^{-1}\sigma(\mathcal{H})$ ) with magnitude equal to 1 are  $1 \times 1$ . Then  $(\rho^{-1}\mathcal{H})^k x^0$  remains bounded for all  $x^0$  and hence  $(\mathcal{H})^k x^0 = \rho^k (\rho^{-1}\mathcal{H})^k x^0$  converges at a linear rate  $\rho = \sigma(\mathcal{H})$ .
2. When  $\sigma(\mathcal{H}) < 1$  and some of the Jordan blocks corresponding to eigenvalues with magnitude equal to  $\sigma(\mathcal{H})$  are not  $1 \times 1$ , we need to choose  $\rho = \sigma(\mathcal{H}) + \varepsilon$  for some arbitrary small  $\varepsilon > 0$ . Then  $\sigma(\rho^{-1}\mathcal{H}) < 1$  and  $(\rho^{-1}\mathcal{H})^k x^0$  converges to 0 for all  $x^0$ . Consequently,  $(\mathcal{H})^k x^0 = \rho^k (\rho^{-1}\mathcal{H})^k x^0$  converges at a linear rate  $\rho = \sigma(\mathcal{H}) + \varepsilon$ .

In this paper, for simplicity we do not want to further look at the Jordan decomposition structure of  $\mathcal{H}$  and hence we always set the rate as  $\rho = \sigma(\mathcal{H}) + \varepsilon$ . One can also use the relationship between spectral radius and other matrix norms to obtain the above convergence rate. See Section 2.2 in [39] for a detailed discussion. The arbitrarily small number  $\varepsilon$  also appears in the argument there.

## B More discussions about Markov jump linear systems

First, we verify that (7) and (8) are equivalent to the LTI model (9).

Rewriting (7) as an LTI model is quite trivial. Rewriting (8) as an LTI system requires applying the vectorization operation to obtain the following formula,

$$\begin{aligned} \begin{bmatrix} \text{vec}(Q_1^{k+1}) \\ \vdots \\ \text{vec}(Q_n^{k+1}) \end{bmatrix} &= \begin{bmatrix} p_{11}H_1 \otimes H_1 & \dots & p_{n1}H_n \otimes H_n \\ \vdots & \ddots & \vdots \\ p_{1n}H_1 \otimes H_1 & \dots & p_{nn}H_n \otimes H_n \end{bmatrix} \begin{bmatrix} \text{vec}(Q_1^k) \\ \vdots \\ \text{vec}(Q_n^k) \end{bmatrix} + \begin{bmatrix} p_{11}p_1^k I_{n_\xi^2} & \dots & p_{n1}p_n^k I_{n_\xi^2} \\ \vdots & \ddots & \vdots \\ p_{1n}p_1^k I_{n_\xi^2} & \dots & p_{nn}p_n^k I_{n_\xi^2} \end{bmatrix} \begin{bmatrix} G_1 \otimes G_1 \\ \vdots \\ G_n \otimes G_n \end{bmatrix} \\ &+ \begin{bmatrix} p_{11}(G_1 \otimes H_1 + H_1 \otimes G_1) & \dots & p_{n1}(G_n \otimes H_n + H_n \otimes G_n) \\ \vdots & \ddots & \vdots \\ p_{1n}(G_1 \otimes H_1 + H_1 \otimes G_1) & \dots & p_{nn}(G_n \otimes H_n + H_n \otimes G_n) \end{bmatrix} \begin{bmatrix} q_1^k \\ \vdots \\ q_n^k \end{bmatrix}. \end{aligned} \quad (\text{B.1})$$

Then we can augment the update rules for  $q^k$  and  $\text{vec}(Q^k)$  to obtain the desired LTI model for  $(q^k, \text{vec}(Q^k))$ .

Now we briefly review how to analyze  $q^k$  and  $Q^k$  using the LTI model (9). We denote  $\mathcal{H} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix}$ .

First, we will have the following closed-form formula for computing  $(q^k, \text{vec}(Q^k))$ :

$$\begin{bmatrix} q^k \\ \text{vec}(Q^k) \end{bmatrix} = \left( \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \right)^k \begin{bmatrix} q^0 \\ \text{vec}(Q^0) \end{bmatrix} + \sum_{t=0}^{k-1} \left( \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \right)^{k-1-t} \begin{bmatrix} u_q^t \\ u_Q^t \end{bmatrix}. \quad (\text{B.2})$$

The first term on the right side of the above equation will be guaranteed to converge to 0 if we have the stability condition  $\sigma(\mathcal{H}) < 1$ . As discussed in A.1, the stability condition  $\sigma(\mathcal{H}) < 1$  is fairly tight. Based on Proposition 3.6 in [14], we know that  $\mathcal{H}$  is Schur stable if and only if  $\mathcal{H}_{22}$  is Schur stable. Therefore, the needed stability condition is  $\sigma(\mathcal{H}_{22}) < 1$ . Under this condition, if we have  $p_i^k \rightarrow p_i^\infty$ , then Statement 2 in Proposition 3 can be used to show

