ECE 586 BH: Interplay between Control and Machine Learning **Fall 2023**

Lecture 16

A Jump System Perspective on Temporal Difference Learning, Part II

Lecturer: Bin Hu, Date:10/12/2023

In this lecture, we will show how the jump system theory can be used to provide finite-time analysis for TD(0) with linear function approximation. We consider $V^\pi(s) \approx \theta^\mathsf{T}\phi(s)$ where ϕ is the feature vector. Recall that TD(0) iterates as

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

where $\{s_k\}$ is the underlying Markov chain under the policy π. As discussed in the last lecture, we can rewrite TD(0) as the following linear stochastic approximation scheme

$$\theta_{k+1} = \theta_k + \varepsilon(A_{i_k}\theta_k + b_{i_k}), \tag{16.1}$$

where i_k is the augmented vector of (s_{k+1}, s_k). If $\{s_k\}$ forms a Markov chain, then $\{i_k\}$ also forms a Markov chain. Suppose θ_π is the solution to the projected Bellman equation[1] for the given policy π. How can we obtain a finite time bound for the mean square TD estimation error $\mathbb{E}\|\theta_k - \theta_\pi\|^2$? Obtaining such a bound under the IID assumption on $\{i_k\}$ is relatively straightforward. However, for TD learning, $\{i_k\}$ is a Markov chain. Obtaining a finite sample bound under such a Markov assumption is not easy. There are several different techniques that can be used to tackle this difficulty. We will present one argument based on the Markov jump linear system (MJLS) theory from the control field. We will follow the outline below:

- Review of linear time-invariant (LTI) system theory
- Review of MJLS theory
- TD learning as MJLS

16.1 Review of LTI Systems

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

$$x_{k+1} = \mathcal{H}x_k + \mathcal{G}u_k, \tag{16.2}$$

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 [3, 1]. Next, we review several useful facts.

[1] The projected Bellman equation and the equation $\sum_{i=1}^n p_i^\infty b_i = 0$ are actually equivalent, i.e. we have $\sum_i p_i^\infty(A_i\theta_\pi + b_i) = 0$ where $p_i^\infty = \lim_{k\to\infty} P(s_k = i)$.

Closed-form formulas for x_k. Given an initial condition x_0 and an input sequence, the state x_k yields the following closed-form expression

$$x_k = (\mathcal{H})^k x_0 + \sum_{t=0}^{k-1} (\mathcal{H})^{k-1-t} \mathcal{G} u_t, \tag{16.3}$$

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

Necessary and sufficient stability condition: It is well-known that the LTI system (16.2) is stable if and only if $\mathcal{H}$ is Schur stable. When $\mathcal{H}$ is Schur stable, we know $(\mathcal{H})^k x^0 \to 0$ 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 $(\mathcal{H})^k x^0 \to \infty$ [3, Section 7.2].

Exact limit for x^k. Suppose the system is stable, i.e. $\sigma(\mathcal{H}) < 1$. If $\mathcal{H}$ is Schur stable and u_k converges to a limit u^∞, then x_k has an exact limit (i.e. $\lim_{k\to\infty} x_k$ exists), and we must have $x^\infty = \lim_{k\to\infty} x_k = (I - \mathcal{H})^{-1} \mathcal{G} u^\infty$. If $u_k = u \ \forall k$ and $\sigma(\mathcal{H}) < 1$, then the closed-form expression for x_k can be simplified as

$$x_k = x^\infty + (\mathcal{H})^k (x_0 - x_\infty), \tag{16.4}$$

which is a sum of a constant steady state term x^∞ and a matrix power term that decays at a linear rate specified by $\sigma(\mathcal{H})$. If u_k converges to u^∞ at a linear rate $\tilde{\rho}$, then x_k will converge to its limit point at a linear rate specified by $\max\{\sigma(\mathcal{H}) + \varepsilon, \tilde{\rho}\}$.

