Characterizing the Exact Behaviors of Temporal Difference Learning Algorithms Using Markov Jump Linear 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 first and second order moments of a large family of TD learning algorithms.
- The evolution of an augmented versions of the mean and covariance matrix of TD learning exactly follows the trajectory of a deterministic linear time-invariant (LTI) dynamical system.
- Tools from Linear Control Systems literature are used to represent the asymptotic and non asymptotic behaviors of TD learning.

Objectives

- Closed-form expressions for the mean and covariance matrix of TD learning at any time step.
- Condition to guarantee the convergence of the covariance matrix of TD learning and perturbation analysis to characterize the dependence of the TD behaviors on learning rate and the other Markov Chain parameters.

TD learning as MJLS

Numerous TD learning algorithms including **TD, TDC, GTD, GTD2, A-TD**, and **D-TD** are just special cases of the following linear stochastic recursion: $\xi^{k+1} = \xi^k + \alpha\left(A(z^k)\xi^k + b(z^k)\right)$
The above model is a special case of a MJLS if we set $H(z^k) = I + \alpha A(z^k)$, $G(z^k) = \alpha b(z^k)$, 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
$$A(z^k) = -\phi(s^k)(\phi(s^k) - \phi(s^{k+1}))^\mathsf{T}$$
$$b(z^k) = \phi(s^k)\left(r(s^k) - (\phi(s^k) - \phi(s^{k+1}))^\mathsf{T}\theta^*\right)$$

Perturbation Analysis: IID Case

Perturbation analysis is aimed at choosing an α such that $\sigma(\mathcal{H}_{22}) < 1$ for some given $\{A_i\}$, $\{b_i\}$, and $\{p_i\}$. Define $\bar{A} = \sum_{i=1}^n p_i A_i$ then under mild technical conditions: $\sigma(\mathcal{H}_{22}) \approx 1 + 2\,\mathrm{real}(\lambda_{\max\,\mathrm{real}}(\bar{A}))\alpha + O(\alpha^2)$
Therefore, as long as $\bar{\mathbf{A}}$ **is Hurwitz**, there exists sufficiently small α such that $\sigma(\mathcal{H}_{22}) < 1$.

Background: LTI Systems and MJLS

The state-space model of a discrete time 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 at time k. $\mathcal{H}$ and $\mathcal{G}$ are the system and input matrices.

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.

Define $q_i^k = \mathbb{E}\left(\xi^k \mathbf{1}_{\{z^k=i\}}\right)$, $Q_i^k = \mathbb{E}\left(\xi^k(\xi^k)^\mathsf{T}\mathbf{1}_{\{z^k=i\}}\right)$, $\mu^k = \mathbb{E}\xi^k$, $\mathbb{Q}^k = \mathbb{E}\left(\xi^k(\xi^k)^\mathsf{T}\right)$, $q^k = \begin{bmatrix}q_1^k & \dots & q_n^k\end{bmatrix}^\mathsf{T}$ and $Q^k = \begin{bmatrix}Q_1^k & Q_2^k & \dots & Q_n^k\end{bmatrix}$

TD learning under Markov assumption

Theorem 1 *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 $\mathrm{vec}(Q^k)$ are governed by the following state-space model:*

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

$$\mathcal{H}_{11} = (P^\mathsf{T}\otimes I_{n_\xi})\,\mathrm{diag}(I_{n_\xi} + \alpha A_i),$$
$$\mathcal{H}_{21} = \alpha\left(P^\mathsf{T}\otimes I_{n_\xi}\right)\mathrm{diag}(S(b_i, A_i))$$
$$\mathcal{H}_{22} = (P^\mathsf{T}\otimes I_{n_\xi^2})\,\mathrm{diag}((I_{n_\xi} + \alpha A_i)\otimes(I_{n_\xi} + \alpha A_i))$$

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

$$q^k = (\mathcal{H}_{11})^k q^0 + \alpha\sum_{t=0}^{k-1}(\mathcal{H}_{11})^{k-1-t}((P^\mathsf{T}\,\mathrm{diag}(p_i^t))\otimes I_{n_\xi})b \qquad \mathrm{vec}(Q^k) = (\mathcal{H}_{22})^k\,\mathrm{vec}(Q^0) + \sum_{t=0}^{k-1}(\mathcal{H}_{22})^{k-1-t}\left(\mathcal{H}_{21}q^t + \alpha^2((P^\mathsf{T}\,\mathrm{diag}(p_i^t))\otimes I_{n_\xi^2})\hat{B}\right)$$

Perturbation Analysis: Markov Case

The desired stability condition is: $\sigma(\mathcal{H}_{22}) < 1$. Perturbation analysis aids in choosing 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 the Markov chain under the ergodicity assumption then the perturbation analysis yields: $\sigma(\mathcal{H}_{22}) \approx 1 + 2\,\mathrm{real}(\lambda_{\max\,\mathrm{real}}(\bar{A}))\alpha + O(\alpha^2)$ Therefore, as long as $\bar{\mathbf{A}}$ **is Hurwitz**, there exists sufficiently small α such that $\sigma(\mathcal{H}_{22}) < 1$.

TD learning under IID assumption

Theorem 2 *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 μ^k and $\mathrm{vec}(\mathbb{Q}^k)$ are governed by the following LTI system:*

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

$$\mathcal{H}_{11} = I + \alpha\bar{A} \qquad \mathcal{H}_{21} = \alpha^2\sum_{i=1}^n p_i(A_i\otimes b_i + b_i\otimes A_i)$$
$$\mathcal{H}_{22} = I_{n_\xi^2} + \alpha(I\otimes\bar{A} + \bar{A}\otimes I) + \alpha^2\sum_{i=1}^n p_i(A_i\otimes A_i)$$

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

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

If $\sigma(\mathcal{H}_{22}) < 1$, $\mu^\infty = \lim_{k\to\infty}\mu^k = 0$ $\mathrm{vec}(\mathbb{Q}^\infty) = -\alpha\left(I\otimes\bar{A} + \bar{A}\otimes I + \alpha\sum_{i=1}^n p_i(A_i\otimes A_i)\right)^{-1}\left(\sum_{i=1}^n p_i(b_i\otimes b_i)\right)$