$$\begin{aligned} u_q^\infty &= \lim_{k \rightarrow \infty} u_q^k = \begin{bmatrix} p_{11}G_1 & \dots & p_{n1}G_n \\ \vdots & \ddots & \vdots \\ p_{1n}G_1 & \dots & p_{nn}G_n \end{bmatrix} \begin{bmatrix} p_1^\infty I_{n_\xi} \\ \vdots \\ p_n^\infty I_{n_\xi} \end{bmatrix}, \\ u_Q^\infty &= \lim_{k \rightarrow \infty} u_Q^k = \begin{bmatrix} p_{11}G_1 \otimes G_1 & \dots & p_{n1}G_n \otimes G_n \\ \vdots & \ddots & \vdots \\ p_{1n}G_1 \otimes G_1 & \dots & p_{nn}G_n \otimes G_n \end{bmatrix} \begin{bmatrix} p_1^\infty I_{n_\xi^2} \\ \vdots \\ p_n^\infty I_{n_\xi^2} \end{bmatrix}, \\ \begin{bmatrix} q^\infty \\ \text{vec}(Q^\infty) \end{bmatrix} &= \lim_{k \rightarrow \infty} \begin{bmatrix} q^k \\ \text{vec}(Q^k) \end{bmatrix} = (I - \mathcal{H})^{-1} \begin{bmatrix} u_q^\infty \\ u_Q^\infty \end{bmatrix}. \end{aligned} \quad (\text{B.3})$$

Finally, if  $\|p^k - p^\infty\| \leq C_p \tilde{\rho}^k$ , then there exists a constant  $C$  such that the following inequality holds.

$$\left\| \begin{bmatrix} u_q^k \\ u_Q^k \end{bmatrix} - \begin{bmatrix} u_q^\infty \\ u_Q^\infty \end{bmatrix} \right\| \leq C \tilde{\rho}^k.$$

If we know  $\sigma(\mathcal{H}_{22}) < 1$ , then we can directly apply Proposition 3 to obtain the following linear convergence result:

$$\left\| \begin{bmatrix} q^k \\ \text{vec}(Q^k) \end{bmatrix} - \begin{bmatrix} q^\infty \\ \text{vec}(Q^\infty) \end{bmatrix} \right\| \leq C_0 \max\{\sigma(\mathcal{H}) + \varepsilon, \tilde{\rho}\}^k,$$

where  $C_0$  is some constant and  $\varepsilon$  is an arbitrarily small positive number. We can see that the convergence rates of  $\text{vec}(Q^k)$  and  $q^k$  depend on both  $\sigma(\mathcal{H})$  and the mixing rate of the underlying Markov jump parameter  $\{z^k\}$  (which is denoted as  $\tilde{\rho}$ ).

Therefore, when the underlying Markov chain  $\{z^k\}$  is aperiodic and irreducible, the mean and covariance of the MJLS just converges to the steady state values at a linear rate specified by  $\max\{\sigma(\mathcal{H}) + \varepsilon, \tilde{\rho}\}$ . This is a powerful result that can be potentially applied to more general stochastic approximation schemes other than (5). We also want to mention that there are other proofs for the convergence of  $(q^k, Q^k)$ . See Proposition 3.36 in [14] for an alternative proof. Here, our result is a little bit stronger than Proposition 3.36 in [14] since we also specify the convergence rate of  $(q^k, Q^k)$ .

Finally, it is worth mentioning that under the assumption  $\sigma(\mathcal{H}_{22}) < 1$ , one can further prove  $\{Q^k\}$  converges to a stationary process in some sense. This is a stronger result. Specifically, Proposition 3.37 In [14] shows that the MJLS is ‘‘asymptotically wide sense stationary’’ under the assumption  $\sigma(\mathcal{H}_{22}) < 1$ . We are not that interested in the correlation between the updates at different steps since our main purpose is analyzing TD learning. Hence we will skip a detailed discussion on this topic. See Chapter 3.4 in [14] for a thorough treatment.

## C Detailed proofs

### C.1 Detailed proofs of Propositions 1, 2, and 3

We believe that all the statements in Propositions 1, 2, and 3 are known in the controls field. Since we are not able to find a reference to exactly match the statements, we provide a proof here for completeness.

We will need the following lemma.

**Lemma C.1.** *Consider the LTI model (1). Suppose  $\sigma(\mathcal{H}) < 1$ . We set  $\rho = \sigma(\mathcal{H}) + \varepsilon$  where  $\varepsilon$  is an arbitrary small positive number. Then there exists a positive definite matrix  $V$  and a positive constant  $\gamma$  s.t. the following inequality holds for all  $k$ ,*

$$(x^{k+1})^\top V x^{k+1} \leq \rho^2 (x^k)^\top V x^k + \gamma \|u^k\|^2. \quad (\text{C.1})$$

*Proof.* We know  $\rho^{-1}\mathcal{H}$  is Schur stable. Based on Theorem 8.4 in [27], there exists a positive definite matrix  $V$  such that

$$\rho^{-2}\mathcal{H}^\top V \mathcal{H} - V < 0,$$

where the matrix inequality holds in the negative definite sense. The above condition is actually equivalent to  $\mathcal{H}^\top V \mathcal{H} - \rho^2 V < 0$ . Choose the matrix  $V$  that satisfies  $\mathcal{H}^\top V \mathcal{H} - \rho^2 V < 0$ . Then there exists a sufficiently large  $\gamma$  such that  $\mathcal{G}^\top V \mathcal{G} - \gamma I < 0$  and  $\mathcal{H}^\top V \mathcal{H} - \rho^2 \mathcal{H} - \mathcal{H}^\top V \mathcal{G}(\mathcal{G}^\top V \mathcal{G} - \gamma I)^{-1} \mathcal{G}^\top V \mathcal{H} < 0$ . By Schur complement lemma, this is equivalent to

$$\begin{bmatrix} \mathcal{H}^\top V \mathcal{H} - \rho^2 V & \mathcal{H}^\top V \mathcal{G} \\ \mathcal{G}^\top V \mathcal{H} & \mathcal{G}^\top V \mathcal{G} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & -\gamma I \end{bmatrix} < 0.$$

