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

Bin Hu and Usman Ahmed Syed

Coordinated Science Lab and Electrical and Computer Engineering Department
University of Illinois Urbana Champaign

Overview

- Analyzing TD learning algorithms with linear function approximators by exploiting their connections to Markov jump linear systems (MJLS)
- Using MJLS theory to characterize the exact behaviors of the mean and covariance of estimation errors for many TD learning algorithms
- Tight matrix spectral radius condition to guarantee the convergence of the covariance matrix of TD estimation error under Markov assumption
- Formula for the exact limit of the Mean Square Error (MSE) of TD
- Convergence rate for TD learning with small or large learning rate
- Computing the upper and lower bounds on the MSE for TD learning.

Background: LTI Systems and MJLS

- A linear time-invariant (LTI) system is given by: $x^{k+1} = \mathcal{H}x^k + \mathcal{G}u^k$ where x^k and u^k are the state and input. Given x^0 and $\{u^k\}$, one has

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

- 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: $\xi^{k+1} = H(z^k)\xi^k + G(z^k)y^k$ where $H(z^k)$ and $G(z^k)$ are matrix functions of z^k. A key result for MJLS is that the exact formulas for mean q^k and covariance Q^k are available,

$$\text{where } q_i^k = \mathbb{E}\left(\xi^k \mathbf{1}_{\{z^k=i\}}\right), \quad Q_i^k = \mathbb{E}\left(\xi^k(\xi^k)^\mathsf{T}\mathbf{1}_{\{z^k=i\}}\right), \quad \mu^k = \mathbb{E}\xi^k,$$

$$\mathbb{Q}^k = \mathbb{E}\left(\xi^k(\xi^k)^\mathsf{T}\right) \quad q^k = \begin{bmatrix} q_1^k & \dots & q_n^k \end{bmatrix}^\mathsf{T} \quad Q^k = \begin{bmatrix} Q_1^k & Q_2^k & \dots & Q_n^k \end{bmatrix}.$$

TD learning as MJLS

TD learning variants such as **TD, TDC, GTD, GTD2, A-TD**, and **D-TD** are special cases of the following linear stochastic recursion:

$$\xi^{k+1} = \xi^k + \alpha\left(A(z^k)\xi^k + b(z^k)\right) \quad (2)$$

which is a MJLS with $H(z^k) = I + \alpha A(z^k)$, $G(z^k) = \alpha b(z^k)$, and $y^k = 1 \forall k$.

Example: TD(0) $\theta^{k+1} = \theta^k - \alpha\phi(s^k)\left((\phi(s^k) - \gamma\phi(s^{k+1}))^\mathsf{T}\theta^k - r(s^k)\right)$

Suppose θ^* is the vector that solves the projected Bellman equation. Let $z^k = \begin{bmatrix} (s^{k+1})^\mathsf{T} & (s^k)^\mathsf{T} \end{bmatrix}^\mathsf{T}$ and then rewrite the TD update as:

$$\theta^{k+1} - \theta^* = \left(I + \alpha A(z^k)\right)(\theta^k - \theta^*) + \alpha b(z^k)$$

where

$$A(z^k) = -\phi(s^k)(\phi(s^k) - \phi(s^{k+1}))^\mathsf{T}$$

$$b(z^k) = \phi(s^k)\left(r(s^k) - (\phi(s^k) - \phi(s^{k+1}))^\mathsf{T}\theta^*\right)$$

TD learning under IID Assumption

Theorem 1 *Consider a MJLS 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 set $\mathcal{H}_{11} = I + \alpha\bar{A}$, $\mathcal{H}_{21} = \alpha^2\sum_{i=1}^n p_i(A_i \otimes b_i + b_i \otimes A_i)$, and $\mathcal{H}_{22} = I_{n_\xi^2} + \alpha(I \otimes \bar{A} + \bar{A} \otimes I) + \alpha^2\sum_{i=1}^n p_i(A_i \otimes A_i)$. We have*

$$\begin{bmatrix} \mu^{k+1} \\ \text{vec}(\mathbb{Q}^{k+1}) \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \begin{bmatrix} \mu^k \\ \text{vec}(\mathbb{Q}^k) \end{bmatrix} + \begin{bmatrix} 0 \\ \alpha^2\sum_{i=1}^n p_i(b_i \otimes b_i) \end{bmatrix} \quad (3)$$

The input for the above LTI model does not change with k. Therefore, if $\sigma(\mathcal{H}_{22}) < 1$, the following exact formula holds

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

where $\mu^\infty = \lim_{k\to\infty}\mu^k = 0$ and

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

Stability Condition: LTI system (3) is stable if and only if $\mathcal{H}_{22}$ is Schur stable. For TD learning to converge, it is important to choose α such that $\sigma(\mathcal{H}_{22}) < 1$ for some given $\{A_i\}$, $\{b_i\}$ and $\{p_i\}$. Assuming α to be small, eigenvalue perturbation analysis to $\mathcal{H}_{22}$ suggests: $\sigma(\mathcal{H}_{22}) \approx 1 + 2\text{real}\left(\lambda_{\max\text{ real}}\left(\bar{A}\right)\right)\alpha + O(\alpha^2)$. Hence **as long as $\bar{A}$ is Hurwitz, there exists sufficiently small α s.t. $\sigma(\mathcal{H}_{22}) < 1$.**