16.2 Review of MJLS

In the control field, the behaviors of MJLS have been extensively studied. A standard MJLS is governed by the following state-space model:

$$\xi_{k+1} = H_{i_k} \xi_k + G_{i_k} u_k \tag{16.5}$$

where $i_k \in \{1, 2, \cdots, N\}$ is the so-called jump parameter. For MJLS, $\{i_k\}$ is assumed to be a Markov chain. For every k, we have $H_{i_k} \in \{H_1, H_2, \cdots, H_N\}$ and $G_{i_k} \in \{G_1, G_2, \cdots, G_N\}$. A key result from the classic MJLS theory states that $\mathbb{E}\|\xi_k\|^2$ can be obtained as an output of a special linear time-invariant (LTI) system [2, Proposition 3.35]. Now we briefly review this result here.

Roughly speaking, the analysis is built upon the fact that some augmented versions of the mean and the covariance matrix of $\{\xi_k\}$ for the MJLS model (16.5) actually follow the dynamics of a deterministic LTI model in the form of (16.2) [2, Chapter 3]. Let us denote the transition probabilities for the Markov chain $\{i_k\}$ as $p_{ij} := \mathbb{P}(i_{k+1} = j | i_k = i)$ 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 for the event $i_k = i$ is denoted as $\mathbf{1}_{\{i_k=i\}}$. Now we define the following key quantities:

$$q_k^i = \mathbb{E}\left(\xi_k \mathbf{1}_{\{i_k=i\}}\right), \quad Q_k^i = \mathbb{E}\left(\xi_k(\xi_k)^\mathsf{T}\mathbf{1}_{\{i_k=i\}}\right).$$

Suppose $u_k = 1\ \forall k$. Based on [2, Proposition 3.35], we must have

$$q_{k+1}^j = \sum_{i=1}^n p_{ij}(H_i q_k^i + G_i p_k^i), \tag{16.6}$$

$$Q_{k+1}^j = \sum_{i=1}^n p_{ij}\left(H_i Q_k^i H_i^\mathsf{T} + 2\,\mathrm{sym}(H_i q_k^i G_i^\mathsf{T}) + p_k^i G_i G_i^\mathsf{T}\right), \tag{16.7}$$

where $p_k^i := \mathbb{P}(i_k = i)$. To see why the above equations hold, we have the following arguments[2]

$$\begin{aligned} q_{k+1}^j &= \sum_{i=1}^n \mathbb{E}\left((H_{i_k}\xi_k + G_{i_k}u_k)\mathbf{1}_{\{i_k=i\}}\mathbf{1}_{\{i_{k+1}=j\}}\right) \\ &= \sum_{i=1}^n \left(H_i \mathbb{E}(\xi_k \mathbf{1}_{\{z^k=i\}}\mathbb{P}(i_{k+1}=j|\mathcal{F}_k)) + G_i \mathbb{E}(\mathbf{1}_{\{i_k=i\}}\mathbb{P}(i_{k+1}=j|\mathcal{F}_k))\right) \\ &= \sum_{i=1}^n p_{ij}(H_i q_k^i + G_i p_k^i) \end{aligned}$$

(The proof for the formula of Q_{k+1}^j is very similar. Please derive it by yourself!)

If we further augment q_k^i and Q_k^i as

$$q_k = \begin{bmatrix} q_k^1 \\ \vdots \\ q_k^n \end{bmatrix}, \quad Q_k = \begin{bmatrix} Q_k^1 & Q_k^2 & \dots & Q_k^n \end{bmatrix},$$

then it is straightforward to rewrite (16.6) (16.7) as the following LTI system

$$\begin{bmatrix} q_{k+1} \\ \mathrm{vec}(Q_{k+1}) \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \begin{bmatrix} q_k \\ \mathrm{vec}(Q_k) \end{bmatrix} + \begin{bmatrix} u_k \\ v_k \end{bmatrix}, \tag{16.8}$$

[2] We use the fact that $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|\mathcal{F}))$ if Y is $\mathcal{F}$-measurable.