Now we left and right multiply the right side of the above matrix inequality with  $\begin{bmatrix} (x^k)^\top & (u^k)^\top \end{bmatrix}$  and  $\begin{bmatrix} (x^k)^\top & (u^k)^\top \end{bmatrix}^\top$ . This leads to

$$\begin{bmatrix} x^k \\ u^k \end{bmatrix}^\top \begin{bmatrix} \mathcal{H}^\top V \mathcal{H} - \rho^2 V & \mathcal{H}^\top V \mathcal{G} \\ \mathcal{G}^\top V \mathcal{H} & \mathcal{G}^\top V \mathcal{G} \end{bmatrix} \begin{bmatrix} x^k \\ u^k \end{bmatrix} + \begin{bmatrix} x^k \\ u^k \end{bmatrix}^\top \begin{bmatrix} 0 & 0 \\ 0 & -\gamma I \end{bmatrix} \begin{bmatrix} x^k \\ u^k \end{bmatrix} \leq 0.$$

One can verify that the first term on the left side of the above inequality is just equal to  $(x^{k+1})^\top V x^{k+1} - \rho^2 (x^k)^\top V x^k$  as follows

$$\begin{aligned} (x^{k+1})^\top V x^{k+1} - \rho^2 (x^k)^\top V x^k &= (\mathcal{H}x^k + \mathcal{G}u^k)^\top V (\mathcal{H}x^k + \mathcal{G}u^k) - \rho^2 (x^k)^\top V x^k \\ &= \begin{bmatrix} x^k \\ u^k \end{bmatrix}^\top \begin{bmatrix} \mathcal{H}^\top V \mathcal{H} - \rho^2 V & \mathcal{H}^\top V \mathcal{G} \\ \mathcal{G}^\top V \mathcal{H} & \mathcal{G}^\top V \mathcal{G} \end{bmatrix} \begin{bmatrix} x^k \\ u^k \end{bmatrix}. \end{aligned}$$

Therefore, we have  $(x^{k+1})^\top V x^{k+1} - \rho^2 (x^k)^\top V x^k - \gamma \|u^k\|^2 \leq 0$ , which is equivalent to (C.1).  $\square$

Now we are ready to prove Proposition 1. Since  $\sigma(\mathcal{H}) < 1$ ,  $x^\infty$  can still be well defined as  $x^\infty = (I - \mathcal{H})^{-1} \mathcal{G}u$ . Notice we have not shown the existence of  $\lim_{k \rightarrow \infty} x^k$  at this point. We will show  $\lim_{k \rightarrow \infty} x^k$  exists and is equal to  $x^\infty$ . Applying the relation  $(I - \mathcal{H})x^\infty = \mathcal{G}u^\infty$ , we still have

$$x^{k+1} - x^\infty = \mathcal{H}(x^k - x^\infty) + \mathcal{G}(u^k - u^\infty).$$

By Lemma C.1, there exists a positive definite matrix  $V$  and a positive number  $\gamma$  such that

$$(x^{k+1} - x^\infty)^\top V (x^{k+1} - x^\infty) \leq \rho^2 (x^k - x^\infty)^\top V (x^k - x^\infty) + \gamma \|u^k - u^\infty\|^2, \quad (\text{C.2})$$

where  $\rho = \sigma(\mathcal{H}) + \varepsilon < 1$ . First we show  $(x^k - x^\infty)^\top V (x^k - x^\infty)$  is bounded for all  $k$  and then we apply lim sup to the above inequality. Clearly there exists a constant  $U$  such that  $\|u^k - u^\infty\|^2 \leq U$  for all  $k$ . Then we have

$$(x^k - x^\infty)^\top V (x^k - x^\infty) \leq \rho^{2k} (x^0 - x^\infty)^\top V (x^0 - x^\infty) + \sum_{t=0}^{k-1} \rho^{2t} \gamma U \leq (x^0 - x^\infty)^\top V (x^0 - x^\infty) + \frac{\gamma U}{1 - \rho^2}.$$

Therefore,  $\limsup_{k \rightarrow \infty} (x^k - x^\infty)^\top V (x^k - x^\infty)$  is finite. Now we take lim sup on both sides of (C.2) and will immediately be able to show  $\limsup_{k \rightarrow \infty} (x^k - x^\infty)^\top V (x^k - x^\infty) = 0$ . Since  $V$  is positive definite, we have  $x^k \rightarrow x^\infty$ , which is the desired conclusion.