Corollary 1 *Consider TD update (2) with $\bar{A}$ being Hurwitz. Suppose $\sigma(\mathcal{H}_{22}) < 1$ and $\mathbb{P}(z^k = i) = p_i \forall i$. Then $\delta^\infty := \lim_{k\to\infty}\mathbb{E}\|\theta^k - \theta^*\|^2$ exists and is given by $\delta^\infty = \text{trace}(\mathbb{Q}^\infty)$. Additionally, the following Mean Square TD error bounds hold for some constant C_0 and any arbitrary small $\varepsilon > 0$ (the rate $\sigma(\mathcal{H})$ is precise):*

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

Key Trade-off: *For small α, one can use perturbation to show $\lim_{k\to\infty}\mathbb{E}\|\theta^k - \theta^*\|^2 = O\left(\alpha\right)$ and $\sigma(\mathcal{H}) \approx 1 + \text{real}(\lambda_{\max\text{ real}}(\bar{A}))\alpha + O(\alpha^2)$. This gives the standard rate v.s. error trade-off.*

TD learning under Markov assumption

Theorem 2 *Consider the MJLS 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 $p^k = \begin{bmatrix} p_1^k & p_2^k & \dots & p_n^k \end{bmatrix}^\mathsf{T}$. Clearly $p^k = (P^\mathsf{T})^k p^0$. Further denote the augmented vectors as $b = \begin{bmatrix} b_1^\mathsf{T} & b_2^\mathsf{T} & \dots & b_n^\mathsf{T} \end{bmatrix}^\mathsf{T}$, $\hat{B} = \begin{bmatrix} (b_1 \otimes b_1)^\mathsf{T} & \dots & (b_n \otimes b_n)^\mathsf{T} \end{bmatrix}^\mathsf{T}$, 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 state-space model:*

$$\begin{bmatrix} q^{k+1} \\ \text{vec}(Q^{k+1}) \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \begin{bmatrix} q^k \\ \text{vec}(Q^k) \end{bmatrix} + \begin{bmatrix} \alpha((P^\mathsf{T}\,\text{diag}(p_i^k)) \otimes I_{n_\xi})b \\ \alpha^2((P^\mathsf{T}\,\text{diag}(p_i^k)) \otimes I_{n_\xi^2})\hat{B} \end{bmatrix} \quad (4)$$

$$\mathcal{H}_{11} = (P^\mathsf{T} \otimes I_{n_\xi})\,\text{diag}(I_{n_\xi} + \alpha A_i), \quad \mathcal{H}_{21} = \alpha\left(P^\mathsf{T} \otimes I_{n_\xi}\right)\text{diag}(S(b_i, A_i))$$

$$\mathcal{H}_{22} = (P^\mathsf{T} \otimes I_{n_\xi^2})\,\text{diag}((I_{n_\xi} + \alpha A_i) \otimes (I_{n_\xi} + \alpha A_i))$$

Key Difference: The input depends on p^k which changes over k. However, if the input converges linearly, the overall convergence behavior is similar.
Exact Solution: The augmented mean q^k and covariance Q^k can still be exactly computed by (1).
Stability Condition: The LTI system (4) is stable iff $\sigma(\mathcal{H}_{22}) < 1$. We need to choose α such that $\sigma(\mathcal{H}_{22}) < 1$ for some given $\{A_i\}$, $\{b_i\}$, P, and $\{p^0\}$. Define, $\bar{A} = \sum_{i=1}^n p_i^\infty A_i$ and let p^∞ be the unique stationary distribution of z^k. The eigenvalue perturbation analysis yields: $\sigma(\mathcal{H}_{22}) \approx 1 + 2\,\text{real}(\lambda_{\max\text{ real}}(\bar{A}))\alpha + O(\alpha^2)$. Therefore, **as long as $\bar{A}$ is Hurwitz, there exists sufficiently small α such that $\sigma(\mathcal{H}_{22}) < 1$.**

Corollary 2 *Consider the TD update (2) with $\bar{A}$ being Hurwitz. Let $\{z^k\}$ be a Markov chain sampled from $\mathcal{N}$ using the transition matrix P. Suppose $\sigma(\mathcal{H}_{22}) < 1$. Assume $p^k \to p^\infty$, then we have:*

$$q^\infty = \lim_{k\to\infty} q^k = \alpha(I - \mathcal{H}_{11})^{-1}((P^\mathsf{T}\,\text{diag}(p_i^\infty)) \otimes I_{n_\xi})b,$$

$$\text{vec}(Q^\infty) = \alpha^2(I_N - \mathcal{H}_{22})^{-1}\left(\alpha^{-2}\mathcal{H}_{21}q^\infty + ((P^\mathsf{T}\,\text{diag}(p_i^\infty)) \otimes I_{n_\xi^2})\hat{B}\right)$$

$$\delta^\infty = \lim_{k\to\infty}\mathbb{E}\|\theta^k - \theta^*\|^2 = (\mathbf{1}_n^\mathsf{T} \otimes \text{vec}(I_{n_\theta})^\mathsf{T})\,\text{vec}(Q^\infty)$$

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

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

where C_0 is a constant and ε is an arbitrary small positive number.

Key Messages:

- The MSE has an exact limit δ^∞. One can show $\delta^\infty = O(\alpha)$.
- For small α, the rate is $\sigma(\mathcal{H}) \approx 1 + \text{real}(\lambda_{\max\text{ real}}(\bar{A}))\alpha + O(\alpha^2)$.
- Trade-off: rate $1 + \text{real}(\lambda_{\max\text{ real}}(\bar{A}))\alpha + O(\alpha^2)$ v.s. error $O(\alpha)$.
- For large α, the rate is $\max\{\sigma(\mathcal{H}) + \varepsilon, \tilde{\rho}\}$ and cannot be faster than $\tilde{\rho}$ (the mixing rate of z^k). IID case does not have such an issue.