where $\mathcal{H}_{11}$, $\mathcal{H}_{21}$, $\mathcal{H}_{22}$, u_k, and v_k are defined as

$$\begin{aligned}
\mathcal{H}_{11} &= \begin{bmatrix} p_{11}H_1 & \dots & p_{n1}H_n \\ \vdots & \ddots & \vdots \\ p_{1n}H_1 & \dots & p_{nn}H_n \end{bmatrix}, \mathcal{H}_{22} = \begin{bmatrix} p_{11}H_1 \otimes H_1 & \dots & p_{n1}H_n \otimes H_n \\ \vdots & \ddots & \vdots \\ p_{1n}H_1 \otimes H_1 & \dots & p_{nn}H_n \otimes H_n \end{bmatrix}, \\
\mathcal{H}_{21} &= \begin{bmatrix} p_{11}(H_1 \otimes G_1 + G_1 \otimes H_1) & \dots & p_{n1}(H_n \otimes G_n + G_n \otimes H_n), \\ \vdots & \ddots & \vdots \\ p_{1n}(H_1 \otimes G_1 + G_1 \otimes H_1) & \dots & p_{nn}(H_n \otimes G_n + G_n \otimes H_n) \end{bmatrix}, \\
u_k &= \begin{bmatrix} p_{11}G_1 & \dots & p_{n1}G_n \\ \vdots & \ddots & \vdots \\ p_{1n}G_1 & \dots & p_{nn}G_n \end{bmatrix} \begin{bmatrix} p_k^1 I_{n_\xi} \\ \vdots \\ p_k^n I_{n_\xi} \end{bmatrix}, \\
v_k &= \begin{bmatrix} p_{11}G_1 \otimes G_1 & \dots & p_{n1}G_n \otimes G_n \\ \vdots & \ddots & \vdots \\ p_{1n}G_1 \otimes G_1 & \dots & p_{nn}G_n \otimes G_n \end{bmatrix} \begin{bmatrix} p_k^1 I_{n_\xi^2} \\ \vdots \\ p_k^n I_{n_\xi^2} \end{bmatrix}.
\end{aligned} \tag{16.9}$$

Therefore, q_k and $\mathrm{vec}(Q_k)$ follow the LTI dynamics (16.8) and can be analyzed using the LTI theory reviewed in the last section.

Exercise problem. Can you apply the above LTI system model to write out an analytical formula for $\mathbb{E}\|\xi_k\|^2$?

16.3 TD Learning as MJLS

Now we can treat the TD scheme as a MJLS:

$$\theta_{k+1} - \theta_\pi = \theta_k - \theta_\pi + \varepsilon A_{i_k}(\theta_k - \theta_\pi) + \varepsilon(A_{i_k}\theta_\pi + b_{i_k}),$$

which is equivalent to

$$\theta_{k+1} - \theta_\pi = (I + \varepsilon A_{i_k})(\theta_k - \theta_\pi) + \varepsilon(A_{i_k}\theta_\pi + b_{i_k}). \tag{16.10}$$

Denote $\xi_k = \theta_k - \theta_\pi$, $G_{i_k} = \varepsilon(A_{i_k}\theta_\pi + b_{i_k})$, and $H_{i_k} = I + \varepsilon A_{i_k}$. The above scheme just becomes a jump system. Then we can easily obtain an analytical formula for the mean square TD estimation error $\mathbb{E}\|\theta_k - \theta_\pi\|^2$ and draw various insights on how the learning rate will affect the system behavior. This is exactly the idea used in [4]. If you are interested, please read [4] for a detailed treatment.

Bibliography

[1] C. Chen. *Linear system theory and design.* Oxford University Press, Inc., 1998.

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

[3] J. Hespanha. *Linear systems theory.* Princeton university press, 2009.

[4] B. Hu and U. Syed. Characterizing the exact behaviors of temporal difference learning algorithms using Markov jump linear system theory. In *Advances in Neural Information Processing Systems*, pages 8479–8490, 2019.