Next, we prove Proposition 2. If  $u^k = u$  for all  $k$ , we have  $\sum_{t=0}^{k-1} (\mathcal{H})^{k-1-t} \mathcal{G}u^t = \sum_{t=0}^{k-1} (\mathcal{H})^t \mathcal{G}u$ . When  $\sigma(\mathcal{H}) < 1$ , we have  $(\mathcal{H})^k \rightarrow 0$  and  $\sum_{k=0}^\infty (\mathcal{H})^k = (I - \mathcal{H})^{-1}$ . Hence we have  $x^\infty = \lim_{k \rightarrow \infty} x^k =$

$(I - \mathcal{H})^{-1}\mathcal{G}u$ . Clearly,  $(I - \mathcal{H})$  is nonsingular due to the fact  $\sigma(\mathcal{H}) < 1$ . Therefore, we have  $(I - \mathcal{H})x^\infty = \mathcal{G}u^\infty = \mathcal{G}u$ . Now it is straightforward to show

$$x^{k+1} - x^\infty = \mathcal{H}x^k + \mathcal{G}u^k - x^\infty = \mathcal{H}(x^k - x^\infty) + \mathcal{G}u^k - (I - \mathcal{H})x^\infty = \mathcal{H}(x^k - x^\infty),$$

which directly leads to the desired conclusion  $x^k = x^\infty + (\mathcal{H})^k(x^0 - x^\infty)$ .

Finally, we will still use (C.2) to prove Proposition 3. It is assumed that the arbitrary small  $\varepsilon$  is chosen in a way that  $\varepsilon + \sigma(\mathcal{H}) \neq \tilde{\rho}$  since one can always decrease  $\varepsilon$  by a tiny bit. Then iterating (C.2) leads to

$$\begin{aligned} (x^k - x^\infty)^\top V(x^k - x^\infty) &\leq \rho^{2k}(x^0 - x^\infty)^\top V(x^0 - x^\infty) + \gamma \sum_{t=0}^{k-1} \rho^{2(k-1-t)} \|u^t - u^\infty\|^2 \\ &\leq \rho^{2k}(x^0 - x^\infty)^\top V(x^0 - x^\infty) + C^2 \gamma \sum_{t=0}^{k-1} \rho^{2(k-1-t)} \tilde{\rho}^{2t} \\ &= \rho^{2k}(x^0 - x^\infty)^\top V(x^0 - x^\infty) + \left( \frac{C^2 \gamma}{\rho^2 - \tilde{\rho}^2} \right) (\rho^{2k} - \tilde{\rho}^{2k}). \end{aligned}$$

Obviously the right side of the above inequality is on the order of  $O((\max\{\rho, \tilde{\rho}\})^{2k})$ . Hence we have

$$\|x^k - x^\infty\|^2 \leq \frac{1}{\lambda_{\min}(V)} (x^k - x^\infty)^\top V(x^k - x^\infty) = O((\max\{\rho, \tilde{\rho}\})^{2k}),$$

which leads to  $\|x^k - x^\infty\| = O((\max\{\rho, \tilde{\rho}\})^k)$ . This completes the proof of this proposition.

An interesting thing is that when  $\rho = \tilde{\rho}$ , the convergence rate is actually on the order of  $O(k\rho^k)$ . Specifically, we have

$$(x^k - x^\infty)^\top V(x^k - x^\infty) \leq \rho^{2k}(x^0 - x^\infty)^\top V(x^0 - x^\infty) + C^2 \gamma k \rho^{2(k-1)}.$$

Of course this rate is always bounded above by  $O((\varepsilon + \rho)^k)$ . In addition, if it happens  $\varepsilon + \sigma(\mathcal{H}) = \tilde{\rho}$ , one can always decrease  $\varepsilon$  by a tiny bit and the convergence rate becomes linear again.

## C.2 A detailed proof for Theorem 1

The underlying probability space is denoted by  $(\Omega, \mathcal{F}, \mathbb{P})$ . We denote by  $\mathcal{F}_k$  the  $\sigma$ -algebra generated by  $(z^0, z^1, \dots, z^k)$ . Clearly,  $z^k$  is  $\mathcal{F}_k$ -adapted and we obtain a filtered probability space  $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}, \mathbb{P})$  on which the stochastic optimization method is defined.

First, we prove  $\mu^{k+1} = (I + \alpha \bar{A})\mu^k$ . Since  $\mathbb{E}b(z^k) = \sum_{i=1}^n p_i b_i = 0$ , we have

$$\mathbb{E}(\xi^{k+1} | \mathcal{F}_{k-1}) = \sum_{i=1}^n p_i \left( (I + \alpha A_i) \xi^k + \alpha b_i \right) = \left( I + \alpha \left( \sum_{i=1}^n p_i A_i \right) \right) \xi^k + \alpha \sum_{i=1}^n p_i b_i = (I + \alpha \bar{A}) \xi^k.$$

Taking full expectation of the above equation leads to  $\mu^{k+1} = (I + \alpha \bar{A})\mu^k$ .

Next, we prove  $\mathbb{Q}^{k+1} = \mathbb{Q}^k + \alpha(\bar{A}\mathbb{Q}^k + \mathbb{Q}^k \bar{A}^\top) + \alpha^2 \sum_{i=1}^n p_i (A_i \mathbb{Q}^k A_i^\top + 2 \text{sym}(A_i \mu^k b_i^\top) + b_i b_i^\top)$ . We can use a similar argument. We have

$$\begin{aligned} &\mathbb{E}(\xi^{k+1} (\xi^{k+1})^\top | \mathcal{F}_{k-1}) \\ &= \sum_{i=1}^n \left( p_i ((I + \alpha A_i) \xi^k + \alpha b_i) ((I + \alpha A_i) \xi^k + \alpha b_i)^\top \right) \\ &= \sum_{i=1}^n p_i (I + \alpha A_i) \xi^k (\xi^k)^\top (I + \alpha A_i)^\top + \sum_{i=1}^n \alpha p_i b_i (\xi^k)^\top (I + \alpha A_i)^\top + \sum_{i=1}^n \alpha p_i (I + \alpha A_i) \xi^k b_i^\top + \alpha^2 \sum_{i=1}^n p_i b_i b_i^\top. \end{aligned}$$

Taking full expectation and applying the fact  $\sum_{i=1}^n p_i b_i = 0$  leads to

$$\begin{aligned} \mathbb{Q}^{k+1} &= \sum_{i=1}^n p_i (I + \alpha A_i) \mathbb{Q}^k (I + \alpha A_i)^\top + \alpha^2 \sum_{i=1}^n p_i \left( b_i (\mu^k)^\top A_i^\top + A_i \mu^k b_i^\top + b_i b_i^\top \right) \\ &= \mathbb{Q}^k + \alpha(\bar{A}\mathbb{Q}^k + \mathbb{Q}^k \bar{A}^\top) + \alpha^2 \sum_{i=1}^n p_i \left( A_i \mathbb{Q}^k A_i^\top + 2 \text{sym}(A_i \mu^k b_i^\top) + b_i b_i^\top \right). \end{aligned}$$



This proves the recursive formula for  $\mathbb{Q}^k$ . Now we can apply the vectorization operation to this formula. For any matrices  $A$ ,  $X$ , and  $B$ , we have  $\text{vec}(AXB) = (B^\top \otimes A) \text{vec}(X)$ . Hence we can directly show

$$\begin{aligned} \text{vec}(\mathbb{Q}^{k+1}) &= \text{vec}(\mathbb{Q}^k) + \alpha(\text{vec}(\bar{A}\mathbb{Q}^k) + \text{vec}(\mathbb{Q}^k\bar{A}^\top)) + \alpha^2 \sum_{i=1}^n p_i \text{vec} \left( A_i \mathbb{Q}^k A_i^\top + 2 \text{sym}(A_i \mu^k b_i^\top) + b_i b_i^\top \right) \\ &= \text{vec}(\mathbb{Q}^k) + \alpha(I \otimes \bar{A} + \bar{A} \otimes I) \text{vec}(\mathbb{Q}^k) + \alpha^2 \left( \sum_{i=1}^n p_i A_i \otimes A_i \right) \text{vec}(\mathbb{Q}^k) \\ &\quad + \alpha^2 \left( \sum_{i=1}^n p_i (b_i \otimes A_i + A_i \otimes b_i) \right) \mu^k + \alpha^2 \sum_{i=1}^n p_i (b_i \otimes b_i). \end{aligned}$$

Therefore, we have  $\text{vec}(\mathbb{Q}^{k+1}) = \mathcal{H}_{22} \text{vec}(\mathbb{Q}^k) + \mathcal{H}_{21} \mu^k + \alpha^2 \sum_{i=1}^n p_i (b_i \otimes b_i)$  where  $\mathcal{H}_{21}$  and  $\mathcal{H}_{22}$  are determined by (15). Putting this together with  $\mu^{k+1} = (I + \alpha \bar{A}) \mu^k$  gives us the LTI model in (14). Then notice we have

$$\left( \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \right)^t \begin{bmatrix} 0 \\ \alpha^2 \sum_{i=1}^n p_i (b_i \otimes b_i) \end{bmatrix} = \alpha^2 \begin{bmatrix} 0 \\ (\mathcal{H}_{22})^t (\sum_{i=1}^n p_i (b_i \otimes b_i)) \end{bmatrix}.$$

Recall that we have  $\mathcal{H} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix}$ . We can apply Proposition 3.6 in [14] to show that  $\mathcal{H}$  is Schur stable if and only if  $\mathcal{H}_{22}$  is Schur stable. Therefore, a direct application of Proposition 3 will lead to (16). This completes the proof for Theorem 1.

### C.3 A detailed proof for Theorem 2

One may prove this theorem as a corollary of Proposition 3.35 in [14]. For completeness, we add more detailed calculations and present the proof in a self-contained manner. One can update  $q_j^{k+1}$  as

$$\begin{aligned} q_j^{k+1} &= \sum_{i=1}^n \mathbb{E} \left( (H(z^k) \xi^k + \alpha b(z^k)) \mathbf{1}_{\{z^k=i\}} \mathbf{1}_{\{z^{k+1}=j\}} \right) \\ &= \sum_{i=1}^n \left( H_i \mathbb{E}(\xi^k \mathbf{1}_{\{z^k=i\}}) \mathbb{P}(\mathbf{1}_{\{z^{k+1}=j\}} | \xi^k \mathbf{1}_{\{z^k=i\}}) + \alpha \mathbb{E}(b(z^k) \mathbf{1}_{\{z^k=i\}}) \mathbb{P}(\mathbf{1}_{\{z^{k+1}=j\}} | \xi^k \mathbf{1}_{\{z^k=i\}}) \right) \\ &= \sum_{i=1}^n p_{ij} \left( (I + \alpha A_i) q_i^k + \alpha p_i^k b_i \right). \end{aligned}$$

This leads to the following update rule for  $q^k$ :

$$\begin{bmatrix} q_1^{k+1} \\ \vdots \\ q_n^{k+1} \end{bmatrix} = \begin{bmatrix} p_{11}(I + \alpha A_1) & \dots & p_{n1}(I + \alpha A_n) \\ \vdots & \ddots & \vdots \\ p_{1n}(I + \alpha A_1) & \dots & p_{nn}(I + \alpha A_n) \end{bmatrix} \begin{bmatrix} q_1^k \\ \vdots \\ q_n^k \end{bmatrix} + \alpha \begin{bmatrix} p_{11} p_1^k I & \dots & p_{n1} p_n^k I \\ \vdots & \ddots & \vdots \\ p_{1n} p_1^k I & \dots & p_{nn} p_n^k I \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix},$$

which can be compactly written as  $q^{k+1} = (P^\top \otimes I) \text{diag}(I + \alpha A_i) q^k + \alpha((P^\top \text{diag}(p_i^k)) \otimes I_{n_\xi}) b$ , where  $b$  is the augmented vector

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

This proves  $q^{k+1} = \mathcal{H}_{11} q^k + \alpha((P^\top \text{diag}(p_i^k)) \otimes I_{n_\xi}) b$ , where  $\mathcal{H}_{11}$  is given by (21).

Next, we perform similar steps to obtain the iterative formula for  $Q^k$ . One can update  $Q_j^{k+1}$  as

$$\begin{aligned} Q_j^{k+1} &= \sum_{i=1}^n \mathbb{E} \left( (H(z^k) \xi^k + \alpha b(z^k)) (H(z^k) \xi^k + \alpha b(z^k))^\top \mathbf{1}_{\{z^k=i\}} \mathbf{1}_{\{z^{k+1}=j\}} \right) \\ &= \sum_{i=1}^n \left( H_i \mathbb{E}(\xi^k (\xi^k)^\top \mathbf{1}_{\{z^k=i\}}) H_i^\top \mathbb{P}(\mathbf{1}_{\{z^{k+1}=j\}} | \mathbf{1}_{\{z^k=i\}}) \right) + \alpha \sum_{i=1}^n \left( H_i \mathbb{E}(\xi^k b(z^k)^\top \mathbf{1}_{\{z^k=i\}}) \mathbb{P}(\mathbf{1}_{\{z^{k+1}=j\}} | \mathbf{1}_{\{z^k=i\}}) \right) \\ &\quad + \alpha \sum_{i=1}^n \left( \mathbb{E}(b(z^k) (\xi^k)^\top \mathbf{1}_{\{z^k=i\}}) H_i^\top \mathbb{P}(\mathbf{1}_{\{z^{k+1}=j\}} | \mathbf{1}_{\{z^k=i\}}) \right) + \alpha^2 \sum_{i=1}^j \mathbb{E}(b(z^k) b(z^k)^\top \mathbf{1}_{\{z^k=i\}}) \mathbb{P}(\mathbf{1}_{\{z^{k+1}=j\}} | \mathbf{1}_{\{z^k=i\}}) \\ &= \sum_{i=1}^n p_{ij} \left( H_i Q_i^k H_i^\top + 2\alpha \text{sym}(H_i q_i^k b_i^\top) + \alpha^2 p_i^k b_i b_i^\top \right). \end{aligned}$$

Now we can apply the vectorization operation to obtain the following equation

$$\text{vec}(Q_j^{k+1}) = \sum_{i=1}^n p_{ij} \left( (H_i \otimes H_i) \text{vec}(Q_i^k) + \alpha(b_i \otimes H_i + H_i \otimes b_i) q_i^k + \alpha^2 p_i^k b_i \otimes b_i \right),$$

which is equivalent to

$$\begin{aligned} \begin{bmatrix} \text{vec}(Q_1^{k+1}) \\ \vdots \\ \text{vec}(Q_n^{k+1}) \end{bmatrix} &= \begin{bmatrix} p_{11}H_1 \otimes H_1 & \dots & p_{n1}H_n \otimes H_n \\ \vdots & \ddots & \vdots \\ p_{1n}H_1 \otimes H_1 & \dots & p_{nn}H_n \otimes H_n \end{bmatrix} \begin{bmatrix} \text{vec}(Q_1^k) \\ \vdots \\ \text{vec}(Q_n^k) \end{bmatrix} + \alpha^2 \begin{bmatrix} p_{11}p_1^k I_{n_\xi} & \dots & p_{n1}p_n^k I_{n_\xi^2} \\ \vdots & \ddots & \vdots \\ p_{1n}p_1^k I_{n_\xi^2} & \dots & p_{nn}p_n^k I_{n_\xi^2} \end{bmatrix} \begin{bmatrix} b_1 \otimes b_1 \\ \vdots \\ b_n \otimes b_n \end{bmatrix} \\ &+ \alpha \begin{bmatrix} p_{11}(b_1 \otimes H_1 + H_1 \otimes b_1) & \dots & p_{n1}(b_n \otimes H_n + H_n \otimes b_n) \\ \vdots & \ddots & \vdots \\ p_{1n}(b_1 \otimes H_1 + H_1 \otimes b_1) & \dots & p_{nn}(b_n \otimes H_n + H_n \otimes b_n) \end{bmatrix} \begin{bmatrix} q_1^k \\ \vdots \\ q_n^k \end{bmatrix}. \end{aligned} \quad (\text{C.3})$$

We can compactly rewrite the above equation as  $\text{vec}(Q^{k+1}) = \mathcal{H}_{22} \text{vec}(Q^k) + \mathcal{H}_{21} q^k + \alpha^2 \text{diag}(p_i^k)(P^\top \otimes I_{n_\xi^2}) \hat{B}$ , where  $\mathcal{H}_{22}$  and  $\mathcal{H}_{21}$  are given by (21), and  $\hat{B} = [(b_1 \otimes b_1)^\top \dots (b_n \otimes b_n)^\top]^\top$ . Putting the recursion formulas for  $q^k$  and  $\text{vec}(Q^k)$  together leads to the desired state-space model (20). The rest of the theorem statement follows from direct applications of Equation (2).

## D Details for perturbation analysis under the Markov assumption

The perturbation analysis in Section 5 relies on a few technical lemmas from matrix perturbation theory. We provide more details here. We will use the following fact.

**Proposition D.1.** *Suppose  $\lambda$  is a simple eigenvalue of  $A$  with left eigenvector  $y$  and right eigenvector  $x$ . Suppose  $B$  and  $A \otimes I_m$  have the same dimension. Let  $c$  be an eigenvalue of the  $m \times m$  matrix  $(y \otimes I_m)B(x \otimes I_m)$ . Then the matrix  $(A \otimes I_m) + \alpha B$  has an eigenvalue yielding the first-order expansion  $\lambda + c\alpha + O(\alpha^2)$  for small  $\alpha$ .*

The above proposition is a special case of Theorem 2.1 in [42]. See the remark placed after Theorem 2.1 in [42] for explanations. Now we can directly apply the above proposition to analyze the spectral radius of  $\mathcal{H}_{11}$  and  $\mathcal{H}_{22}$ . First recall that  $\mathcal{H}_{11} = (P^\top \otimes I_{n_\xi}) \text{diag}(I_{n_\xi} + \alpha A_i) = P^\top \otimes I_{n_\xi} + \alpha (P^\top \otimes I_{n_\xi}) \text{diag}(A_i)$ . Based on the ergodicity assumption on  $\{z^k\}$ , 1 is a simple eigenvalue of  $P^\top$  with left eigenvector  $y = [1 \ 1 \ \dots \ 1]$  and right eigenvector  $p^\infty$  which is the unique stationary distribution of  $\{z^k\}$ . Since we have  $(y \otimes I_{n_\xi})(P^\top \otimes I_{n_\xi}) \text{diag}(A_i)(p^\infty \otimes I_{n_\xi}) = \sum_{i=1}^n p_i^\infty A_i = \bar{A}$ , we can directly apply the above proposition to show

$$\lambda_{\max}(\mathcal{H}_{11}) = 1 + \lambda_{\max \text{ real}}(\bar{A})\alpha + O(\alpha^2). \quad (\text{D.1})$$

Therefore, we have

$$\sigma(\mathcal{H}_{11}) = \sqrt{(1 + \alpha \text{real}(\lambda_{\max \text{ real}}(\bar{A})))^2 + (\text{imag}(\lambda_{\max \text{ real}}(\bar{A})))^2} \approx 1 + \text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha + O(\alpha^2).$$

Next, we do a similar perturbation analysis to show  $\mathcal{H}_{22} = 1 + 2 \text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha + O(\alpha^2)$ . Recall that we have

$$\begin{aligned} \mathcal{H}_{22} &= (P^\top \otimes I_{n_\xi^2}) \text{diag}((I_{n_\xi} + \alpha A_i) \otimes (I_{n_\xi} + \alpha A_i)) \\ &= P^\top \otimes I_{n_\xi^2} + \alpha ((P^\top \otimes I_{n_\xi^2}) \text{diag}(A_i \otimes I_{n_\xi} + I_{n_\xi} \otimes A_i)) + O(\alpha^2). \end{aligned}$$

Since we have  $(y \otimes I_{n_\xi^2})(P^\top \otimes I_{n_\xi^2}) \text{diag}(A_i \otimes I_{n_\xi} + I_{n_\xi} \otimes A_i)(p^\infty \otimes I_{n_\xi^2}) = \bar{A} \otimes I_{n_\xi} + I_{n_\xi} \otimes \bar{A}$ , we can directly apply the above matrix perturbation proposition to show

$$\lambda_{\max}(\mathcal{H}_{22}) = 1 + 2\lambda_{\max \text{ real}}(\bar{A})\alpha + O(\alpha^2), \quad (\text{D.2})$$

which leads to the desired first order expansion of  $\sigma(\mathcal{H}_{22})$ . Another fact that we used in the above argument is that all the eigenvalues of  $\bar{A} \otimes I_{n_\xi} + I_{n_\xi} \otimes \bar{A}$  are in the form of a sum of two eigenvalues of  $\bar{A}$ .

**A remark on the IID case.** For the IID case, we have  $\mathcal{H}_{11} = I + \alpha \bar{A}$ . Hence the eigenvalues of  $\mathcal{H}_{22}$  are in the form of  $1 + \alpha \lambda$  where  $\lambda$  is an eigenvalue of  $\bar{A}$ . So there is no need to even perform a perturbation analysis here. We directly have

$$\sigma(\mathcal{H}_{11}) = \sqrt{(1 + \alpha \text{real}(\lambda_{\max \text{ real}}(\bar{A})))^2 + (\text{imag}(\lambda_{\max \text{ real}}(\bar{A})))^2} \approx 1 + \text{real}(\lambda_{\max \text{ real}}(\bar{A}))\alpha + O(\alpha^2).$$

To analyze  $\sigma(\mathcal{H}_{22})$ , first recall that we have  $\mathcal{H}_{22} = I_{n_\xi^2} + \alpha(I \otimes \bar{A} + \bar{A} \otimes I) + \alpha^2 \sum_{i=1}^n p_i(A_i \otimes A_i)$ . If we assume  $\lambda_{\max}(I_{n_\xi^2} + \alpha(I \otimes \bar{A} + \bar{A} \otimes I))$  is a semisimple eigenvalue, then we can apply Proposition D.1 to obtain  $\lambda_{\max}(\mathcal{H}_{22}) = 1 + 2\lambda_{\max \text{ real}}(\bar{A}) + O(\alpha^2)$ .

## E Connections to existing finite sample bounds on mean square errors

Most existing finite sample bounds for TD learning with a constant learning rate have the following form:

$$\mathbb{E}\|\xi^k\|^2 \leq C_0 \rho^{2k} + C_1, \quad (\text{E.1})$$

where  $C_0$  is a constant,  $\rho^2$  is the convergence rate, and  $C_1$  quantifies the final error level. Typically one proves  $\rho^2 = 1 - c\alpha + O(\alpha^2)$  for some  $c$ , and  $C_1 = O(\alpha)$ . One most relevant result of this nature for the Markov noise model was presented as Theorem 7 in [45]. Our result justifies the tightness of the result in [45] for the following reasons.

- In [45], the constant  $c$  in the rate  $\rho^2$  is a constant determined by  $\bar{A}$ . Using our perturbation analysis, we can see eventually  $c$  is going to be determined by the real part of  $\lambda_{\max \text{ real}}(\bar{A})$ . Actually one could modify the argument in [45] to match the constant  $c$  with our perturbation analysis result by choosing a slightly better Lyapunov function based on the solution of an linear matrix inequality  $\bar{A}^\top V + V\bar{A} + 2\rho V \preceq 0$ .
- In [45], the constant  $C_1$  is at the order of  $O(\alpha)$  which matches the perturbation analysis result obtained in our paper. It is possible that an exact formula for the constant  $C_1$  can be obtained to match the steady state mean square error  $\lim_{k \rightarrow \infty} \text{trace}(\mathbb{Q}^k)$  more accurately, although we have not pursued such an analysis.
- In [45], the rate  $\rho$  does not depend on the mixing time property. This is consistent with our theory. Based on our theory, as  $\alpha$  gets smaller, the rate  $\rho$  becomes independent of the mixing rate  $\bar{\rho}$ , although the constant  $C_0$  still has some dependence on  $\bar{\rho}$ .
- Our results further show that  $\mathbb{E}\|\xi^k\|^2$  converges to an exact limit at a linear rate. Hence both upper and lower bounds for the mean square TD error are simultaneously provided.

It is worth mentioning that the bounds in the form of (E.1) capture the behaviors of TD learning quite well for small  $\alpha$ , but can be conservative for large  $\alpha$ . Our formulas are exact for all  $\alpha$ . The generality comes at the price of loosing some interpretability for the large learning rate region. How to interpret  $\sigma(\mathcal{H}_{22})$  for larger  $\alpha$  in a better way remains unclear at this moment.

## F More discussions on jump system formulations for variants of TD(0)

Finally, we present some extra details for the jump system formulations of several TD learning algorithms other than TD(0). Specifically, all the methods that can be analyzed using the ODE method has the form  $\xi^{k+1} = (I + \alpha A(z^k))\xi^k + \alpha b(z^k)$ . Then taking expectation of  $A(z^k)$  and  $b(z^k)$  under the stationary distribution and making  $\alpha$  arbitrarily small leads to the ODE  $\dot{\xi} = \bar{A}\xi$ . As commented in Section 3, the linear stochastic approximation scheme  $\xi^{k+1} = (I + \alpha A(z^k))\xi^k + \alpha b(z^k)$  is just a MJLS. Now we give detailed references for this type of formulations for various TD learning algorithms. The detailed linear stochastic approximation form for GTD is given in Section 4 of [50]. The detailed linear stochastic approximation form for GTD2 is given in Section 5 of [49]. TDC yields a similar formulation. The double temporal difference (DTD) learning method and the average temporal difference (ATD) learning method are proposed in [38]. The ODE formulations for both DTD and ATD are presented in the supplementary material of [38], yielding straightforward jump system formulations.

It is also possible to model two time-scale methods or off-policy TD learning using the general jump system model (4). One needs to slightly modify the definitions of  $\{H_i\}$  and  $\{G_i\}$ . Then one can immediately apply the LTI model (9) to obtain closed-form formulas for the mean square error of these methods.